A Mathematical Philosophy of Explanations in Mechanistic Interpretability

Kola Ayonrinde[1]*, **Louis Jaburi**[2]

[1]UK AI Security Institute
[2]Independent Researcher

Abstract

Mechanistic Interpretability aims to understand neural networks through causal explanations. We argue for the *Explanatory View Hypothesis*: that Mechanistic Interpretability research is a principled approach to understanding models because neural networks contain implicit explanations which can be extracted and understood. We hence show that Explanatory Faithfulness, an assessment of how well an explanation fits a model, is well-defined. We propose a definition of Mechanistic Interpretability (MI) as the practice of producing *Model-level*, *Ontic*, *Causal-Mechanistic*, and *Falsifiable* explanations of neural networks, allowing us to distinguish MI from other interpretability paradigms and detail MI's inherent limits. We formulate the *Principle of Explanatory Optimism*, a conjecture which we argue is a necessary precondition for the success of Mechanistic Interpretability.

1 Introduction

ML artifacts are *strange* objects. ML researchers have produced models with a wide range of cognitive capabilities that no human knows how to program a machine to do, from playing Go and poker at a superhuman level (Schrittwieser et al. 2020; Brown and Sandholm 2019) to folding proteins (Jumper et al. 2021), and solving advanced mathematical problems (Glazer et al. 2024)[1].

However, we did not design these systems. No human wrote the blueprint for how AI systems ought to perform a given task. Instead, neural networks organically learn to solve problems via gradient descent, given large quantities of data. Neural networks aren't built; they're grown.

Because we don't design neural networks, ML researchers typically do not know how their models perform a given task. Additionally, neural networks often solve problems in unintuitive ways, relying on concepts that are not obvious to humans (Widdicombe, Julier, and Kim 2018; Hosseini et al. 2018; Goodfellow, Shlens, and Szegedy 2014; Ilyas et al. 2019). This situation of relative ignorance about the processes that give rise to a neural network's capabilities leaves us with a scientific problem analogous to the natural sciences. A physicist might observe some natural dynamical system, like the weather, and seek an explanation allowing them to understand, predict, and possibly even steer the system. Similarly, *neural network interpretability* (henceforth just *interpretability*) is the process of understanding artificial neural networks using the scientific method.

In this way, we characterise Interpretability as *The Strange Science*: Interpretability is the science of understanding artificial neural phenomena, just as the natural sciences seek to understand natural phenomena. Interpretability researchers study formal systems using empirical methods — making observations, generating conjectures, and refuting those conjectures — to understand complex neural systems. Since interpretability is analogous to the natural sciences, a Philosophy of Interpretability should be inspired by our best understanding of the philosophy of science.

Recent works like Bereska and Gavves (2024), Sharkey et al. (2025), and Geiger, Potts, and Icard (2023) have explored the methods and assumptions of Mechanistic Interpretability (MI). Other works have explored the philosophically relevant components of MI (Millière and Buckner 2025; Harding 2023; Kästner and Crook 2024). In this work, we foreground the philosophical role of *explanation* in *Mechanistic Interpretability* specifically, and how this differs from previous interpretability paradigms.[2] In particular, inspired by the Information-Theoretic perspective, we can understand explanations in terms of their compressive power and ability to communicate *understanding which generalises*.

Understanding neural networks can provide affordances for interventions which are important for AI Safety, AI Ethics, and AI Cognitive Science (Bengio et al. 2025; Anwar et al. 2024; Chalmers 2025; Olah et al. 2020; Amodei 2025). Such understanding also allows us to improve the performance of neural networks and debug their failings (Lindsay and Bau 2023; Sharkey et al. 2025; Amodei 2025).

Contributions. Our contributions are as follows:

- Firstly, we show that producing compressive explanations frames a potential solution to the Interpretability Problem.

*Correspondence: koayon@gmail.com

[1]to deceiving humans in games (Golechha and Garriga-Alonso 2025; Bakhtin et al. 2022)

[2]Lipton (2018); Gilpin et al. (2018); Fleisher (2022); Doshi-Velez and Kim (2017); Leavitt and Morcos (2020); Erasmus, Brunet, and Fisher (2021) provide an overview of classical (pre-Mechanistic) Interpretability.

We hence define *explanatorily faithfulness* as the goal of Mechanistic Interpretability.

- Secondly, we provide a technical definition of MI and leverage this definition to highlight both the possibilities and limitations of MI.
- Thirdly, we formulate the *Principle of Explanatory Optimism*, a conjecture at the heart of MI which states that the algorithmic structure of generalising neural networks is human-understandable. We show that without the Principle of Explanatory Optimism the project of MI is intractable.

2 The Strange Science

Is Interpretability a natural science or a formal science? *Natural sciences* like physics, biology, and earth sciences seek to explain physical phenomena, creating hypotheses and running experiments. *Formal sciences* like algebra, decision theory, formal linguistics, and music theory are concerned with abstract systems and structures, using deductive methods to construct proofs from axioms. The defining strange property of neural network interpretability is that it is a natural science whose objects of study are both *artificial* (constructed by humans rather than naturally occurring) and *formal* (inherently abstract rather than physical systems).

Neural networks are mathematical objects; they are deterministic functions from the input embedding space to the output embedding space. Yet, the strange science of interpretability is not a formal science. Interpretability researchers are not primarily interested in proving theorems about neural networks, but in understanding the empirical properties of neural networks. They seek to understand "why" questions about generalisation: *why does the network generalise in this specific way?* or *what are the reusable representations that a neural network is using?* and so on.

We a priori know the model's weights, architecture and formal specification and we can compute the output behaviour corresponding to any given input. Where other sciences might be limited by the precision of their measuring tools or the fidelity of observations, in Interpretability our observations are exactly precise and our experiments perfectly reproducible. Furthermore, we can intervene on the network at any point and any time to arbitrarily high precision. However, despite this formal knowledge and potential for intervention, understanding neural networks remains elusive. Here we find a peculiar reversal of the sciences: in natural science, we pursue mathematical formalism to describe empirically observed phenomena; yet in interpretability, we pursue empirical methods to understand formalism.

To make progress in understanding ML models, we approach them as naturalists studying a complex system. We can see neural networks as an exemplar of the pitfalls of reductionism: we know all the material and formal causes of the network, and we understand each individual part; however, we do not understand the system. We have complete access to all formal properties of the system, yet our *scientific knowledge* of the system is incomplete. The *strange* paradox of interpretability is thus: we have a complete understanding of neural networks at the base, formal, implementation level up to arbitrary precision.[3] However, it's not immediately obvious how this low-level knowledge translates into high-level understanding and the ability to predict and control the complex system's behaviour. We might say that the properties of a neural network *supervene* on low-level mathematical facts about the system. That is, the neural network is entirely defined by its low-level facts, but yet there are *emergent* mental phenomena that are not apparent purely by analysing the low-level facts.

To understand a neural network, we would like to understand the relevant variables in model's computation. These variables, which we might call *features*, are generally not neurons or network parameters; they are instead unseen entities that we must posit and discover like subatomic particles in physics.[4] Though we have perfect formal knowledge of the system, we are nonetheless explaining what we see (the network's behaviour) in terms of what we cannot immediately see (the features). In interpretability, as in other natural sciences, understanding of the seen is revealed through the unseen (Deutsch 2011; Marion 2008; Girard, Oughourlian, and Lefort 1987; Aquinas 1273).

