The Dead Salmons of AI Interpretability

Anonymous Authors[1]

Abstract

In a striking neuroscience study, the authors placed a dead salmon in an MRI scanner and showed it images of humans in social situations. Astonishingly, standard analyses of the time reported brain regions *predictive* of social emotions. The explanation, of course, was not supernatural cognition but a cautionary tale about misapplied statistical inference. In AI interpretability, reports of similar "dead salmon" artifacts abound: feature attribution, probing, sparse auto-encoding, and even causal analyses can produce plausible-looking explanations for randomly initialized neural networks. In this work, we examine this phenomenon and argue for a pragmatic **statistical–causal reframing**: explanations of computational systems should be treated as parameters of a (statistical) model, inferred from computational traces. This perspective goes beyond simply measuring statistical variability of explanations due to finite sampling of input data; interpretability methods become statistical estimators, and findings should be tested against explicit and meaningful **alternative computational hypotheses**, with uncertainty quantified with respect to the postulated statistical model. It also highlights important theoretical issues, such as the identifiability of common interpretability queries, which we argue is critical to understand the field's susceptibility to false discoveries, poor generalizability, and high variance.

1. Introduction

In 2009, researchers placed a dead salmon in an MRI scanner, showed it photographs of humans in social situations, and ostensibly asked it to judge their emotions (Bennett et al., 2009). Standard analysis pipelines commonly used at the time surprisingly returned brain voxels as significantly predictive of emotional situations. The error arose from a failure to correct for multiple comparisons within the statistical analysis pipeline. The "dead salmon" demonstration of false positives contributed to a larger reckoning in the field of neuroscience. For instance, an influential study showed that different research groups obtained different results even when analyzing the same dataset and the same research question (Botvinik-Nezer et al., 2020). Subsequent work identified several sources of *statistical fragility*. Widely used statistical procedures embedded in standard analysis pipelines were shown to inflate false-positive rates (Eklund et al., 2016), an effect worsened by non-independent analyses producing spuriously large brain–behavior correlations (Vul et al., 2009). Also, early neuroimaging research was constrained by small samples and limited data availability (Button et al., 2013; Marek et al., 2022), exacerbating overfitting and spurious associations. Moreover, fMRI had been criticized for offering predictive explanations, rather than functional ones, resulting in little clinical relevance (Lyon, 2017). Finally, reverse inference emerged as a central interpretative problem, given that individual neural systems are not uniquely associated with specific cognitive functions (Poldrack, 2006; Duncan & Owen, 2000).

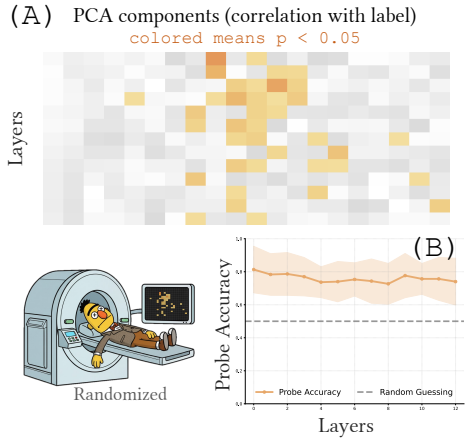

Figure 1. **Minimal dead salmon artifacts**. We extract token representations from a randomly initialized BERT for 300 IMDb sentences and average over sequence length. (A) Several principal components are spuriously correlated with sentiment labels. (B) A simple probe trained on layer representations achieves nontrivial cross-validated accuracy.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

AI interpretability now faces its own *dead salmon* issues, similarly begging for a larger reevaluation of its statistical foundations. A growing body of work has shown that many influential methods, including feature attribution (Adebayo et al., 2018), probing classifiers (Ravichander et al., 2021), sparse autoencoders (Heap et al., 2025), circuit discoveries (Méloux et al., 2025), and causal abstractions (Sutter et al., 2025), can yield plausible-looking explanations even when applied to **random neural networks**. In Figure 1, we report a minimum dead salmon artifact from analyzing activations of a fully randomized BERT model in a sentiment analysis task where both correlation analysis and probing find highly significant *explanations*. Such striking failure modes are particularly troubling as modern AI systems are increasingly deployed in high-stakes domains where AI interpretability should be essential for transparency, accountability, and error diagnosis (Mehrabi et al., 2021; Barnes & Hutson, 2024; Ramachandram et al., 2025). Interpretability methods have the potential surface critical failure modes (Kim & Canny, 2017; Zech et al., 2018; Caruana et al., 2015; Meng et al., 2022; Monea et al., 2024; Nguyen et al., 2025) and offer levers for mitigating bias and systematic errors (Arrieta et al., 2019; Kristofik, 2025; Lepori et al., 2025).

Yet, despite frequent analogies to mature sciences like *neuroscience* (Barrett et al., 2019), *biology* (Lindsey et al., 2025), or *physics* (Allen-Zhu & Li, 2023; Allen-Zhu, 2024) of neural networks, the practice of AI interpretability remains in its early foundational stages. Striking dead-salmon artifacts are accompanied by a general statistical fragility: small perturbations to inputs (Ghorbani et al., 2019; Kindermans et al., 2019; Zhang et al., 2025) or changes in random initialization (Adebayo et al., 2018; Zafar et al., 2021) can radically change explanations. Explanations often fail to generalize to new settings and input distributions (Hoelscher-Obermaier et al., 2023). Also, multiple incompatible explanations can be *discovered* for the same behavior (Méloux et al., 2025; Dombrowski et al., 2019). While the dead salmon study demonstrated a simple statistical oversight correctable through multiple comparison adjustments, AI interpretability's difficulties stem from more fundamental issues. In particular, we argue that, for common interpretability queries, computational traces do not uniquely determine explanations.

Beyond neuroscience and AI, such challenges are not unprecedented. Psychology and the social sciences faced a similar reckoning during the replication crisis, when questionable research practices produced widespread false positives (Collaboration, 2015; Simmons et al., 2011; Ioannidis, 2005; Schimmack, 2020). These fields responded with methodological reforms: pre-registration, registered reports, increased statistical power, and explicit multiple-comparison corrections (Munafò et al., 2017; Korbmacher et al., 2023). Likewise, econometrics used causal inference (Pearl, 2009) to formalize the distinction between correlation and causation, developing identification criteria, sensitivity analyses, and robustness tests (Imbens & Rubin, 2015; Angrist & Pischke, 2009; Heckman, 2007).

Now, AI interpretability can also begin to build its own methodological guardrails. As argued before, this requires both technical innovation and philosophical clarity (Miller, 2019; Williams et al., 2025). This means clarifying our epistemic goals by answering: what does it mean to "explain" a neural network? (Lillicrap & Kording, 2019; Lipton, 2018) Mechanistic interpretability embodies a type of *scientific realism*, aiming to discover the *one true* explanatory algorithm (Psillos, 2005; Chakravartty, 2011). However, there is a significant push-back against the feasibility of this research project (Rudin, 2019; Páez, 2019; Saphra & Wiegreffe, 2024), motivating a shift toward pragmatic approaches prioritizing the utility of the explanations for specific downstream goals (Zou et al., 2025). Here, we align with the pragmatic stance (Dewey, 1948; Chang, 2004; Potochnik, 2017), where explanations are seen as useful models that enable prediction, manipulation, and control (Van Fraassen, 1980; Cartwright, 1983).

This work. We analyze failure modes of contemporary AI interpretability methods, ranging from striking dead-salmon false positives to broader forms of statistical fragility, including poor generalization and high variance. We argue that these pathologies share a common root cause: the non-identifiability of many interpretability queries, compounded by the lack of principled uncertainty quantification, where non-identifiability manifests as high-variance estimates that should be reflected in large uncertainty. Diagnosing and addressing these issues, as well as articulating a coherent pragmatic research direction for interpretability, requires reframing AI interpretability as a problem of statistical (causal) inference. Accordingly, we propose one such statistical–causal reframing in which explanations are treated as parameters inferred from computational traces, enabling uncertainty-aware evaluation against meaningful alternative computational hypotheses.

