THE DEAD SALMONS OF AI INTERPRETABILITY: THE NEED FOR A STATISTICAL INFERENCE PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

A striking neuroscience study once placed a dead salmon in an fMRI scanner and showed it images of humans in social situations. Astonishingly, standard analyses 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 argue for a fundamental **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**. It also highlights important theoretical issues, such as the identifiability of explanations, which we argue is critical to understand the field's susceptibility to false discoveries. We illustrate this reframing with a toy scenario recasting probing as hypothesis testing against null distributions derived from random computation. The statistical–causal perspective opens many avenues for future work aiming to turn AI interpretability into a rigorous science.

1 INTRODUCTION

AI systems are increasingly deployed in high-stakes real-world settings, where their decisions directly affect human lives. In such contexts, interpretability is essential for transparency, accountability, and error diagnosis (Mehrabi et al., 2021; Barnes & Hutson, 2024; Ramachandram et al., 2025). Regulators have recognized these risks: under both the GDPR and the EU AI Act, explainability is now a legal requirement[1] (Lorè et al., 2023; Fresz et al., 2024; Kiseleva et al., 2022; Bengio et al., 2024).

Interpretability methods have already uncovered critical failure modes in safety-sensitive domains such as autonomous driving (Kim & Canny, 2017), medical decision support (Zech et al., 2018; Caruana et al., 2015), and large language models (Meng et al., 2022; Monea et al., 2024). They promise levers for mitigating bias and correcting errors (Arrieta et al., 2019; Kristofik, 2025), and they enable human–AI collaboration, where operators can meaningfully supervise, understand, and override automated decisions when needed (Felzmann et al., 2020; Kim et al., 2018). At the societal level, interpretability supports communication and trust: non-experts are more willing to accept, contest, or regulate AI systems when their behavior can be faithfully explained (Poursabzi-Sangdeh et al., 2021; Buçinca et al., 2021; for Economic Co-operation & Development, 2023).

Despite these urgent needs, our ability to explain modern AI systems remains in what may be called a *pre-scientific era*. A now-classic cautionary tale from neuroscience illustrates the danger of misapplied statistical inference. Researchers placed a dead salmon in an fMRI scanner, showed it photographs of humans, and ostensibly *asked* it to judge their emotions (Bennett et al., 2009). Standard pipelines identified several brain regions as predictive. The explanation, of course, was not supernatural cognition but a failure to correct for multiple comparisons.

[1] Regulation (EU) 2016/679, `https://eur-lex.europa.eu/eli/reg/2016/679/oj`; Regulation (EU) 2024/1689, `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`

In AI interpretability, "dead salmons" take the form of randomly initialized neural networks, and our field collects report of examples where plausible explanations are extracted from random networks. This phenomenon has been documented across feature attribution (Adebayo et al., 2018), probing (Ravichander et al., 2021), sparse autoencoders (Heap et al., 2025), and even causal approaches like circuit discovery or causal abstraction (Méloux et al., 2025; Sutter et al., 2025).

Beyond these striking false positives, interpretability methods also display statistical fragility. Small perturbations to the input (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 drastically change the explanations. In mechanistic interpretability, interventions on identified circuits often fail to generalize: edits that succeed in one context may break in another (Hoelscher-Obermaier et al., 2023). Conversely, multiple incompatible explanations can be found for the exact same behavior (Méloux et al., 2025; Dombrowski et al., 2019).

In this work, we synthesize evidence to argue that existing methods are particularly prone to *false discovery* with dead salmon artifacts as the most extreme and embarrassing instances. As road forward, we propose a unifying statistical-causal perspective: interpretability should be framed as a task of *statistical-causal inference*, where explanations are treated as parameters to be inferred from computational traces. This reframing yields the principled tools of *identifiability*, uncertainty estimation, and hypothesis testing. It also opens promising avenues for future work, such as Bayesian approaches to interpretability. To illustrate this perspective, we present a toy example showing how hypothesis testing against a null distribution of random computations can trivially eliminate dead salmon artifacts.

2 The Dead Salmons of AI Interpretability

Adebayo et al. (2018) showed that *saliency maps* can survive randomization and still look visually plausible. Several benchmarking efforts also demonstrated that many feature attribution methods perform little better than random (Kindermans et al., 2019; Nguyen et al., 2021). Complementing these negative results, Dombrowski et al. (2019) demonstrated that gradient-based explanations can be manipulated by small perturbations that leave predictions unchanged, also producing spurious attributions. From a theoretical perspective, Bilodeau et al. (2024) proved impossibility results showing that no attribution method can simultaneously satisfy intuitive desiderata across broad model classes. This formal perspective helps explain the empirical tendency toward false positives and instability.

Unsupervised concept-discovery pipelines such as sparse autoencoders (SAEs) (Cunningham et al., 2023; Yun et al., 2021) also display analogous pathologies. Heap et al. (2025) showed that SAEs recover apparently *interpretable* components even in randomly initialized transformers. Additional studies confirm that SAE-discovered features often fail to generalize (Heindrich et al., 2025; Kantamneni et al., 2025), hinting they may not capture fundamental properties of the representations computed by the neural network.