Paper Structure. The rest of this paper is organised as follows:

- In Section 3, we provide an exposition of the *Explanatory View* of neural networks: a model's internal structures admit explanations of model behaviour. We argue that the Explanatory View provides additional justification for why MI researchers can productively seek Causal-Mechanistic explanations of ML models.
- Section 4 provides a technical definition of Mechanistic Interpretability as seeking model explanations that are *Model-level*, *Ontic*, *Causal-Mechanistic*, and *Falsifiable*.
- From this definition, we analyse the inherent limits of Mechanistic Interpretability in Section 5 and discuss implications of the Explanatory View in Section 6.
- We conclude in Section 7 by articulating the implicit conjecture at the heart of interpretability, which we call *The Principle of Explanatory Optimism* (EO). EO states that the generalising algorithms learned by neural networks are human-understandable.

3 Explanations and Interpretability

Much of the human experience, both social and personal, is made up of explanations. We explain why vegetables are good to our children, why product A is likely to sell more units than product B to our boss, why to pursue a given

[3] This would be comparable to a physicist having perfect knowledge of the fundamental particles and forces of the universe, being able to measure and manipulate each atom at will and see subatomic particles with the naked eye. Or a biologist knowing the precise reactions occurring in every cell in the body and knowing the structure and shape of every relevant molecule (though see also Jonas and Paul Kording (2016)).

[4] The Sparse Autoencoder paradigm has a specific linearly accessible interpretation of these features; in general we make no such commitment to any particular instantiation of how features are represented in the network.

research topic to ourselves, and so on. But what do we mean by the term 'explanation' in science?

3.1 Scientific Explanation

The epistemic aim of science is to understand phenomena by way of explaining these phenomena (Regt 2017). A scientific explanation, then, is an answer to a "why" question (Lipton 2001). The fundamental question that explanations answer is "why did the phenomenon occur?" (Hempel and Oppenheim 1948). We can view explanations as a solution to a problem: there's a gap between our current best theory and the phenomena that we would like to explain. Good explanations close this gap. With a good explanation, we can say that the phenomenon was indeed expected — and crucially — here's why. Explanations are vehicles for understanding; someone understands a phenomenon when they grasp an accurate explanation of the phenomenon (Strevens 2013; Khalifa 2013).

Understanding and Compression. Wilkenfeld (2019) describes the close relationship between understanding and compression.[5] Given a series of observations which characterise a phenomenon, we understand the phenomenon if we have an explanation that compresses the data into a more concise form such that we could reproduce the data from the explanation or use the explanation to predict future data. Good explanations exploit regularities in the data for compression. A series of observations is incomprehensible if it cannot be compressed, that is, if it contains no regularities to exploit and is purely random (Li, Vitányi et al. 2008). Explanations are not *merely* compressions however. It is not obvious that a compressed zip file engenders more understanding than the original data file in general. Explanations are compressions of a particular kind: those compressions that facilitate the understanding of phenomena (Ayonrinde, Pearce, and Sharkey 2024).

3.2 From Induction To Explanation

Andrews (2023) details a classical view of machine learning as a process of induction: "ML models use evidence, or training data, to form predictions or classifications, which generalise what they have learned from their training set to unseen instances (i.e., novel data). The field of ML strives to automate inductive inference." When we view ML models in this behaviourist fashion as black boxes with only inputs and outputs, we may think of them as providing predictions without explanations. Such *explanationless predictions* are of the same variety as a prophecy from an oracle.

Suppose that we are to think of an ML model qua oracle as providing reasons to believe some proposition p. Since oracles do not provide explanations in terms of the prediction's *content* (subject matter), we may only believe such an explanationless oracle-style claim if we have extrinsic reasons to believe the model's prediction (for example, that the model has been generally correct before or that we have some knowledge of the model's training or similar). For MI researchers, however, believing a proposition p from an ML model should be based on the *content* of model explanations rather than merely on extrinsic reasons like the model's track record.

Mechanistic Interpretability researchers do not view neural networks as incomprehensible inductive black boxes. In this paper, we offer a philosophical substantiation of Mechanistic Interpretability as practised by scientists. We will argue that MI researchers take an alternative Deutschian (Deutsch 2011) *Explanatory View* of neural networks, which we may contrast with the classical inductive perspective expressed by Andrews (2023, inter alia).

Where the classical view understands neural networks as black boxes that inductively infer from data, the Explanatory View would have us consider knowledge of the internal mechanisms of the model as necessary for understanding the model's behaviour. Here the internal mechanisms themselves can be seen as implicit explanations of the model's behaviour - the reason that a model makes a prediction is contained within its internal mechanisms. This is a white-box (cognitivist) view of neural networks. In this way, we can view models as *proto-explainers* rather than merely predictors.

Generalisation: What We're Explaining When We're Explaining Neural Networks When we are interested in explaining neural networks, what we would like to explain is how, and in which ways, they *generalise*. By generalise, we mean that the model can leverage regularities & structure in the training data to solve some task with respect to unseen data. As models learn to generalise, *internal structure* forms within the model.[6]

A system contains structure to the system's generating process can be expressed more concisely (i.e., in fewer bits) than the observations of the system. That is, structure is *compressibility*. Systems that follow general principles contain structure and regularities such that they are (at least in principle) more predictable than structureless systems like random noise generators. For example, consider an idealised pendulum's position over time. The pendulum's position follows a predictable pattern which can be expressed as a mathematical function and hence we only need to store the equation and the initial conditions to reproduce the observations of the pendulum's position over time.

In this sense, the natural world contains structural patterns — there are natural laws which allow us to compress and understand the world. And the training data for ML models, which is sampled from the world, inherits such structure from the natural world. ML models generalise to the extent that they can learn or approximate the structure in the world through the training data.[7] A good explanation should expose the model's internal learned structures. We define

[5] See also Chaitin (2002); Li, Vitányi et al. (2008); MacKay (2003); Solomonoff (1964); Hutter, Catt, and Quarel (2024).

[6] Here we are interested in explanations of neural networks as objects of scientific and philosophical interest. Much previous work has been interested in explanations of the results of neural networks as it pertains to some task (e.g., medical diagnosis). Here the application domain is of secondary interest and we are examining explanations *of neural networks themselves.*

[7] Models trained on randomised data may memorise but never learn to generalise (Zhang et al. 2017; Lehalleur et al. 2025; Lin, Tegmark, and Rolnick 2017; Deletang et al. 2024).

ur-explanations[8] as the idealised explanations of model behaviour on an input distribution, given in terms of its learned internal structures.[9]

The *ur-explanations* of a neural network can be seen as *internal computations* over *learned representations* that compute the output from the input. The network learns these representations during training in a process of automated Conceptual Engineering. These network computations, outputs, and intermediate activations together constitute not only a prediction of some answer, but also an explanation of the process by which the model came to such a result.

Explanatory Faithfulness The Explanatory View takes seriously the idea that there is structure in the model to be interpreted. Under this view, there is a target to the interpretability program: we are not merely looking for explanations that appear to correlate with model behaviour, we are looking to extract the internal explanations from neural networks. Explanations can be more than confabulatory just-so stories that provide the illusion of understanding the model's behaviour.

Under the Explanatory View, we can now define *explanatory faithfulness*. An explanation is explanatorily faithful to the model to the extent that it matches the model's ur-explanation.[10] Interpretability researchers would like to say that their explanations are faithful to the model and (approximately) describe the same algorithmic mechanisms that the model uses. Note the difference here between our notion of *explanatory faithfulness* and *(behavioural) faithfulness* (e.g., from Wang et al. (2023)): *behavioural faithfulness* says that the explanation and model produce the same outputs; *explanatory faithfulness* says that the step-by-step explanation matches the model's internal mechanisms, not just the input-output behaviour. Note that defining explanatory faithfulness is not possible under the classical view of Machine Learning without ur-explanations. Under the classical view, statements about circuit equivalence can only be understood as statements about behavioural statistics (Shi et al. 2024).