2. The Statistical Fragility of AI Interpretability

Reports documenting the failure modes of interpretability methods are frequent and highlight a recurring theme: a general statistical fragility, most strikingly illustrated by dead salmon artifacts. We provide here a non-exhaustive overview of such issues.

Feature Attribution. Attribution methods (Simonyan et al., 2014; Sundararajan et al., 2017) aim to highlight input features most relevant to model predictions. However, Adebayo et al. (2018) demonstrated that saliency maps can remain

visually plausible even after model weights are randomized. Further, Dombrowski et al. (2019) showed that gradient-based explanations can be manipulated by adversarial perturbations, leaving predictions unchanged, while Ghorbani et al. (2019) revealed that explanations are unstable under minor data transformations. From a theoretical standpoint, Bilodeau et al. (2024) established impossibility results showing that no attribution method can simultaneously satisfy intuitive desiderata across broad model classes.

Probing. Probing methods train a classifier to predict a target label from internal activations. Early studies showed that both linear and structural probes could recover information with surprisingly high accuracy from randomized contextualized embeddings (Conneau et al., 2018; Hewitt & Manning, 2019), and syntactic probes do not generalize (Hall Maudslay & Cotterell, 2021). Later, Ravichander et al. (2021) demonstrated that probes can extract features merely encoded (e.g., inherited from embeddings) even if unused during inference; probing asks whether a concept is encoded in an activation, not whether it is computationally relevant. Capacity-controlled probes (Voita & Titov, 2020; Zhu & Rudzicz, 2020; Pimentel et al., 2020; Belinkov, 2022) or amnesic probing (Elazar et al., 2021) attempt to mitigate such false discoveries.

Sparse Autoencoders. Unsupervised concept-discovery pipelines such as sparse autoencoders (SAEs) (Cunningham et al., 2023; Yun et al., 2021; Bricken et al., 2023; Templeton et al., 2024) display analogous pathologies. Heap et al. (2025) showed that SAEs can recover apparently interpretable components even in randomly initialized transformers. Additional studies show that SAEs often fail to generalize across settings or tasks (Heindrich et al., 2025; Kantamneni et al., 2025). Also, Li et al. (2025a) show SAE are sensitive to adversarial input perturbations.

Concept-Based Explanations. Concept-based methods (Kim et al., 2018; Bau et al., 2017) aim to identify human-interpretable concepts that align with model representations (e.g., concept activation vectors (Kim et al., 2018) or network dissection (Bau et al., 2017)). These methods also face documented limitations (Sinha & Zhang, 2025; Aysel et al., 2025). Already, Bolukbasi et al. (2021) showed *interpretability illusion* arising where activations of individual neurons in BERT may spuriously appear to encode a concept. Then, Nicolson et al. (2025) showed that concept activation scores can produce inconsistent explanations, and Ramaswamy et al. (2023) show poor generalization and high sensitivity to the dataset used to infer concepts. Finally, Piratla et al. (2024) further demonstrated high variance and recommended incorporating uncertainty estimation.

Causal Approaches. To address issues with prediction-based explanations, a shift toward causality-based interpretability has emerged through the use of causal mediation analysis (Pearl, 2012; Elazar et al., 2021; Vig et al., 2020b; Meng et al., 2022; Finlayson et al., 2021; Syed et al., 2023; Monea et al., 2024; Mueller et al., 2024). These methods intervene on intermediate representations to quantify causal effects of components on model outputs. Yet recent work documents substantial fragilities and trade-offs (Canby et al., 2025): Zhang & Nanda (2024) showed that such approaches are sensitive to experimental design. Then, McGrath et al. (2023) discovered a "hydra effect," where ablating components identified as causally important fail to change behavior due to redundant causal pathways. This phenomenon, known as **overdetermination**, occurs when multiple redundant, independently sufficient causal pathways exist (Schaffer, 2003; Sider, 2003; Dyrkolbotn, 2017). Rather than isolating simple mechanisms, interventions tend to reveal overdetermined causal structures.

Mechanistic Interpretability. Causal approaches culminate in *mechanistic interpretability* (MI), which aims to reverse-engineer networks into human-interpretable algorithms (Olah et al., 2020). One family of approaches (*where-then-what*) first identifies circuits carrying information from inputs to outputs and then interprets their components (Dunefsky et al., 2024; Davies & Khakzar, 2024; Conmy et al., 2023). The second (*what-then-where*) instead starts from high-level candidate algorithms and searches for causally aligned neural subspaces, using *causal abstraction* metrics (Geiger et al., 2022a;b; Beckers & Halpern, 2019). Despite promising demonstrations, both categories have the typical issues (Sharkey et al., 2025). Subspace patching can produce *interpretability illusions* by activating alternate pathways (Makelov et al., 2023), also a problem of overdetermination. Circuit explanations often fail to generalize (Wang et al., 2022; Li et al., 2025b) and are sensitive to minor experimental choices (Méloux et al., 2025). Exhaustive studies on toy models reveal multiple incompatible explanations for both strategies, even for random networks (Méloux et al., 2025). Finally, Sutter et al. (2025) proved that, in general, existing causal abstraction methods can produce explanations for random networks.

Natural Language Explanations. Generating natural language rationales is also a possible approach (Marasovic et al., 2022; Wiegreffe et al., 2022). However, Ajwani et al. (2024) showed that LLM-generated explanations can be systematically unfaithful, confidently providing plausible-sounding justifications for predictions made for entirely different reasons. Moreover, chain-of-thought (self-)explanations are typically unfaithful to the model's computation (Lanham et al., 2023; Arcuschin et al., 2025; Turpin et al., 2023). Since many plausible stories can rationalize any behavior post hoc, natural language explanations are particularly susceptible to confabulation and false positives.

3. The Deeper Statistical Issue

The problems documented in Section 2 point to a broad statistical fragility. Here, we identify the common structure underlying these failures: the non-identifiability of interpretability queries.

Behavior-based approaches that study input–output relationships (e.g., feature attributions, behavioral testing) are fundamentally limited by **underspecification**: multiple, distinct explanations can equally well account for the same input–output patterns (Jacovi et al., 2021; Rogers et al., 2021; Hagendorff et al., 2023). Similar observations in cognitive science motivated the development of brain imaging as a complement to purely behavioral data, with the goal of measuring neural computation and thereby obtaining objective, measurable, and more generalizable quantities (Kosslyn, 1999; Logothetis, 2008; Churchland & Sejnowski, 1988). AI interpretability has followed a related trajectory moving toward analyzing internal computation (Mueller et al., 2024). However, predictive approaches based on internal states (probing, SAEs) inherit standard machine-learning pathologies such as overfitting and poor generalization (Belinkov, 2022). These failure modes are instances of **underspecification**: many predictive models can fit the training data equally well, leaving it unclear which ones posit generalizable causal mechanisms (Teney et al., 2022; D'Amour et al., 2022).

Causal approaches, introduced in response to the shortcomings of predictive methods, appear at first to provide the scientific rigor needed for generalizable explanations. However, AI systems are large, distributed systems with many interacting components, which gives rise to redundant and context-dependent causal pathways (Frankle & Carbin, 2019). This creates **overdetermination**, where multiple distinct causal mechanisms are each independently sufficient to produce the same behavior (Tononi et al., 1994; Loosemore, 2012; Sarkar, 2022). Then, finding mechanistic stories within complex computational systems can become *too easy*: many different, incompatible explanations can be produced for the same phenomenon (Lindsay & Bau, 2023; Méloux et al., 2025).