Explanatory Faithfulness

An explanation E is **explanatorily faithful** to a model M over some data distribution D, to the extent that intermediate activations $\mathbf{s}_i$ at each layer i that are given by the algorithmic explanation E closely match the intermediate activations $\mathbf{x}_i$ of the model M for input data in D.

[8]Here we use the *ur-* prefix to indicate primacy or origin as in the German language.

[9]Considering the model internals as ur-explanations does not necessarily mean that all such ur-explanations will be interesting explanations of generalisation per se. Models may have memorised certain answers or resort to bags of faulty heuristics in some cases. However, in at least some cases, we have evidence of models learning genuine algorithms which represent explanatory knowledge within the network (Wang et al. 2024; Nanda et al. 2023; Wu et al. 2024).

[10]We argue for the uniqueness of the ur-explanation in **??**.

3.3 Neural Networks Perform Computations over Representations

We described the ur-explanation of a model (the idealised explanation of model behaviour) in terms of *Computations* over *Representations*. We now provide more details on what is meant by each of these terms.

Computation. Marr (1982) describes 3 levels of analysis for understanding a machine carrying out an information-processing task (McClamrock 1990; Angelou 2025):

1. **Computational Level:** What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
2. **Algorithmic/Representational Level:** What is the algorithm being used to perform the computation? How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
3. **Implementation Level:** What is the physical implementation of the algorithm? How can the representation and algorithm be realized physically on some computational substrate?

For neural networks, the *Implementation* Level of Analysis corresponds to matrix multiplications with the model weights and how these are implemented on the substrate of hardware computational accelerators (Angelou 2025). We know essentially all there is to know formally about the Implementation level.

When we speak of a neural network carrying out computations, we are referring to the *Algorithmic* and *Computational* Levels of Analysis. We would like to have useful compressive causal explanations at the Algorithmic and Computational levels which are detailed enough to be "runnable" (Cao and Yamins 2024).[11] Explanatory Faithfulness is an Algorithmic level property: the stages of the explanation should match the model layers and be "locatable". It is not sufficient for Explanatory Faithfulness for the outputs to agree if algorithms producing the outputs differs.

Representation. A pattern of neural activations is a *representation* when it represents *something*, that is it has some appropriate correspondence with features of the input data (and hence the external world). Representations are representations *of* a feature.[12] We paraphrase Harding (2023)'s three criteria for activations to qualify as representations below.

Consider a pattern of activations $h(\mathbf{x})$ for $\mathbf{x} \in X$, where X is the domain of a model (e.g. natural language). Then $h(\mathbf{x})$ *represents* a property Z if the following three criteria hold:

[11]By "runnable" here we mean that the explanation that we provide should be pseudocode that we could imagine formalising such that it would compile and run on a computer.

[12]In other words, representations have *intentionality*. Note that we use the word intentionality here in the philosophical sense of "aboutness". This usage of the term is not to be confused with the psychological sense of "intention" (as in I intend to get the next train home) or any claims about some conscious relationship to representations.

- **Information:** The activations $h(\mathbf{x})$ correlate with the property Z. More formally, the random variable $h(\mathbf{x})$ has sufficiently high Shannon Mutual Information with the property Z, $I(h(\mathbf{x}); Z)$, such that we could train a successful probe $g_z : h(\mathbf{x}) \to \mathcal{P}(Z)$. Intuitively, Information says *representations are **Causal Results** of features contained within the input* $\mathbf{x}$.
- **Use:** The model uses the information in activations $h(\mathbf{x})$ about Z to perform its task. That is to say that if we were to remove the relevant information from the activations through a causal intervention, the model's performance on the relevant downstream tasks would decrease. Intuitively, Use says *representations are **Causes** of the model's behaviour*.
- **Misrepresentation:** It should be possible for the activation vector $h(\mathbf{x})$ to misrepresent Z. Suppose that we have activations $h(\mathbf{x})$ which do not contain useful information about Z and $h(\mathbf{s})$ which does contain information about the property Z. Then we say that Z is misrepresentable if we can perform an intervention which patches the information $h(\mathbf{s})$ into $h(\mathbf{x})$, and predictably increase the likelihood of our model mistaking our input $\mathbf{x}$ for having property Z. Intuitively, Misrepresentation says *representations can be **causally intervened** on*.[13] To be able to represent, you must be able to misrepresent.

A pattern of neural activations that satisfies the Information, Use and Misrepresentation criteria can be called a *representation*.

3.4 The Goal of Interpretability

ML methods are often applied to problems in other fields like predicting weather patterns, classifying legal cases and allocating scarce resources. Interpretability, then, can be viewed as applying ML methods and analysis to an epistemic problem: "how does a neural network perform computations over representations to produce useful answers to queries?"

Interpretability researchers don't want to only know what a neural network predicts. We would also like to understand the structures, features, regularities, and knowledge which cause the neural network to make such and such a prediction. We would like to extract *explanatory knowledge* from neural networks, *uncovering* the ur-explanations that are always-already present within a trained, generalising model. Most ML researchers are in the prediction business. Interpretability researchers, however, are in the explanation business.

The Explanatory View treats neural networks as containing explanations rather than as being purely behaviourist oracles, moving from black-box induction to white-box computations. The Explanatory View is the first step in understanding ML models not in terms of prediction but in terms of *explanation*.

4 Demarcating Mechanistic Interpretability

There has been much discussion about what makes some interpretability research 'Mechanistic' rather than another form of interpretability (Saphra and Wiegreffe 2024; Chalmers 2025). Gieryn (1999) describes the problem of demarcating where one science starts and another begins —'boundary-work'— analogously to the Demarcation Problem between Science and Pseudo-Science (Laudan 1983; Popper 1935). The definition of a given science can be seen as a grab-bag of associations like Wittgensteinian language games (Wittgenstein 1953). Another way to define a science is as (social) culture (Latour 1987). Under this view "Science is what scientists do" (Bridgman 1980).

To formalise Mechanistic Interpretability, we instead provide a technical definition that focuses on the goal of Mechanistic Interpretability compared to other adjacent disciplines (Olah et al. 2020; Saphra and Wiegreffe 2024). We define Mechanistic Interpretability as the study of *Model-level, Ontic, Causal-Mechanistic, and Falsifiable Explanations* of Neural Networks. This definition clearly delineates Mechanistic Interpretability from other paradigms like Concept-Based Interpretability.[14] We can understand these properties of explanations by way of contrast with other forms of explanation.

Model-level Explanations. An autoregressive language *model* is a neural network that returns a probability distribution over possible next tokens when conditioned on some input tokens (i.e., textual prompt). A language model *system* (LM system) is a software object that contains a language model as part of the control flow. The LM system leverages the language model to produce some useful output, like a text completion or an image, rather than a single next-token probability distribution. An LM system may be as simple as augmenting a language model with a sampling method or greedy decoding. More complex LM systems may use meta-decoding strategies, tool use, automated prompting, multiple language models, and more (Arditi 2024; Zaharia et al. 2024; Khattab et al. 2023; Guo et al. 2024; Dafoe et al. 2020; Welleck et al. 2024).

Capability evaluations are typically system-level evaluations. Model performance depends substantially on prompting strategies like Chain of Thought reasoning (Wei et al. 2022) and tool use (Schick et al. 2023). Systems-level explanations might seek to explain whole system performance by, for example, reading a model's Chain of Thought (Perez et al. 2023). Conversely, Model-level explanations seek to understand the neural network part of the system in isolation to explain why the output distribution is as it is.

Ontic Explanations. Ontic explanations consist of real, physical entities (Salmon 1984). We may contrast Ontic explanations with epistemic explanations which focus on making phenomena understandable or predictable to the interpreter, potentially using idealizations, models, or abstractions that may not directly correspond to reality. Non-ontic, "epistemic" explanations may give useful rules of thumb for prediction or intuition but may not be well supported by reality.[15] Varma et al. (2023)'s work on circuit efficiency can be seen

[13] in the sense of Causal Abstractions Theory (Geiger, Potts, and Icard 2023; Pearl 2009; Beckers and Halpern 2019)

[14] This delineation is useful for researchers but we do not intend to imply that non-Mechanistic Interpretability is not useful.

[15] Note that scientific non-realists may only produce epistemic explanations, as they may not believe that the entities referred to in scientific theories actually exist in reality.

as giving non-ontic explanations, through the hypothesised efficiency metric.

Causal-Mechanistic Explanations. Causal-Mechanistic Explanations (Woodward 2003; Salmon 1989; Lewis 1986) identify the causal processes that produce phenomena rather than just describing statistical correlations or general laws. Here, we are interested in the relevant components of a system, how they are organised, and how they interact to produce phenomena. Causal-Mechanistic theorists refer to these explanations as "explaining why by explaining how" (Bechtel and Abrahamsen 2005): explaining why a phenomenon occurred involves identifying the underlying mechanisms that give rise to observed phenomena. Causal-Mechanistic Explanations go step by step to explain the end-to-end process: they provide a continuous causal chain from cause to effect, without any unexplained gaps. Causal-Mechanistic Explanations explain the end-to-end process (Salmon 1984; Lipton 2003).[16]

We can contrast Causal-Mechanistic Explanations with:

- **Statistically Relevant Explanations** (Salmon, Jeffrey, and Greeno 1971; Salmon 1989). X explains Y if and only if $P(Y|X) \neq P(Y)$ — that is, if and only if the conditional probability of Y given X differs from the probability of Y. For example, we might explain ice cream sales by high correlation with temperature.[17]
- **Telic Explanations** (Sosa 2021). We explain a phenomenon by reference to its purpose, aims, or function rather than in terms of a causal chain of events. For example, we might explain the heart as being for the purpose of transmitting and pumping blood.
- **Nomological Explanations** (Myers 2012; Scheibe 2002). We explain a phenomenon by reference to general laws or principles rather than in terms of a causal chain of events. For example, linguistic theory might appeal to universal grammar "laws" to explain the structure of human languages.

Definition of Mechanistic Interpretability

Interpretability explanations are **valid** as Mechanistic Interpretability explanations if they are **Model-level**, **Ontic**, **Causal-Mechanistic** and **Falsifiable**.

Causal-Mechanistic and Ontic explanations are necessary to the empirical practice of Mechanistic Interpretability amongst active researchers (Bereska and Gavves 2024; Sharkey et al. 2025). Though it is feasible to imagine a causal-mechanistic approach to understand system-level behaviours, it is a historically contingent fact that the field has coalesced around methods for model-level explanations (Saphra and Wiegreffe 2024).[18]

[16]Several works within Machine Learning that also provide a good introduction to causal modelling include Jin and Garrido (2024); Schölkopf et al. (2021); Liu et al. (2024); Pearl (2009).

[17]or unhelpfully, our explanation could show correlation with the number of shark attacks

[18]Since system-level explanations are of practical and academic interest, there is currently an opportunity for a new field to emerge focusing on system-level explanations, possibly building on the work of the mechanistic interpretability community.

Perhaps the nascent field of LLM-ology might be a candidate to fill this gap (Trott 2023). Note that many Benchmarks and Evaluations researchers could be seen as working in this space already. Chain of Thought interpretability (Perez et al. 2023) is another example of system-level explanation.

5 The Limits of Mechanistic Interpretability

We have analysed the type of explanations that Mechanistic Interpretability researchers seek, namely those which are Model-level, Ontic, Causal-Mechanistic and Falsifiable Explanations of a model's internal mechanisms. We now turn our attention to the extent of the limits and challenges of such explanations.

5.1 Value-Ladenness & Theory-Ladenness of Explanations

We would like explanations that are accurate and human-understandable compressed representations of observations. With this goal in mind, the best explanation of a phenomenon is interpreter-relative in the following two senses. *Firstly*, the ideal explanation is relative to the interpreter's initial set of concepts, their priors and what types of explanation are easy for them to understand. In this sense, explanations are **Theory-Laden**. *Secondly*, the ideal explanation depends on what the interpreter would like to *do* with such an explanation. The interpreter may be satisfied with a different level of granularity of explanation depending on whether they are seeking an explanation of a model's behaviour to be able to make crude interventions, or to make guarantees about model performance, or for scientific curiosity. In other words, the ideal explanation is also relative to the interpreter's *values*; explanations are **Value-Laden**.

Value-Ladenness of Explanations Weber (1949) argued for the Value-Free Ideal in the sciences, the principle that scientists should be value neutral. We can articulate the Value-Free Ideal as "*Scientists should strive to minimize the influence of contextual values on scientific reasoning, e.g., in gathering evidence and assessing/accepting scientific theories*" (Reiss and Sprenger 2020).

For the normative statement of the Value-Free Ideal to be considered a reasonable ideal, it must be attainable (at least to some degree). That is, ought implies can. So we may first analyse whether value-freeness is possible which can be expressed in the Value-Neutrality Thesis as follows: "*Scientists can—at least in principle—gather evidence and assess/accept theories without making contextual value judgments*" (Reiss and Sprenger 2020).

In science generally, and interpretability research particularly, it is difficult to hold Value-Neutrality. Hence the Value-Free Ideal seems to be unattainable and likely undesirable as a goal (Douglas 2009). The choice of methods and which results are particularly interesting for researchers has a close dependence on what researchers might hope to achieve. Many researchers in Mechanistic Interpretability are interested in the applications to AI Safety, AI Ethics, AI Cognitive Science and AI Governance (Bengio et al. 2025; Anwar et al. 2024;

Olah et al. 2020), all of which affect researchers' contextual value judgements. Evidential standards for accepting theories are highly influenced by such application-guided values.

Given the increasing importance of AI systems in society and their potential benefits and harms, it is perhaps more instructive to understand (the lack of) value-freeness in Mechanistic Interpretability as we would in Climate Science or Public Health, rather than in Theoretical Physics.[19] Researchers must share reproducible, quantitative results for the community to assess but it is unavoidable that the choice of study and what counts as sufficiently convincing to such and such a conclusion is highly value-laden (Sharkey et al. 2025; Casper, Krueger, and Hadfield-Menell 2025).

Expressing an interpretability-flavoured notion of Value-Ladenness, Dmitry's Koan (Vaintrob 2025) states:

> *There is no such thing as interpreting a neural network. There is only interpreting a neural network at a given scale of precision.*

We can view Dmitry's Koan as a direct consequence of the Value-Ladenness of explanations in Mechanistic Interpretability: what counts as a good explanation is highly dependent on the level of precision that the interpreter desires. Explanations at a maximal precision would capture lots of noise as well as useful explanatory signal.

For some use cases (e.g., determining the safety of critical AI systems), higher fidelity explanations may be required for providing guarantees about model behaviour and so we would like highly precise explanations. In other cases, the interpreter may seek sufficient understanding with which to monitor or steer the model, which might be possible with a lower fidelity explanation.[20] The ideal precision depends on the (human) interpreter's goals and hence the ideal explanation is inherently value-laden.

Noting that we do not always have an appropriate definition of what 'precision' itself means in our interpreter's context, we can extend Dmitry's Koan to Dmitry's Koan^{++} [21] stating:

> There is no such thing as interpreting a neural network. There is only interpreting a neural network at a given scale of precision *and a given metric for defining what precision means.*

Dmitry's Koan^{++} highlights that the choice of precision metric itself is value-laden.