Identifiability. These failure modes can be formalized using the concept of identifiability. Informally, identifiability is the property of a statistical inference task stating that the parameters (explanatory variables) of a statistical model can be uniquely recovered from available observations (Casella & Berger, 2024). Identifiability is typically a prerequisite for reliable inference in the natural sciences; without it, inferred explanations remain ambiguous. Therefore, substantial work in statistics, unsupervised learning, and causal inference has focused on characterizing identifiability conditions and designing identifiable tasks (Casella & Berger, 2024; Allman et al., 2009; Locatello et al., 2019; Khemakhem et al., 2020; Shpitser & Pearl, 2008). For interpretability, both underspecification and overdetermination produce non-identifiability, explaining most of the statistical fragilities: (i) **Poor generalization:** when multiple explanations fit the observed data equally well, their explanatory claims can diverge arbitrarily on unseen data. Selecting among these explanations, therefore, depends on arbitrary inductive biases that are rarely validated. (ii) **Sensitivity to design choices:** non-identifiability implies a manifold of explanations that achieve a *good fit*. Different algorithmic choices (datasets, optimization procedures, hyperparameters) traverse this manifold differently, and thus produce different explanations. (iii) **False discovery:** when explanations are non-identifiable, the probability of recovering a spurious explanation that happens to fit the data increases with the size and complexity of the hypothesis space.

Currently, identifiability is just a conceptual analogy, because interpretability has not yet been formalized as an explicit statistical inference task. Making this formal connection and casting interpretability queries as well-specified statistical estimation problems is a necessary first step toward developing methods whose limitations and assumptions can be explicitly characterized.

4. The Statistical–Causal Inference Perspective

A straightforward way to address dead-salmon artifacts across interpretability methods is to compare findings on a trained target network against a randomized alternative: the same architecture with randomized weights analyzed by the same method. This leads to a principled hypothesis test against a null hypothesis of randomized computation, an idea foreshadowed in early work on probing (Conneau et al., 2018; Hewitt & Manning, 2019; Ravichander et al., 2021) and circuit discovery (Shi et al., 2024). We formalize such a test in Appendix A and show that, for probing, it eliminates some false discoveries and substantially reduces effect sizes in standard analyses.

While effective, directly correcting dead-salmon artifacts is a very low bar for interpretability. The goal is to address the deeper statistical issues that give rise to these failures in the first place. Nevertheless, hypothesis testing against computationally meaningful null alternatives naturally motivates a broader statistical–causal reframing of interpretability. Here, we sketch one such formalization, viewing interpretability as a problem of *statistical–causal inference*. In this view, explanations are *surrogate models* constructed to answer distributions of causal queries about a computational system. An explanation is useful insofar as it supports prediction and manipulation, generalizes under intervention, and remains robust to noise. This perspective aligns with a growing pragmatist approach to interpretability (Páez, 2019).

4.1. Background: Statistical–Causal Inference

Statistical inference provides the rigorous framework through which empirical observations become scientific knowledge (Cox, 2006; Lehmann & Casella, 1998). We argue that interpretability, like every empirical science, must be grounded in these principles. We provide here a brief overview.

Statistical Models and Identifiability. A *statistical model* is a family of probability laws $\{\mathbb{P}_\theta^{\mathbf{V}} : \theta \in \Theta\}$ on a sample space $\mathcal{V}$, indexed by parameters $\theta \in \Theta$. Here, $\mathbf{V}$ denotes observed data. Intuitively, we assume data arises from some process indexed by unknown parameters θ, and the goal is to recover θ from observations. Sound inference requires *identifiability*: distinct parameters must induce distinct distributions over observables. Formally, a model is identifiable if $\theta \neq \theta' \implies \mathbb{P}_\theta^{\mathbf{V}} \neq \mathbb{P}_{\theta'}^{\mathbf{V}}$. Without identifiability, hypotheses cannot be distinguished from data, rendering inference ill-posed.

Estimators and Uncertainty Quantification. Given finite observations $\mathcal{D}_n = \{\mathbf{v}^{(i)}\}_{i=1}^n$, an *estimator* T produces an estimate $\hat{\theta} := T(\mathcal{D}_n)$ of unknown parameters θ. Its quality can be assessed through various statistical properties: (i) **Bias**: Does it recover the correct parameter on average? (ii) **Variance**: How much does the estimate vary across datasets? (iii) **Consistency**: Does it converge to the correct parameter as $n \to \infty$? Beyond point estimates, **confidence sets** provide uncertainty quantification under finite sampling.

Causal Inference. Many scientific questions go beyond prediction, seeking explanations of *how* variables influence one another. This requires enriching statistical models with a causal structure (Pearl, 2009). Let $\mathbf{V} = \{V_1, \ldots, V_d\}$ denote *endogenous variables*, quantities computed within the system. A directed graph $\mathcal{G}$ over nodes $\mathbf{V}$ encodes direct causal relationships: an edge $V_i \to V_j$ indicates that V_i directly causes V_j. A *structural causal model* (SCM) is the tuple $\mathfrak{C} = (\mathbf{V}, \mathbf{U}, \mathbf{f}, P_{\mathbf{U}})$, where:

- $\mathbf{U}$ collects *exogenous* (external) inputs representing unobserved causes or environmental randomness, $P_{\mathbf{U}}$ is their joint distribution
- $\mathbf{f} = \{f_1, \ldots, f_d\}$ are *structural assignments*, functions that deterministically compute each variable from its causes:

$$V_i = f_i(\mathbf{PA}_i, U_i), \qquad i = 1, \ldots, d, \tag{1}$$

where $\mathbf{PA}_i \subseteq \mathbf{V}$ denotes the parents of V_i in $\mathcal{G}$, and $U_i \in \mathbf{U}$ is its exogenous input.

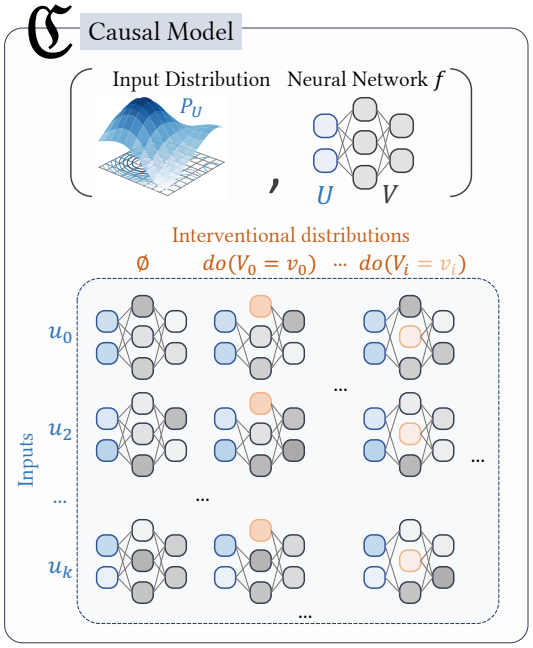

Figure 2. The tuple of a target behavior $P_{\mathbf{U}}$ and the computational system with its internal components form an SCM.

SCMs enable reasoning about interventions and counterfactuals. A *hard intervention* $\mathrm{do}(\mathbf{V}_I = \mathbf{v}_I)$ on a subset $\mathbf{V}_I \subseteq \mathbf{V}$ replaces the structural assignments for variables in $\mathbf{V}_I$ with constants $\mathbf{v}_I$, overriding their causal mechanisms. The intervened model $\mathfrak{C}; \mathrm{do}(\mathbf{V}_I = \mathbf{v}_I)$ induces an *interventional distribution* $P_{\mathbf{V}}^{\mathfrak{C};\mathrm{do}(\mathbf{V}_I = \mathbf{v}_I)}$, which captures how the system behaves under this external manipulation.

A *causal query* $q(\mathfrak{C})$ is any well-defined question about the SCM, such as "What is the marginal distribution of V_i?" or "What is the average effect of setting $V_i = v$ on outcome V_m?" Thus, a causal query is any measurable functional of the SCM, possibly involving conditioning or intervention. Central to causal inference is *query identifiability*: whether $q(\mathfrak{C})$ can be uniquely determined from available observational or interventional data.

4.2. Neural Networks as Structural Causal Models

Returning to modern AI interpretability, we first state a standard framing of computational systems as SCMs. Let f be a computational system, typically a neural network, with internal computational elements $\mathbf{V}$ and input distribution $P_{\mathbf{U}}$. The input distribution $P_{\mathbf{U}}$ represents the *behavior of interest* that we aim to explain. For instance, $P_{\mathbf{U}}$ might represent arithmetic prompts to a language model, images from a particular domain, or factual questions about a specific topic.