Theory-Ladenness of Explanations We might hope to understand ML systems on their own terms, in their ontology, removing all human "biases" from the explanation. After all, if we have enough data, perhaps the data will speak for itself, we might think. This (unfortunate) desire is known as the Theory-Free Ideal (Andrews 2023).

Our explanations always contain underlying (human) theory. Indeed, so-called "unsupervised" learning cannot occur without either pre-defined inductive biases or external supervision (Andrews 2023; Wolpert and Macready 1997; Goldblum et al. 2023; Locatello et al. 2019). All observations (and interpretations) are theory-laden (Kuhn 1962; Duhem 1954; Popper 1935). Interpreter theory seeps into the explanation in all stages of an interpretability workflow: from problem formulation and model design to model selection and semantic interpretation.

Example: Theory-Ladenness of Sparse Autoencoder Explanations Sparse Autoencoders (SAEs) are a method for unsupervised interpretability, aiming to extract concepts (or "features") from the activation space of neural networks. Concept representations are generally entangled and difficult to access in the neural activation space. We may hope for SAEs to disentangle these representations into a linear combination of monosemantic concepts by mapping the neural activations to a feature basis. Empirically, it has been shown that disentangled concept representations are more amenable to human interpretation (Bricken et al. 2023).

We might believe that we are doing completely unsupervised learning with no human theory when producing SAE derived explanations of neural activations. However, note that in using the SAE, we are committing to the theory that features are sparsely activated and linearly represented (Bricken et al. 2023). Similarly, in choosing a particular SAE architecture like TopK (Gao et al. 2024) or Jump-ReLU SAEs (Rajamanoharan et al. 2024), we are committing to the Monotonic Importance Heuristic (Ayonrinde 2024), the conjecture that feature activations are not typically both small and important simultaneously. Hindupur et al. (2025) further describe the theoretical assumptions that come with different choices of SAE architecture. Following Locatello et al. (2019), we note that unsupervised disentanglement learning in the general case is not possible; we must first hold some theoretical commitment to the structure of the data. We choose theoretical commitments because we have reason to believe that they are good inductive priors for the data distribution, or because we believe that the structures will be more easily human-understandable.

Interpretability is not and cannot be a purely engineering affair, devoid of theory. We require theory for two reasons: Firstly, as we have seen with SAEs and disentanglement learning, holding the wrong theoretical commitments[22] leads to the intractability of unsupervised learning.[23] And secondly, the human (interpreter's) priors are a key part of the theory, as indeed it's humans that we would like to make interpretations for! In this sense, Interpretability is a fundamentally socio-technical problem which may be best addressed by a combination of understanding humans, machines and the interactions between the two. Hence we see a key role for Human-Computer Interaction (HCI) and the Social Sciences

[19] Some philosophers have further argued that the Value-Free Ideal is not even tenable or desirable in Theoretical Physics either.

[20] Sharkey (2024) provide a careful analysis of the trade-offs between explanation precision and complexity. We also note the close analogy to rate-distortion theory in Information Theory.

[21] We may also refer to Dmitry's Koan^{++} as Nora's Koan, after Interpretability Researcher Nora Belrose who brought this to our attention.

[22] which is very easy if researchers believe that they are holding no theoretical commitments at all!

[23] Since interpretability aims to understand neural systems which we do not yet understand, it is inherently an unsupervised learning problem: we don't know what we don't know about the system.

in MI. We suggest that increased focus on the criteria which make explanations accessible to humans (Schut et al. 2023; Ayonrinde, Pearce, and Sharkey 2024), especially to diverse humans (Himmelsbach et al. 2019), is likely to prove fruitful for future interpretability work.[24]

5.2 Limits of Model-level vs System-level Explanations

As detailed in Section 4, explanations in Mechanistic Interpretability are inherently Model-level explanations. However, when interacting with AI, we are typically interacting with AI *systems*, not *models*. Though models may think, only systems behave: it is systems that perform actions in the world. MI explanations, then, have limited explanatory power when the system-model relationship is complex or not well-understood. In systems with meta-decoding processes such as those with inference-time compute loops, ensembling methods, or similar, then we might expect model-level explanations to be insufficient for understanding system-level behaviour.

Systems which can well be described as having an Extended Mind, in the sense of Clark and Chalmers (1998) or as embodied/embedded agents (Demski and Garrabrant 2018; Shapiro and Spaulding 2021) may also be difficult to understand with model-level explanations. It may be difficult to pick out some feature if the feature as it appears in the model is simply a pointer to some cognitive process distributed elsewhere in the system. A particularly notable case of such systems is multi-agent AI systems, which may have emergent properties not well explained by the analysis of individual agents. For example, consider a flock of birds or a well-functioning marketplace which may not be easily understood by the analysis of individual agents within the system (Hyland et al. 2024).

5.3 Limits of Low Abstraction Explanations

One other possible, though surmountable, limit to Mechanistic Interpretability explanations is that low-level explanations may be difficult, or even impossible, to turn into explanations at higher levels of abstraction.[25] For example, in the natural sciences, it is not generally known whether the laws of thermodynamics can be derived from lower-level particle physics laws. If MI provides low-level explanations akin to quantum mechanics, research questions that more closely resemble chemistry, biology or, social science questions may not be obviously derivable in a reductionist way from MI explanations.

[24]Theory-ladenness of explanations is also relevant in the context of the *Construct Validity problem* (Cronbach and Meehl 1955) - the problem of whether the explanation is measuring what it purports to measure. (As Heisenberg (1958, p.58) put it: "what we observe is not nature in itself but nature exposed to our method of questioning.") It is very possible for researchers to agree on the data reported and the statistical validity of hypothesis tests and yet disagree on the interpretation because their underlying theories are different.

[25]Note that by levels of abstraction here we do not mean the semantic abstraction level of a given feature, where some features might represent more higher-level concepts than others (for example those at later layers of a model). We instead mean that features themselves are low-level units compared to higher-level abstractions like components or circuits which they may compose.

6 Discussion

Given a data distribution $\mathcal{D}$, suppose that we would like to explain the behaviour of a neural network M over a subset of the data distribution $D \subset \mathcal{D}$. Then the goal of Mechanistic Interpretability is to provide a Model-level, Ontic, Causal-Mechanistic, Falsifiable explanation E of the model's behaviour on D which is explanatorily faithful to M. Where behavioural faithfulness requires that the end predictions of the models agree, *explanatory faithfulness* is a stronger condition that requires the causal structure of the explanation E to be faithful to the causal structure of M at each stage of the causal chain.

In this work, we have argued that approaching the Problem of Interpretability through the Explanatory View of Mechanistic Interpretability is likely to be fruitful because neural networks naturally admit explanations of their behaviour through their internal structures, where structure corresponds to compressibility. Hence, our explanatory methods can and should look to uncover causal structure in the model M rather than merely producing confabulatory descriptions of model behaviour. The Explanatory View of Neural Networks provides a justification for using explanatory faithfulness as the goal of explanations in Mechanistic Interpretability.

However, there are significant limitations to this approach. In particular, it is infeasible to have a general algorithm for finding explanations E for all models M and all data distributions D which are optimal for all purposes. Solutions to the interpretability problem are Theory-Laden: they require some theoretical priors about neural networks and/or the data distribution to find good explanations. Similarly, the problem of finding good explanations is Value-Laden: what makes for a good explanation depends on our goals as interpreters.

A core problem to address in future work is how to appropriately characterise what makes a *good explanation* in the context of Mechanistic Interpretability. Here, we have argued for necessary criteria for an explanation to be validly considered 'Mechanistic' (namely that it is Model-level, Ontic, Causal-Mechanistic and Falsifiable). We would further like to understand how some explanations are better than others in terms of their usefulness and likelihood to point towards truth.

7 Coda: Explanatory Optimism

> *"Have you persuaded yourself that there are knowledges and truths beyond your grasp, things that you simply cannot learn? ... If you have allowed this to happen, you have arbitrarily imposed limits on your intellectual freedom, and you have smothered the fires from which all other freedoms arise."*
>
> — Scott Buchanan, 1958

We conclude by introducing a conjecture at the heart of (Mechanistic) Interpretability, that we call **The Principle of Explanatory Optimism**. We would like to explicate this conjecture, argue for its importance for MI research, and raise a Call to Action for further research into clarifying both the

statement and its veracity. Here we provide no arguments for the truth of Explanatory Optimism (EO) as a conjecture; we leave further arguments, proofs, or refutations to future work.

7.1 Alien Concepts

Suppose that an AI M (Machine) and the interpreter H (Human) each have some set of concepts which are understandable to them, C_M and C_H respectively (Schut et al. 2023; Hewitt, Geirhos, and Kim 2025). If $C_M \subset C_H$, then intuitively all concepts that the machine uses for its computations can be immediately understood by the human interpreter. However, if $C_M \setminus C_H$ is large, then there are Machine-concepts that are not natively Human-understandable.

Some Machine-concepts that aren't intuitively understandable by humans may be human-understandable with some human effort, effective translation, or good explanations. However, a core problem remains if there are concepts that are understandable to the model but which are fundamentally alien and incomprehensible to humans. We call such concepts **Alien Concepts** and label these C_A. Alien concepts are Machine-concepts in $C_M \setminus C_H$ that are effectively untranslatable into Human-concept terms (see Figure 1).

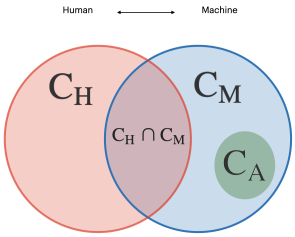

Figure 1: A Venn diagram showing the relationship between the concept spaces of the machine M and human interpreter H. The machine and human have some shared concepts which they can use to communicate ($C_M \cap C_H$) but there are many concepts that the machine uses that the human does not understand ($C_M \setminus C_H$). The set of *Alien Concepts* $C_A \subset (C_M \setminus C_H)$, is a subset of the Machine-concepts. Alien Concepts are causally relevant for the model's computation but are fundamentally incomprehensible to humans. If this set is large or important, then Interpretability may be highly limited.

To the extent that Alien Concepts are present in the model, and are important for the model's computation, the project of Interpretability may be fundamentally flawed. The conjecture at the heart of (Mechanistic) Interpretability then is that artificial neural network-based general intelligences have few, or no, alien concepts that are both vital to model behaviour and not human-understandable. A version of this conjecture appears to be a prerequisite for interpretability research: if most of the model's concepts aren't understandable to MI researchers, then it is not clear how MI research can proceed.[26]

[26]Conversely, if model concepts *are* possible to be made understandable by MI researchers, then MI research is a tenable research direction (though it may still be difficult).

7.2 The Principle of Explanatory Optimism

As phrased above, the "no alien concepts" conjecture is about neural networks. It may be more instructive, however, to rephrase this conjecture to put human intelligence at the center and understand it as a claim about a class of general intelligence (such as neural-network based AI systems). Rephrasing then: *Everything that is important for the behaviour of an intelligence with implicit explanatory knowledge within some explanatory complexity class is human-understandable.* This statement is a strong form of the conjecture that we call **The Principle of Explanatory Optimism**.[27] Weaker forms of Explanatory Optimism might claim that "most" rather than all neural network behaviour is human-understandable (in terms of variance explained for example).

Humans have certain cognitive limitations compared to future generally intelligent systems: we have limited memory and processing power, we are somewhat limited in bandwidth and speed, and we may lack attention. Strong Explanatory Optimism suggests that, given sufficient time, we could understand these artificial intelligences, *if* augmented with good, concise explanations, memory devices, cognitive tools and such like. The Explanatory Optimism conjecture implies that explanations for understanding Machine-Concepts exist.

The Principle of Explanatory Optimism

The Strong Principle of Explanatory Optimism (SEO):

Everything important for the behaviour of an intelligence with implicit explanatory knowledge within some explanatory complexity class is human-understandable.

The Weak Principle of Explanatory Optimism (WEO):

Most important behaviour of an intelligence with implicit explanatory knowledge within some explanatory complexity class is human-understandable.

Explanatory Optimism (EO) can also be understood, as in Deutsch (2011), as a view of explanatory universality, defined analogously to computational universality as given in the Church-Turing thesis (Turing 1936; Church 1936). In the same sense that any Turing machine can simulate any other Turing machine, we would like a theory that maintains that some intelligences are explanatorily universal in the sense that they can understand and explain any other intelligence of an equivalent explanatory complexity class. We leave the question of what such an explanatory complexity class might look like and how to prove explanatory class equivalence and universality to future work.

The truth value of Weak Explanatory Optimism is a load-bearing question for the field of interpretability: if models aren't human-understandable, then the field will face unassailable roadblocks in its mission to explain neural networks

[27]This idea is closely analogous to Deutsch (2011)'s theory of Optimism.

to humans. Hence, an appropriate disproof of, or otherwise sufficiently convincing arguments against, (W)EO should motivate researchers working on Mechanistic Interpretability to consider reorienting their research focus.[28]

Call to Action for Explanatory Optimism

The Call To Action for future work is twofold:

- Firstly, to formalise the above conjectures (SEO and WEO). We would like to develop core definitions for the explanatory complexity classes and the appropriate notion of explanatory universality. This formalisation would likely involve understanding the explanatory complexity classes of different intelligences as well as operationalising the quantitative notion of understanding "most" of a model's behaviour.
- Secondly, to prove the formalised conjectures. We would like to assess the truth value of the Principle of Explanatory Optimism. We believe that such results would be of great interest, both to interpretability researchers and scientists who expect to use non-transparent computational models.

Complexity theorists, theoretical computer scientists, analytic philosophers, and computational mechanics theorists may be able to use the tools from their fields to make progress on this conjecture.[29]

7.3 The Importance of Explanatory Optimism

As ML models begin to do more cognitive work in the world, the frontiers of knowledge in mathematics, the (natural, social, and computational) sciences, and the humanities may be known to machines before humans.[30] The default state of the world when living alongside such cognitively advanced machines is that humans are *epistemically disempowered* and subjected to living in a world built by knowledge that no human understands. However, Explanatory Optimism offers an alternative future for humanity. The upshot of Explanatory Optimism is that as machines learn more about the world, through interpretability, humans can learn more about the world too. Any Machine Knowledge can become Human Knowledge.

[28]MI researchers with a downstream goal of AI Safety, AI Ethics, or AI Cognitive Science may be interested in whether EO holds. Absent EO, MI may not be an effective way to reach their goals.

[29]Explanatory Optimism is another area where we might expect fruitful collaboration between Philosophy and Computational Complexity (and Machine Learning) as in Aaronson (2013).

[30]Arguably AI automation of science may be the final stage of the long-continuing Crisis of the European Sciences depicted by Husserl and Carr (1970). For Husserl, the fact that the sciences have become so disconnected from the phenomenological world of everyday experience results in scientific disciplines becoming increasingly specialised such that no human has a unified understanding of science as a whole. With sufficiently intelligent AI, we can imagine a world where no human has an understanding of the furthest advances in *even a single scientific discipline*.