The tuple $(f, P_{\mathbf{U}})$ naturally defines an SCM $\mathfrak{C} =$

$(\mathcal{G},\ \mathbf{V},\ \mathbf{U},\ \mathbf{f},\ P_{\mathbf{U}})$, where:

- **Endogenous variables $\mathbf{V}$** are the network's computational variables (e.g., hidden states, attention patterns, outputs).
- **Exogenous variables $\mathbf{U}$** are inputs sampled from $P_{\mathbf{U}}$, representing the behavior we seek to explain.
- **Structural assignments $\mathbf{f}$** are the deterministic functions defining the network's computation (layers, attention mechanisms, nonlinearities).
- **Causal graph $\mathcal{G}$**: the network's computation graph.

The SCM induces a unique *observational distribution* over $\mathbf{V}$: sampling corresponds to drawing inputs from $P_{\mathbf{U}}$, executing a forward pass, and recording desired activations. Also, the SCM encodes *interventional* and *counterfactual* distributions based on external modifications of the inner computation. This perspective is standard within mechanistic interpretability (Olah et al., 2018; Cammarata et al., 2020; Geiger et al., 2022b; 2025) and is illustrated in Figure 2. Then, a *causal query* is any well-specified quantity about $\mathfrak{C} := (f, P_{\mathbf{U}})$, such as "What distribution would the network produce if we forced activation V_i to value v?" or "How much does attention head V_a causally contribute to correct factual recall?"

4.3. Explanations as Surrogate Models

In an attempt to provide a general statistical-causal perspective on interpretability, we formalize explanations as *surrogate models*: simpler computational descriptions designed to answer chosen collections of causal queries about a target system. This perspective treats interpretability as a form of model compression, where we seek a simpler model that faithfully approximates a complex system's behavior for queries we care about. In this perspective, every interpretability method is characterized by three ingredients:

1. **Query space Q with distribution μ**: The set of causal queries to be answered by the explanation. It dictates what aspects of $\mathfrak{C}$ should be explained. This encodes our explanatory goals.
2. **Surrogate set $\mathcal{E}$**: The class of admissible explanations. It dictates what forms the explanation can take, e.g., circuits, sparse subgraphs, linear probes, concept vectors, causal graphs, ...
3. **Discrepancy measure D**: How we measure whether a surrogate (member of $\mathcal{E}$) *correctly* answers queries.

This framework is (non-rigorously) illustrated with the example of circuit discovery in Figure 3. While standard

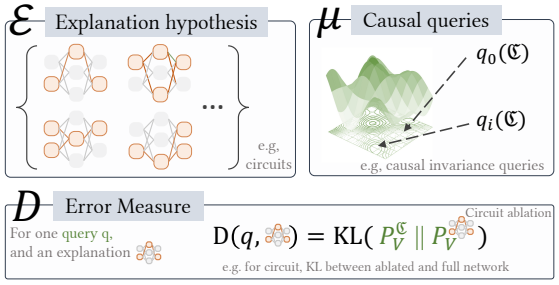

Figure 3. An interpretability task is defined by three elements: $\mathcal{E}$, the hypothesis space; μ, the distribution over causal queries about the SCM (model and behavior); and D, the error measure.

causal inference often concentrates μ on a *single* causal query (e.g., average treatment effect), interpretability aims to answer *many diverse queries* drawn from a non-trivial distribution μ. For example, μ might distribute probability over interventional queries and counterfactual queries across network components, or any functionals of the interventional and counterfactual distributions.

For instance, we can view each candidate explanation $e \in \mathcal{E}$ as defining a *query-answering map* $S_e : Q \to \mathcal{R}$, where $S_e(q)$ is the surrogate's predicted answer to query q, and $\mathcal{R}$ is the space of possible answers for query q (e.g., probability distributions, scalar effects, or discrete predictions). The surrogate's *fidelity* is measured by the discrepancy function $D : \mathcal{R} \times \mathcal{R} \to \mathbb{R}_+$ quantifying the error between the answer from $q(\mathfrak{C})$ and the surrogate's prediction $S_e(q)$. Then, we can define the *population risk* of a candidate explanation as the expected error over queries:

$$L_\mu(e) = \mathbb{E}_{q\sim\mu}\big[D\big(q(\mathfrak{C}),\ S_e(q)\big)\big]. \tag{2}$$

An ideal explanation $e^* \in \mathcal{E}$ minimizes this risk: $e^* \in \arg\min_{e\in\mathcal{E}} L_\mu(e)$.

Interpretability Task and Identifiability. We call the triple $(\mu, \mathcal{E}, D)$ an *interpretability task*, fully specifying what we aim to explain (query distribution μ), what explanations are admissible (hypothesis class $\mathcal{E}$), and how we measure success (discrepancy D). The task is *identifiable* if L_μ admits a unique minimizer in $\mathcal{E}$ (potentially up to predefined acceptable symmetries). Identifiability captures whether the surrogate set can, in principle, be distinguished using the queries deemed relevant by μ. Without identifiability, multiple incompatible explanations achieve the same population risk, making inference fundamentally ambiguous. The analysis of Section 2 indicates that the most common tasks are not identifiable. Finally, this reframing highlights that explanations are *pragmatic computational summaries* of the structure encoded by $\mathfrak{C} := (f, P_{\mathbf{U}})$. They are inferred models useful for specific explanatory purposes.

Estimation with Finite Data. In practice, we face two types of finite sampling difficulties. First, we observe only finitely many queries $q_1, \ldots, q_n \sim \mu$ from the query distribution. Second, for each query q_j, we can only collect a finite amount of computational traces by sampling inputs $\mathbf{U} \sim P_{\mathbf{U}}$ and recording the corresponding activations and outputs, potentially under interventions. Let $\mathcal{T}_n$ denote the complete dataset of query-trace pairs. An interpretability method M acts as an *estimator*, mapping this finite dataset to an estimated surrogate explanation:

$$\hat{e} := M(\mathcal{T}_n). \tag{3}$$

A natural estimator is given by empirical risk minimization: choose $\hat{e}$ to minimize the empirical risk $\widehat{L_\mu}(e) = \frac{1}{n}\sum_{j=1}^{n} \widehat{D}(q_j(\mathfrak{C}), S_e(q_j))$, where $\widehat{D}$ is estimated from finite traces.

Relevant to this exposition, Senetaire et al. (2023) proposed a statistical framing for feature attribution. Then, previous works already explored hypothesis testing and uncertainty quantification for circuit discovery (Shi et al., 2024; Méloux et al., 2025).

4.4. Re-Interpreting Documented Issues

The framework does not prescribe which $(\mu, \mathcal{E}, D)$ researchers should adopt. Rather, it provides a *shared language* for making assumptions explicit and rooted in the tools of statistical inference. Different research programs will choose different hypothesis classes or query distributions; the framework ensures that such choices are transparent and their implications are analyzable. Table 1 in the appendix illustrates how existing interpretability methods can be mapped into this formulation, each implicitly making assumptions about queries, surrogates, and error metrics. Under this view, the issues described in Section 2 can be understood as problems of *non-identifiability*.

Behavioral Benchmarks. Benchmarks that evaluate model outputs against a gold standard (averaged success over input distributions) ask an identifiable question: *how well does the model perform under a specific task distribution and error metric?* This is arguably the simplest form of interpretability and drove most of the progress in AI. Its usefulness depends on the construction of the benchmark, but the inference problem is well-posed.

Concept-Based Approaches. Predictive methods (probes, SAEs) inherit non-identifiability issues from the underlying underspecification of machine learning tasks (D'Amour et al., 2022). For example, methods like Concept Activation Vectors (Kim et al., 2018; Cunningham et al., 2023) postulate that internal states $\mathbf{v}$ are generated by interpretable concepts $\mathbf{z}$ via $\mathbf{v} = g(\mathbf{z})$. This is an instance of (causal) representation learning, which is **non-identifiable** without auxiliary information (Locatello et al., 2019; Khemakhem et al., 2020). It is therefore unsurprising that proposed improvements mirror standard remedies for underspecification in machine learning: regularization in the form of capacity control for probes (Belinkov, 2022) or cross-validation to assess generalization (Kantamneni et al., 2025).