Hence, if Explanatory Optimism is true, Interpretability may be one of the most important projects in the history of modern science. Explanatory Optimism implies that *all explanatory knowledge is accessible to people* through interpretability and human-computer interaction: we are sitting at but the beginning of an explosion in human understanding.

Acknowledgments

Thanks to Nora Belrose, Matthew Farr, Sean Trott, Elsie Jang, Evžen Wybitul, Andy Artiti, Owen Parsons, Kristaps Kallaste and Egg Syntax for comments on early drafts. We appreciate Daniel Filan and Joseph Miller's helpful feedback. Thanks to Mel Andrews, Alexander Gietelink Oldenziel, Jacob Pfau, Michael Pearce, Catherine Fist, Lee Sharkey, Jesse Hoogland, Jason Gross, Joseph Bloom, Nick Shea, Barnaby Crook, Eleni Angelou, Dashiell Stander and attendees of the ICML2024 MechInterp Social for useful conversations. We're grateful to Kwamina Orleans-Pobee for additional support. We also thank our anonymous reviewers for their thoughtful feedback. This project was supported by a Foresight Institute AI Safety Grant.

References

Aaronson, S. 2013. Why Philosophers Should Care About Computational Complexity. *Computability: Turing, Gödel, Church, and Beyond*, 261–328.

Amodei, D. 2025. The urgency of interpretability.

Andrews, M. 2023. The Devil in the Data: Machine Learning & the Theory-Free Ideal.

Angelou, E. 2025. Three levels for large language model cognition.

Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Aquinas, T. D. 1273. *Summa Theologica*. Hayes Barton Press.

Arditi, A. 2024. AI as systems, not just models.

Ayonrinde, K. 2024. Adaptive Sparse Allocation with Mutual Choice and Feature Choice Sparse Autoencoders. arXiv:2411.02124.

Ayonrinde, K.; Pearce, M. T.; and Sharkey, L. 2024. Interpretability as Compression: Reconsidering SAE Explanations of Neural Activations with MDL-SAEs. arXiv:2410.11179.

Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074.

Bechtel, W.; and 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.

Beckers, S.; and Halpern, J. Y. 2019. Abstracting Causal Models. In *Proceedings of the 33Rd Aaai Conference on Artificial Intelligence*, 2678–2685.

Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; et al. 2025. International AI Safety Report. *arXiv preprint arXiv:2501.17805*.

Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety – A Review. ArXiv:2404.14082 [cs].

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.; McLean, B.; Burke, J. E.; Hume, T.; Carter, S.; Henighan, T.; and Olah, C. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.

Bridgman, P. W. 1980. *Reflections of a Physicist*. New York: Arno Press.

Brown, N.; and Sandholm, T. 2019. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890.

Cao, R.; and Yamins, D. 2024. Explanatory models in neuroscience, Part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 101244.

Casper, S.; Krueger, D.; and Hadfield-Menell, D. 2025. Pitfalls of Evidence-Based AI Policy. *arXiv preprint arXiv:2502.09618*.

Chaitin, G. 2002. *On the intelligibility of the universe and the notions of simplicity, complexity, and irreducibility*. na.

Chalmers, D. J. 2025. Propositional interpretability in artificial intelligence. *arXiv preprint arXiv:2501.15740*.

Church, A. 1936. An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, 58(2): 345–363.

Clark, A.; and Chalmers, D. J. 1998. The Extended Mind. *Analysis*, 58(1): 7–19.

Cronbach, L. J.; and Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4): 281–302.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*.

Deletang, G.; Ruoss, A.; Duquenne, P.-A.; Catt, E.; Genewein, T.; Mattern, C.; Grau-Moya, J.; Wenliang, L. K.; Aitchison, M.; Orseau, L.; Hutter, M.; and Veness, J. 2024. Language Modeling Is Compression. In *The Twelfth International Conference on Learning Representations*.

Demski, A.; and Garrabrant, S. 2018. Embedded agents.

Deutsch, D. 2011. *The beginning of infinity: Explanations that transform the world*. penguin uK.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Douglas, H. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Duhem, P. M. M. 1954. *The Aim and Structure of Physical Theory*. Princeton,: Princeton University Press.

Erasmus, A.; Brunet, T. D.; and Fisher, E. 2021. What is interpretability? *Philosophy & Technology*, 34(4): 833–862.

Fleisher, W. 2022. Understanding, Idealization, and Explainable AI. *Episteme*, 19(4): 534–560.

Gao, L.; la Tour, T. D.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024. Scaling and evaluating sparse autoencoders. ArXiv:2406.04093 [cs] version: 1.

Geiger, A.; Potts, C.; and Icard, T. 2023. Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*.

Gieryn, T. F. 1999. *Cultural Boundaries of Science: Credibility on the Line*. University of Chicago Press.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. IEEE.

Girard, R.; Oughourlian, J.; and Lefort, G. 1987. *Things Hidden Since the Foundation of the World*. Stanford University Press.

Glazer, E.; Erdil, E.; Besiroglu, T.; Chicharro, D.; Chen, E.; Gunning, A.; Olsson, C. F.; Denain, J.-S.; Ho, A.; Santos, E. d. O.; et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.

Goldblum, M.; Finzi, M.; Rowan, K.; and Wilson, A. G. 2023. The no free lunch theorem, kolmogorov complexity, and the role of inductive biases in machine learning. *arXiv preprint arXiv:2304.05366*.

Golechha, S.; and Garriga-Alonso, A. 2025. Among Us: A Sandbox for Agentic Deception. arXiv:2504.04072.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Harding, J. 2023. Operationalising Representation in Natural Language Processing. *arXiv preprint arXiv:2306.08193*.

Heisenberg, W. 1958. Physics and Philosophy - The Revolution in Modern Science.

Hempel, C. G.; and Oppenheim, P. 1948. Studies in the Logic of Explanation. *Philosophy of Science*, 15(2): 135–175.

Hewitt, J.; Geirhos, R.; and Kim, B. 2025. We Can't Understand AI Using our Existing Vocabulary. *arXiv preprint arXiv:2502.07586*.

Himmelsbach, J.; Schwarz, S.; Gerdenitsch, C.; Wais-Zechmann, B.; Bobeth, J.; and Tscheligi, M. 2019. Do We Care About Diversity in Human Computer Interaction: A Comprehensive Content Analysis on Diversity Dimensions in Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.

Hindupur, S. S. R.; Lubana, E. S.; Fel, T.; and Ba, D. 2025. Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry. arXiv:2503.01822.

Hosseini, H.; Xiao, B.; Jaiswal, M.; and Poovendran, R. 2018. Assessing Shape Bias Property of Convolutional Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2004–20048. IEEE.

Husserl, E.; and Carr, D. 1970. *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*. Northwestern University studies in phenomenology & existential philosophy. Northwestern University Press. ISBN 9780810104587.

Hutter, M.; Catt, E.; and Quarel, D. 2024. *An Introduction to Universal Artificial Intelligence*. Chapman & Hall/CRC Artificial Intelligence and robotics series. Chapman & Hall/CRC Press. ISBN 9781003460299.

Hyland, D.; Gavenčiak, T.; Costa, L. D.; Heins, C.; Kovarik, V.; Gutierrez, J.; Wooldridge, M. J.; and Kulveit, J. 2024. Free-Energy Equilibria: Toward a Theory of Interactions Between Boundedly-Rational Agents. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

Jin, Z.; and Garrido, S. 2024. Tutorial Proposal: Causality for Large Language Models.

Jonas, E.; and Paul Kording, K. 2016. Could a Neuroscientist Understand a Microprocessor? *bioRxiv*.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Kästner, L.; and Crook, B. 2024. Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, 14(4): 52.