Causal Mediation Analysis. Methods like causal mediation analysis (Meng et al., 2022; Vig et al., 2020a) estimate the indirect effect of a component on observed outputs. As the intervention and model are fully specified, the mediation estimand is unique and **identifiable**. However, the explanatory claim that a component with high effect is the *locus* of a mechanism is **not identifiable**, because of the overdetermined causal structure.

Circuit Discovery and Causal Approaches. Circuit discovery seeks a subgraph $G' \subset G$ that preserves the model's performance. This task faces the "Hydra effect" (McGrath et al., 2023) and causal overdetermination. If parallel pathways A and B are sufficient, circuits containing only A or only B both satisfy fidelity criteria. Thus, even *correct* causal methods may recover many different explanations consistent with the same behavior (Méloux et al., 2025). Addressing this requires formulating identifiable causal questions. Causal abstraction (Beckers & Halpern, 2019; Geiger et al., 2025) offers a promising direction, as it operates at a coarser representational level where overdetermination can be absorbed into the abstracted representations. However, current operational metrics demonstrate empirical non-identifiability (Méloux et al., 2025; Sutter et al., 2025).

5. Discussion

The systematic failures documented in Section 2 demanded an explanation. We have argued that these pathologies share a common root cause: **non-identifiability**. Most current interpretability tasks attempt to infer explanatory structures that are not uniquely determined by available computational traces. To trace a path forward, we proposed one formalization of interpretability as statistical-causal inference. This framework is tentative rather than definitive; we encourage the community to improve upon it. The important aspect is the *methodological commitment* to making assumptions explicit and quantifying uncertainty rigorously.

5.1. Advantages of the Statistical-Causal Perspective

Drawing on the philosophy of science (Chang, 2004; Woodward, 2003; Potochnik, 2017) and recent calls for a pragmatic approach to interpretability (Davies & Khakzar, 2024; Williams et al., 2025), the framework naturally distinguishes the *explanandum* (what is to be explained, encoded in μ)

from the *explanans* (what does the explaining, encoded in $\mathcal{E}$). Researchers and practitioners have substantial freedom in choosing both. There is no single "correct" explanation of a neural network. The appropriate type of description depends on one's purposes (Potochnik, 2017). Descriptive understanding corresponds to queries about observational distributions; predictive goals involve queries requiring surrogates to generalize to new input distributions; control and intervention require queries about counterfactual or interventional distributions.

However, once the explanatory project is specified, i.e., once $(\mu, \mathcal{E}, D)$ are fixed, the explanation becomes an **objective inference problem**. The best surrogate $e^* \in \mathcal{E}$ is the one minimizing $L_\mu(e)$, and is a property of the system itself and the interpretability task. If the task is identifiable, this explanation is unique (up to permissible symmetries, e.g., rotation invariance in representation space). This reconciles pluralism about explanatory goals with rigor about explanatory claims.

Perhaps most critically, the statistical framing demands that interpretability methods report not just point estimates but *confidence sets* or *posterior distributions* over explanations (in case of Bayesian framing). Just as we would not trust a clinical trial reporting effect sizes without confidence intervals, we may not trust interpretability claims without uncertainty quantification. When explanations are non-identifiable, this uncertainty will be large; when they are identifiable with finite data, uncertainty shrinks as observations accumulate.

5.2. Towards Useful and Identifiable Interpretability Tasks

Identifiability is not an intrinsic property of the model under study but of the interaction between μ, $\mathcal{E}, D$, the model f_{NN}, and the behavior of interest $P_{\mathbf{U}}$. We might wonder what choices to make in order to improve the identifiability and usefulness of interpretability queries.

Query richness. The queries in the support of μ must be sufficiently discriminative to distinguish candidate explanations in $\mathcal{E}$. There is a fundamental trade-off between discriminative power and sample efficiency. If μ spreads probability mass over a large support, accurately estimating L_μ may require prohibitive amounts of interventional data. Conversely, concentrating μ on too few queries risks not singling out one explanation in $\mathcal{E}$.

Expressivity vs. parsimony in $\mathcal{E}$. Conversely, the hypothesis class must have sufficient capacity to approximate the queries well (low bias) but not so much flexibility that many distinct explanations all achieve low error (large equivalence classes, high variance, non-identifiability). This is akin to the classical bias-variance tradeoff, pointing toward standard fixes like regularization of the hypothesis class (Belinkov, 2022).

Human cognitive constraints. Interpretability is meant to facilitate *human understanding*. Empirical studies suggest people can mentally simulate models with only a handful of interacting components (Lombrozo, 2006; Wilkenfeld, 2013; Keil, 2006; Hassija et al., 2024). Explanations exceeding these structural limits may be technically *correct* yet fail to provide insight. Designing $\mathcal{E}$ with human simulability in mind ensures that understanding remains the end goal.

5.3. Opportunities for Future Work

Characterizing identifiability conditions. A systematic theoretical program could characterize when specific $(\mu, \mathcal{E}, D)$ triplets are identifiable, mirroring similar efforts in causal inference (Shpitser & Pearl, 2008) and unsupervised learning (Locatello et al., 2019; Khemakhem et al., 2020). What symmetries and invariances are unavoidable in representation space, and when is identifiability up to such equivalences acceptable? Constructing a taxonomy of identifiable interpretability tasks would provide actionable guidance for practical scenarios.

Bayesian interpretability and uncertainty quantification. Bayesian approaches offer an elegant framework for handling non-identifiability and quantifying uncertainty (Gelman et al., 2013). Specifically, one could specify a **prior distribution** $\pi(e)$ over the explanation class $\mathcal{E}$, encoding structural preferences (e.g., sparsity, modularity) or incorporating prior information from related studies. Then, the **likelihood model** $P(\mathcal{T}_n \mid e)$ describes how computational traces are generated given explanation e. Finally, the **posterior updates** via Bayes' rule: $\pi(e \mid \mathcal{T}_n) \propto P(\mathcal{T}_n \mid e)\pi(e)$, refines beliefs as observations accumulate. Then, **credible sets** can quantify uncertainty. When explanations are non-identifiable, the posterior remains diffuse across an equivalence class; uncertainty quantification naturally reflects this fundamental ambiguity. Conversely, as more discriminative queries are observed, the posterior concentrates. This further provides a principled framework for active setup: strategically selecting queries from μ that maximally reduce posterior uncertainty.

Meta-analysis and cumulative science. Meta-analytic methods (Borenstein et al., 2021) could coherently aggregate evidence across studies, accounting for heterogeneity in μ, $\mathcal{E}$, and experimental conditions. Standardized effect size measures, pre-registration of analyses, and open sharing of collected computational traces would enable interpretability to become a cumulative science where knowledge systematically builds over time. In general, responses proposed by other fields (Poldrack et al., 2017; Korbmacher et al., 2023) become available for interpretability.

Impact Statement

This work aims to improve the scientific rigor and reliability of Mechanistic Interpretability (MI). As MI techniques are increasingly proposed for safety auditing, model alignment, and regulatory compliance, it is critical that these methods produce stable and statistically valid explanations. Our research highlights the risks of relying on unstable point-estimates, which can lead to unjustified confidence in a model's safety properties or internal mechanisms. By advocating for statistical robustness and best practices in circuit discovery, this work contributes to the development of more trustworthy AI systems and helps ensure that future interpretability tools provide a solid foundation for policy and safety decisions.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Ajwani, R., Javaji, S. R., Rudzicz, F., and Zhu, Z. Llm-generated black-box explanations can be adversarially helpful, 2024. URL https://arxiv.org/abs/2405.06800.

Allen-Zhu, Z. ICML 2024 Tutorial: Physics of Language Models, July 2024. Project page: https://physics.allen-zhu.com/.

Allen-Zhu, Z. and Li, Y. Physics of Language Models: Part 1, Learning Hierarchical Language Structures. *SSRN Electronic Journal*, May 2023. Full version available at https://ssrn.com/abstract=5250639.

Allman, E. S., Matias, C., and Rhodes, J. A. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009.

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful. In *Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=L8094Whth0.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. URL https://arxiv.org/abs/1910.10045.