Khalifa, K. 2013. The Role of Explanation in Understanding. *British Journal for the Philosophy of Science*, 64(1): 161–187.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; et al. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Latour, B. 1987. *Science in action*. Cambridge, Massachusetts: Harvard University Press. ISBN 978-0-674-79291-3 and 0-674-79290-4 and 0-674-79291-2 and 978-0-674-79290-6. Literaturverzeichnis: Seite 266-270 ; Hier auch später erschienene, unveränderte Nachdrucke.

Laudan, L. 1983. The Demise of the Demarcation Problem. In Cohen, R. S.; and Laudan, L., eds., *Physics, Philosophy and Psychoanalysis: Essays in Honor of Adolf Grünbaum*, 111–127. D. Reidel.

Leavitt, M. L.; and Morcos, A. 2020. Towards falsifiable interpretability research. arXiv:2010.12016.

Lehalleur, S. P.; Hoogland, J.; Farrugia-Roberts, M.; Wei, S.; Oldenziel, A. G.; Wang, G.; Carroll, L.; and Murfet, D. 2025. You Are What You Eat–AI Alignment Requires Understanding How Data Shapes Structure and Generalisation. *arXiv preprint arXiv:2502.05475*.

Lewis, D. 1986. Causal Explanation. In Lewis, D., ed., *Philosophical Papers, Volume II*, 214–240. Oxford University Press.

Li, M.; Vitányi, P.; et al. 2008. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer.

Lin, H. W.; Tegmark, M.; and Rolnick, D. 2017. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168: 1223–1247.

Lindsay, G. W.; and Bau, D. 2023. Testing methods of neural systems understanding. *Cogn. Syst. Res.*, 82: 101156.

Lipton, P. 2001. What good is an explanation? In *Explanation: Theoretical approaches and applications*, 43–59. Springer.

Lipton, P. 2003. The Causal Model. In *Inference to the Best Explanation*, 25. Routledge, 2 edition. ISBN 9780203470855. EBook.

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.

Liu, X.; Xu, P.; Wu, J.; Yuan, J.; Yang, Y.; Zhou, Y.; Liu, F.; Guan, T.; Wang, H.; Yu, T.; et al. 2024. Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey. *arXiv preprint arXiv:2403.09606*.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124. PMLR.

MacKay, D. J. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Marion, J. 2008. *The Visible and the Revealed*. Fordham University Press. ISBN 9780823228836.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. USA: Henry Holt and Co., Inc. ISBN 0716715678.

McClamrock, R. 1990. Marr?s Three Levels: A Re-Evaluation. *Minds and Machines*, 1(2): 185–196.

Millière, R.; and Buckner, C. 2025. Interventionist Methods for Interpreting Deep Neural Networks. In Piccinini, G., ed., *Neurocognitive Foundations of Mind*. Routledge. Forthcoming.

Myers, J. 2012. Cognitive styles in two cognitive sciences. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Nanda, N.; Chan, L.; Lieberum, T.; Smith, J.; and Steinhardt, J. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3): e00024–001.

Pearl, J. 2009. *Causality*. Cambridge university press.

Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434.

Popper, K. R. 1935. *The Logic of Scientific Discovery*. London, England: Routledge.

Rajamanoharan, S.; Lieberum, T.; Sonnerat, N.; Conmy, A.; Varma, V.; Kramár, J.; and Nanda, N. 2024. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders. arXiv:2407.14435.

Regt, H. W. D. 2017. *Understanding Scientific Understanding*. New York: Oup Usa.

Reiss, J.; and Sprenger, J. 2020. Scientific Objectivity. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition.

Salmon, W. 1989. Four decades of scientific explanation.

Salmon, W. C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press. ISBN 9780691101705.

Salmon, W. C.; Jeffrey, R. C.; and Greeno, J. G. 1971. *Statistical Explanation*, 29–88. University of Pittsburgh Press. ISBN 9780822952251.

Saphra, N.; and Wiegreffe, S. 2024. Mechanistic? arXiv:2410.09087.

Scheibe, E. 2002. *Between Rationalism and Empiricism: Selected Papers in the Philosophy of Physics*. Springer Verlag.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609.

Schut, L.; Tomasev, N.; McGrath, T.; Hassabis, D.; Paquet, U.; and Kim, B. 2023. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero. *arXiv preprint arXiv:2310.16410*.

Shapiro, L.; and Spaulding, S. 2021.

Sharkey, L. 2024. Sparsify: A mechanistic interpretability research agenda.

Sharkey, L.; Chughtai, B.; Batson, J.; Lindsey, J.; Wu, J.; Bushnaq, L.; Goldowsky-Dill, N.; Heimersheim, S.; Ortega, A.; Bloom, J.; Biderman, S.; Garriga-Alonso, A.; Conmy, A.; Nanda, N.; Rumbelow, J.; Wattenberg, M.; Schoots, N.; Miller, J.; Michaud, E. J.; Casper, S.; Tegmark, M.; Saunders, W.; Bau, D.; Todd, E.; Geiger, A.; Geva, M.; Hoogland, J.; Murfet, D.; and McGrath, T. 2025. Open Problems in Mechanistic Interpretability. arXiv:2501.16496.

Shi, C.; Beltran-Velez, N.; Nazaret, A.; Zheng, C.; Garriga-Alonso, A.; Jesson, A.; Makar, M.; and Blei, D. 2024. Hypothesis Testing the Circuit Hypothesis in LLMs. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Solomonoff, R. 1964. A formal theory of inductive inference. Part II. *Information and Control*, 7(2): 224–254.

Sosa, E. 2021. *Epistemic Explanations: A Theory of Telic Normativity, and What It Explains*. Oxford: Oxford University Press.

Strevens, M. 2013. No understanding without explanation. *Studies in history and philosophy of science Part A*, 44(3): 510–515.

Trott, S. 2023. In cautious defense of LLM-ology.

Turing, A. M. 1936. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42): 230–265.

Vaintrob, D. 2025. Dmitry's Koan.

Varma, V.; Shah, R.; Kenton, Z.; Kramár, J.; and Kumar, R. 2023. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*.

Wang, B.; Yue, X.; Su, Y.; and Sun, H. 2024. Grokked Transformers are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization. *arXiv preprint arXiv:2405.15071*.

Wang, K. R.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2023. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*.

Weber, M. 1949. " Objectivity" in social science and social policy. *The methodology of the social sciences*, 49–112.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Welleck, S.; Bertsch, A.; Finlayson, M.; Schoelkopf, H.; Xie, A.; Neubig, G.; Kulikov, I.; and Harchaoui, Z. 2024. From Decoding to Meta-Generation: Inference-time Algorithms for Large Language Models. *Transactions on Machine Learning Research*.

Widdicombe, A.; Julier, S.; and Kim, B. 2018. Saliency Maps Contain Network" Fingerprints". In *ICLR 2022 Workshop on PAIR {\textasciicircum} 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*.

Wilkenfeld, D. A. 2019. Understanding as compression. *Philosophical Studies*, 176(10): 2807–2831.

Wittgenstein, L. 1953. *Philosophical Investigations*. New York, NY, USA: Wiley-Blackwell.

Wolpert, D. H.; and Macready, W. G. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1): 67–82.

Woodward, J. F. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Wu, W.; Jaburi, L.; Drori, J.; and Gross, J. 2024. Unifying and Verifying Mechanistic Interpretations: A Case Study with Group Operations. *arXiv preprint arXiv:2410.07476*.

Zaharia, M.; Khattab, O.; Chen, L.; Davis, J. Q.; Miller, H.; Potts, C.; Zou, J.; Carbin, M.; Frankle, J.; Rao, N.; and Ghodsi, A. 2024. The Shift from Models to Compound AI Systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.