Aysel, H. I., Cai, X., and Prugel-Bennett, A. Concept-based explainable artificial intelligence: Metrics and benchmarks, 2025. URL https://arxiv.org/abs/2501.19271.

Barnes, E. and Hutson, J. Navigating the complexities of ai: The critical role of interpretability and explainability in ensuring transparency and trust. *International Journal of Multidisciplinary and Current Educational Research*, 6 (3), 2024.

Barrett, D. G., Morcos, A. S., and Macke, J. H. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology*, 55:55–64, 2019.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Beckers, S. and Halpern, J. Y. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2678–2685, Jul. 2019. doi: 10.1609/aaai.v33i01.33012678. URL https://ojs.aaai.org/index.php/AAAI/article/view/4117.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7.

Bennett, C. M., Miller, M. B., and Wolford, G. L. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for multiple comparisons correction. *Neuroimage*, 47(Suppl 1):S125, 2009.

Bilodeau, B., Jaques, N., Koh, P. W., and Kim, B. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

Bolukbasi, T., Pearce, A., Yuan, A., Coenen, A., Reif, E., Viégas, F., and Wattenberg, M. An interpretability illusion for bert, 2021. URL https://arxiv.org/abs/2104.07143.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. *Introduction to meta-analysis*. John wiley & sons, 2021.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 2020.

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. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5): 365–376, 2013.

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., and Lim, S. K. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

Canby, M., Davies, A., Rastogi, C., and Hockenmaier, J. How reliable are causal probing interventions?, 2025. URL https://arxiv.org/abs/2408.15510.

Cartwright, N. *How the laws of physics lie*. Oxford University Press, 1983.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Casella, G. and Berger, R. *Statistical inference*. CRC press, 2024.

Chakravartty, A. Scientific realism. *The Stanford Encyclopedia of Philosophy*, (Summer 2017 Edition), 2011. URL https://plato.stanford.edu/archives/sum2017/entries/scientific-realism.

Chang, H. *Inventing temperature: Measurement and scientific progress*. Oxford University Press, 2004.

Churchland, P. S. and Sejnowski, T. J. Perspectives on cognitive neuroscience. *Science*, 242(4879):741–745, 1988.

Collaboration, O. S. Estimating the reproducibility of psychological science. *Science*, 349(6251): aac4716, 2015. doi: 10.1126/science.aac4716. URL https://www.science.org/doi/abs/10.1126/science.aac4716.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198.

Cox, D. R. *Principles of statistical inference*. Cambridge university press, 2006.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(1), January 2022. ISSN 1532-4435.

Davies, A. and Khakzar, A. The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms, 2024. URL https://arxiv.org/abs/2408.05859.

Dewey, J. *Reconstruction in Philosophy*. Dover Publications, Mineola, N.Y., 1948.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.

cc/paper_files/paper/2019/file/bb836c01cdc9120a9c984c525e4b1a4a-Paper.pdf.

Duncan, J. and Owen, A. M. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in neurosciences*, 23(10):475–483, 2000.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable llm feature circuits, 2024. URL https://arxiv.org/abs/2406.11944.

Dyrkolbotn, S. K. On preemption and overdetermination in formal theories of causality. *Electronic Proceedings in Theoretical Computer Science*, 259:1–15, October 2017. ISSN 2075-2180. doi: 10.4204/eptcs.259.1. URL http://dx.doi.org/10.4204/EPTCS.259.1.

Eklund, A., Nichols, T. E., and Knutsson, H. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905, 2016.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10.

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., and Belinkov, Y. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1828–1843, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*. OpenReview.net, 2019. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2019.html#FrankleC19.

Geiger, A., Wu, Z., D'Oosterlinck, K., Kreiss, E., Goodman, N. D., Icard, T., and Potts, C. Faithful, interpretable model explanations via causal abstraction. Stanford AI Lab Blog, 2022a. URL https://ai.stanford.edu/blog/causal-abstraction/.

Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., and Potts, C. Inducing causal structure for interpretable neural networks. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/geiger22a.html.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., and Icard, T. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2025. URL https://arxiv.org/abs/2301.04709.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science Series. CRC, Boca Raton, Florida, third edition, 2013. ISBN 9781439840955 1439840954. URL https://stat.columbia.edu/~gelman/book/.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, Jul. 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://ojs.aaai.org/index.php/AAAI/article/view/4252.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C., Lampinen, A., Wang, J. X., Akata, Z., and Schulz, E. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2023.

Hall Maudslay, R. and Cotterell, R. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 124–131, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.11. URL https://aclanthology.org/2021.naacl-main.11.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.

Heap, T., Lawson, T., Farnik, L., and Aitchison, L. Sparse autoencoders can interpret randomly initialized transformers, 2025. URL https://arxiv.org/abs/2501.17727.

Heckman, J. J. The economics, technology, and neuroscience of human capability formation. *Proceedings of the national Academy of Sciences*, 104(33):13250–13255, 2007.

Heindrich, L., Torr, P., Barez, F., and Thost, V. Do sparse autoencoders generalize? a case study of answerability, 2025. URL https://arxiv.org/abs/2502.19964.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

Hoelscher-Obermaier, J., Persson, J., Kran, E., Konstas, I., and Barez, F. Detecting edit failures in large language models: An improved specificity benchmark, 2023. URL https://arxiv.org/abs/2305.17553.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Ioannidis, J. P. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., and Goldberg, Y. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021.

Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., and Nanda, N. Are sparse autoencoders useful? a case study in sparse probing, 2025. URL https://arxiv.org/abs/2502.16681.

Keil, F. C. Explanation and understanding. *Annual Reviews of Psychology*, 57(1):227–254, 2006.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pp. 2207–2217. PMLR, 2020.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Kim, J. and Canny, J. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. *The (Un)reliability of Saliency Methods*, pp. 267–280. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., et al. The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, 1(1):3, 2023.

Kosslyn, S. M. If neuroimaging is the answer, what is the question? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 354 (1387):1283–1294, 1999.

Kristofik, A. Bias in ai (supported) decision making: Old problems, new technologies. In *International Journal for Court Administration*, volume 16, pp. 1. HeinOnline, 2025.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer, 1998.

Lepori, M. A., Mozer, M. C., and Ghandeharioun, A. Racing thoughts: Explaining contextualization errors in large language models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3020–3036, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.155. URL https://aclanthology.org/2025.naacl-long.155/.

Li, A. J., Srinivas, S., Bhalla, U., and Lakkaraju, H. Interpretability illusions with sparse autoencoders: Evaluating robustness of concept representations, 2025a. URL https://arxiv.org/abs/2505.16004.

Li, V. R., Kaufmann, J., Wattenberg, M., Alvarez-Melis, D., and Saphra, N. Can interpretation predict behavior on unseen data?, 2025b. URL https://arxiv.org/abs/2507.06445.

Lillicrap, T. P. and Kording, K. P. What does it mean to understand a neural network?, 2019. URL https://arxiv.org/abs/1907.06374.

Lindsay, G. W. and Bau, D. Testing methods of neural systems understanding. *Cognitive Systems Research*, 82: 101156, 2023.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Locatello, F., Bauer, S., Lucic, M., Ratsch, G., Gelly, S., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 4114–4124. PMLR, 2019.

Logothetis, N. K. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

Lombrozo, T. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Loosemore, R. P. The complex cognitive systems manifesto. In *Nanotechnology, the Brain, and the Future*, pp. 195–217. Springer, 2012.

Lyon, L. Dead salmon and voodoo correlations: should we be sceptical about functional mri? *Brain*, 140(8):e53–e53, 2017.

Makelov, A., Lange, G., and Nanda, N. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching, 2023. URL https://arxiv.org/abs/2311.17030.

Marasovic, A., Beltagy, I., Downey, D., and Peters, M. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 410–424, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-naacl.31.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022.

McGrath, T., Rahtz, M., Kramar, J., Mikulik, V., and Legg, S. The hydra effect: Emergent self-repair in language model computations, 2023. URL https://arxiv.org/abs/2307.15771.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Méloux, M., Maniu, S., Portet, F., and Peyrard, M. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *The Thirteenth International Conference on Learning Representations*, 2025.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Monea, G., Peyrard, M., Josifoski, M., Chaudhary, V., Eisner, J., Kıcıman, E., Palangi, H., Patra, B., and West, R. A glitch in the matrix? locating and detecting language model grounding with fakepedia, 2024. URL https://arxiv.org/abs/2312.02073.

Mueller, A., Brinkmann, J., Li, M., Marks, S., Pal, K., Prakash, N., Rager, C., Sankaranarayanan, A., Sharma, A. S., Sun, J., Todd, E., Bau, D., and Belinkov, Y. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability, 2024. URL https://arxiv.org/abs/2408.01416.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A manifesto for reproducible science. *Nature human behaviour*, 1(1):0021, 2017.

Méloux, M., Portet, F., and Peyrard, M. Mechanistic interpretability as statistical estimation: A variance analysis of eap-ig, 2025. URL https://arxiv.org/abs/2510.00845.

Nguyen, N., Deng, M., Gala, D., Naruse, K., Virgo, F. G., Byun, M., Hazra, D., Gorton, L., Balsam, D., McGrath, T., Takei, M., and Kaji, Y. Deploying interpretability to production with rakuten: Sae probes for pii detection. *Goodfire Research*, 2025. https://www.goodfire.ai/blog/deploying-interpretability-to-production-with-rakuten.

Nicolson, A., Schut, L., Noble, J. A., and Gal, Y. Explaining explainability: Recommendations for effective use of concept activation vectors, 2025. URL https://arxiv.org/abs/2404.03713.

North, B. V., Curtis, D., and Sham, P. C. A note on the calculation of empirical p values from monte carlo procedures. *The American Journal of Human Genetics*, 71(2): 439–441, 2002.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Pearl, J. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention science*, 13:426–436, 2012.

Phipson, B. and Smyth, G. K. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9:Article39, 2010.

Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., and Cotterell, R. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420.

Piratla, V., Heo, J., Collins, K. M., Singh, S., and Weller, A. Estimation of concept explanations should be uncertainty aware, 2024. URL https://arxiv.org/abs/2312.08063.

Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences*, 10(2): 59–63, 2006.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, 18(2):115–126, 2017.

Potochnik, A. Idealization and the aims of science. In *Idealization and the Aims of Science*. University of Chicago Press, 2017.

Psillos, S. *Scientific realism: How science tracks truth*. Routledge, 2005.

Páez, A. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3): 441–459, May 2019. ISSN 1572-8641. doi: 10.1007/s11023-019-09502-w. URL http://dx.doi.org/10.1007/s11023-019-09502-w.

Ramachandram, D., Joshi, H., Zhu, J., Gandhi, D., Hartman, L., and Raval, A. Transparent ai: The case for interpretability and explainability, 2025. URL https://arxiv.org/abs/2507.23535.

Ramaswamy, V. V., Kim, S. S., Fong, R., and Russakovsky, O. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10932–10941, 2023.

Ravichander, A., Belinkov, Y., and Hovy, E. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295.

Rogers, A., Kovaleva, O., and Rumshisky, A. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866, 2021.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Saphra, N. and Wiegreffe, S. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 480–498, 2024.

Sarkar, A. Is explainable ai a race against model complexity?, 2022. URL https://arxiv.org/abs/2205.10119.

Schaffer, J. Overdetermining causes. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 114(1/2):23–45, 2003.

Schimmack, U. A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne*, 61(4):364, 2020.

Senetaire, H. H. J., Garreau, D., Frellsen, J., and Mattei, P.-A. Explainability as statistical inference. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30584–30612. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/senetaire23a.html.

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. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Shi, C., Beltran-Velez, N., Nazaret, A., Zheng, C., Garriga-Alonso, A., Jesson, A., Makar, M., and Blei, D. M. Hypothesis testing the circuit hypothesis in llms, 2024. URL https://arxiv.org/abs/2410.13032.

Shpitser, I. and Pearl, J. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Sider, T. What's so bad about overdetermination?, 2003.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Sinha, S. and Zhang, A. A comprehensive survey on the risks and limitations of concept-based models, 2025. URL https://arxiv.org/abs/2506.04237.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Sutter, D., Minder, J., Hofmann, T., and Pimentel, T. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability?, 2025. URL https://arxiv.org/abs/2507.08802.

Syed, A., Rager, C., and Conmy, A. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Teney, D., Peyrard, M., and Abbasnejad, E. Predicting is not understanding: Recognizing and addressing underspecification in machine learning. In *ECCV 2022: 17th European Conference on Computer Vision*, pp. 458–476, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20049-6. doi: 10.1007/978-3-031-20050-2_27. URL https://doi.org/10.1007/978-3-031-20050-2_27.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452.

Tononi, G., Sporns, O., and Edelman, G. M. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91 (11):5033–5037, 1994. doi: 10.1073/pnas.91.11.5033. URL https://www.pnas.org/doi/abs/10.1073/pnas.91.11.5033.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

Van Fraassen, B. C. *The scientific image*. Oxford University Press, 1980.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020a. URL https://arxiv.org/abs/2004.12265.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Voita, E. and Titov, I. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. Voodoo correlations in social neuroscience. *Perspectives on psychological Science*, 4(3):274–290, 2009.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., and Choi, Y. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 632–658, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.47.

Wilkenfeld, D. A. Understanding as representation manipulability. *Synthese*, 190(6):997–1016, 2013.

Williams, I., Oldenburg, N., Dhar, R., Hatherley, J., Fierro, C., Rajcic, N., Schiller, S. R., Stamatiou, F., and Søgaard, A. Mechanistic interpretability needs philosophy, 2025. URL https://arxiv.org/abs/2506.18852.

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

Yun, Z., Chen, Y., Olshausen, B., and LeCun, Y. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 1–10, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.deelio-1.1. URL https://aclanthology.org/2021.deelio-1.1.

Zafar, M. B., Donini, M., Slack, D., Archambeau, C., Das, S., and Kenthapadi, K. On the lack of robust interpretability of neural text classifiers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3730–3740, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.327. URL https://aclanthology.org/2021.findings-acl.327.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods, 2024. URL https://arxiv.org/abs/2309.16042.

Zhang, H., Figueroa, F. T., and Hermanns, H. Saliency Maps Give a False Sense of Explanability to Image Classifiers: An empirical evaluation across methods and metrics. In Nguyen, V. and Lin, H.-T. (eds.), *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pp. 479–494. PMLR, 05–08 Dec 2025. URL https://proceedings.mlr.press/v260/zhang25a.html.

Zhu, Z. and Rudzicz, F. An information theoretic view on selecting linguistic probes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9251–9262, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.744. URL https://aclanthology.org/2020.emnlp-main.744.

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., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A. Fixing Dead Salmons with Hypothesis Testing

Consider an interpretability method M that aims to explain a neural network f for some input behavior $P_{\mathbf{U}}$, we note $\mathfrak{C}$ the tuple $(f, P_{\mathbf{U}})$ as done in the main paper. The method produces an explanation $\hat{e}$ from finite observations from $\mathfrak{C}$, possibly under interventions. This tentative explanation could take the form of a circuit, a set of important features, concept activation vectors, or any other hypothesis class. We might wonder how to prevent dead salmon artifacts from arising with the interpretability method M?

A simple, direct, and natural solution is to frame this question as a hypothesis test against a null hypothesis where the observed explanation arises from random computation. Already, for probing, Ravichander et al. (2021) discusses the possibility of comparing the probe against a probe trained on random embeddings. The general idea is to construct a family of null models represented by a distribution $P_{\tilde{\mathfrak{C}}}$, which preserves the network's architectural properties while disrupting the specific computational mechanisms we aim to explain. Such null models can be obtained, for example, via full weight randomization, random orthogonal transformations of representations, or label shuffling (recovering the standard permutation test).

For a given interpretability method, we define a test statistic $T(\hat{e}, \mathfrak{C})$ that quantifies explanatory fit for the interpretability task at hand. For example, for probing methods, T could be test accuracy; for circuit discovery, T could measure behavioral fidelity; for attribution methods, T could quantify the correlation between attribution scores and actual intervention effects.

Applying the intepretability method M to one null model $\tilde{\mathfrak{C}}^{(b)}$ from the randomized family yields explanations $\tilde{e}^{(b)} = M(\tilde{\mathfrak{C}}^{(b)})$ and corresponding null statistics $T^{(b)}_{\text{null}} = T(\tilde{e}^{(b)}, \tilde{\mathfrak{C}}^{(b)})$. Then, following standard procedure, the Monte Carlo estimated p-value is:

$$\hat{p} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\{T^{(b)}_{\text{null}} \geq T_{\text{obs}}\}}{B + 1}, \tag{4}$$

where $T_{\text{obs}} = T(\hat{e}, \mathfrak{C})$. The addition of 1 to both the numerator and denominator ensures Type I error control: $\Pr(\hat{p} \leq \alpha \mid H_0) \leq \alpha$, where H_0 is the null hypothesis (North et al., 2002; Phipson & Smyth, 2010). By design, when the randomization includes full weight reinitialization, no dead salmon artifacts can remain.

A.1. Experiments

To illustrate the hypothesis test, we experiment with three probing tasks.

Sentiment Analysis (IMDb). We reuse the IMDB sentiment classification setup from Figure 1. For each layer of `BERT-base-uncased`, we extract the average sentence embedding and train a linear probe to predict binary sentiment. We also train probes on k=20 random reinitializations of the model, and evaluate statistical significance using the hypothesis test described above. All probes are trained and evaluated on 1000 sentences with 10-fold cross-validation. Figure 4(A) reports (i) the average probe accuracy at each layer for the pretrained model, the randomized models, and a random guessing baseline, and (ii) the corresponding effect sizes relative to random guessing and to randomized models. While all pretrained layers outperform random guessing with large effect sizes, none are statistically distinguishable from the random reinitializations under the new test. Later layers, however, show a clear upward trend in effect size relative to randomized models.

Syntactic Structure (POS Tagging). We next assess token-level syntactic information using POS-tagging probes (Tenney et al., 2019). For each layer of `BERT-base-uncased`, we extract contextual token embeddings and train logistic regression probes on a subset of CoNLL-2003, one probe per layer that should work for all tokens and all POS tags. As above, we also train probes on k=20 random reinitializations and apply the same statistical test. Probes are evaluated with 10-fold cross-validation on 500 sentences. Figure 4(B) reports the layer-wise probe accuracy and effect sizes relative to a majority baseline and to randomized models. Consistent with prior work, POS accuracy peaks in middle layers (Tenney et al., 2019). However, when tested against randomized models rather than random guessing, only the middle layers remain statistically above chance, and the effect sizes are substantially reduced. This shows that testing against random computations eliminates many positive findings while still allowing for genuine positive discoveries where structure is robust.

World Models (Space and Time). Finally, we investigate the emergence of linear representations of space using the "world places" dataset from (Gurnee & Tegmark, 2024). Using `pythia-160m`, we extract average residual stream activations on each token of place names and train linear ridge regression probes to predict their geospatial coordinates (latitude and longitude). We compare the pretrained model against $k = 20$ baselines where transformer block weights are randomized while embeddings remain fixed. Probes are evaluated using R^2 scores with 10-fold cross-validation. Figure 4(C) reveals that raw embeddings (Layer 0) contain latent spatial structure ($R^2 \approx 0.12$), significantly outperforming random guessing ($Z \approx 100$). Passing these embeddings through randomized transformer blocks decreases linear readout ($R^2 \approx 0.38$). In contrast, the pretrained model's layers slightly improve this spatial linearity relative to the random baseline, suggesting that deeper layers progressively construct a more coherent spatial representation.

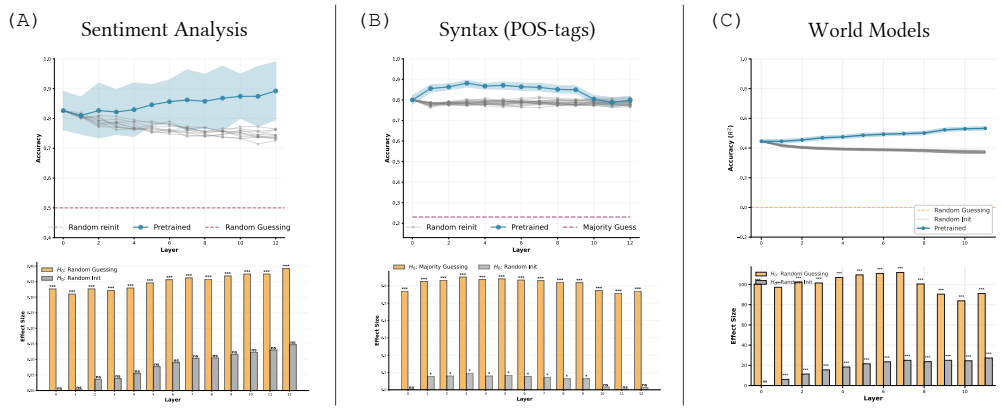

Figure 4. (A) Sentiment analysis experiment where probes on pretrained BERT are compared against probes trained on random computation. (B) Same experiment based on predicting syntactic labels (POS tags). (C) Reproducing the first experiment of Table 2 in (Gurnee & Tegmark, 2024), probing for indications of world models on pythia-160m.

By the final layers, the learned structure statistically surpasses the random baseline ($Z \approx 25$), confirming that the model eventually learns to encode space explicitly beyond the geometry inherent in the embeddings.

Method	Hypothesis space $\mathcal{E}$ (surrogate model)	Typical causal query $q(\mathfrak{C})$ and error criterion D
Performance benchmarking	Single scalar summarizing predictive performance (e.g., accuracy, calibration, perplexity).	*Observational query:* output score distribution under $P_{\mathbf{U}}$. Error: difference in expected performance metrics, e.g., $\lvert\mathbb{E}[f(\mathbf{U})] - \mathbb{E}[\hat{f}(\mathbf{U})]\rvert$.
Probing (linear / diagnostic classifiers)	Linear or shallow classifiers mapping internal activations to target variables (e.g., part-of-speech tags).	*Observational query:* conditional distribution $P(Y \mid \text{activations})$. Error: classification loss.
Feature attributions (saliency, SHAP, Integrated Gradients)	Input-level additive surrogates assigning contribution scores so that $f(x)$ is approximated by $\sum_i e_i(x)\, x_i$ relative to a baseline.	*Counterfactual queries:* local (additive) approximation of model behavior around inputs x. Error: fidelity loss between model predictions and surrogate reconstruction
Concept-based methods (e.g., TCAV, ACE, concept bottleneck models)	Surrogates mapping internal activations to interpretable concept variables and modeling f's dependence on them.	*Interventional queries:* model sensitivity or dependence on interpretable concept activations within latent space. Error: deviation between surrogate-predicted and model-predicted sensitivities (e.g., directional derivative mismatch).
Circuit discovery	Subgraphs of the computational graph representing causal mechanisms.	*Interventional queries:* outputs distribution under targeted ablations encoded by the circuit. Error: consistency in output distribution, e.g., $KL\big(P_{\mathfrak{C}}(Y \mid U) \,\|\, P_{\text{circuit}}(Y \mid U)\big)$
Causal tracing (patching, mediation analysis)	Scalar importance scores over units or connections inferred from intervention or mediation effects.	*Counterfactual mediation queries:* total, direct, or indirect effect of node V_i on target Y. Error: difference between predicted and empirical effects.
Causal abstraction / model-level alignment	High-level structural causal model with mappings from low-level network variables $\mathbf{V}$ to abstract variables $\mathbf{Z}$.	*Interventional invariance queries:* the actions of abstracting from V to Z and intervening should commute. Error: causal abstraction error, measuring causal alignment as violation of commutative properties of abstraction and intervention.

Table 1. Interpretability methods as instances of the statistical–causal framework of *surrogate models*. Each method specifies a hypothesis class $\mathcal{E}$, causal query family $q(\mathfrak{C})$, and associated error measure D quantifying how faithfully the surrogate answers the queries.