Probing (Adi et al., 2017; Conneau et al., 2018), an operationally simple and widely used method, suffers from related limitations. Probes ask whether a representation *contains* information, not whether the model *uses* it. Ravichander et al. (2021) show that probes can extract features merely encoded (e.g., inherited from embeddings) even if unused during inference. Defenses such as probe-capacity controls (Belinkov, 2022) mitigate but do not resolve this core issue: *predicting is not understanding* (D'Amour et al., 2022; Teney et al., 2022). Probing thus remains prone to false positives and dead-salmon style errors, where signals are detected without computational relevance.

Recent work has shifted toward intervention-based approaches (Vig et al., 2020; Meng et al., 2022; Goyal et al., 2019; Feder et al., 2021; Monea et al., 2024). These methods place models in counterfactual states, allowing researchers to isolate and characterize causal factors (Meng et al., 2022; Kramár et al., 2024; Monea et al., 2024), often through causal mediation analysis (Vig et al., 2020; Mueller et al., 2024). This line of work has coalesced into what is often termed *mechanistic interpretability* (MI): the effort to reverse-engineer the computation of complex neural networks into simple human-understandable algorithm (Olah et al., 2020). One of the main task concerns the identification of sparse subnetworks or "circuits" responsible for particular behaviors (Olah et al., 2020; Syed et al., 2023; Hanna et al., 2024b). A related strand relies on the theory of causal abstraction,

which formalizes when a high-level causal model validly represents a more complex system, namely, when their causal effects under interventions align (Chalupka et al., 2016; Rubenstein et al., 2017; Beckers & Halpern, 2019; Beckers et al., 2020; Geiger et al., 2022a). Applied to interpretability, the goal is to discover **high-level abstractions** that capture essential computational mechanisms of a network while remaining *simple* (Geiger et al., 2022a; Wu et al., 2023).

Yet these interventionist approaches also exhibit critical failure modes. Makelov et al. (2023) showed that subspace patching can produce *interpretability illusion* by activating alternate pathways rather than the hypothesized mechanism. Exhaustive studies on toy models reveal that both circuit discovery and causal abstraction often return multiple incompatible explanations, even for random networks, highlighting non-identifiability issues (Méloux et al., 2025). This is also true of larger networks, in which applying multiple circuit discovery methodologies yields inconsistent results (Zhang & Nanda, 2024). Finally, Sutter et al. (2025) proved that without the linear representation assumption, existing causal-abstraction methods will output explanations for random networks. Despite their grounding in causal reasoning, causal pattern analyses still suffer from false positives.

2.1 Discussion: statistical framing, non-identifiability, and uncertainty

The concrete pathologies documented above are not mere engineering quirks; they reflect a coherent statistical and philosophical diagnosis arising from the lack of proper grounding within rigorous inferential procedure. For predictive approaches like probing or SAE, their problem is inherited from standard issues in machine learning like overfitting (Belinkov, 2022) and underspecification (D'Amour et al., 2022; Teney et al., 2022), making them not generalize to settings where interventions are performed on the computations (Elazar et al., 2021). A predictive explanation method can predict a behavior without being a meaningful explanation for it (Ravichander et al., 2021). Solving this requires going beyond absolute predictive power towards relative predictive performance against meaningful computational alternatives (as the ones explored in Sec 4).

These observed problems remain puzzling for causal analysis methods, which, at first glance, appear to implement the causal inference framework at the heart of scientific rigor that should enable generalizable explanations. A likely explanation is **causal complexity**: AI systems are large, distributed computational systems with many interacting components, creating many causal pathways linking inputs to outputs. Philosophers and complex systems scientists have already argued that complex, nonlinear systems commonly admit many interacting pathways and context-dependent mechanisms, which makes it practically infeasible to single out one causal path (Potochnik, 2017), e.g., the *complex system problem (CSP)* in neuroscience (Tononi et al., 1994; Loosemore, 2012; Sarkar, 2022). Because of causal complexity, finding stories within complex computational systems becomes too easy (Lindsay & Bau, 2023).

This diagnosis, for both predictive and causal approaches, can be made more formal through the notion of *identifiability*, a statistical concept stating that the parameters of an explanatory statistical model are identifiable if they can be uniquely determined from observations (Casella & Berger, 2024). Identifiability is typically a prerequisite for inference in the natural sciences; without it, explanations remain ambiguous. Since identifiability is **not** a property of the system itself, but of the type of questions we ask before observing it, a lot of work has been dedicated to building identifiable models and queries in statistics (Casella & Berger, 2024; Allman et al., 2009), unsupervised learning (Locatello et al., 2019; Khemakhem et al., 2020), and causal inference (Shpitser & Pearl, 2008).

Currently, identifiability serves only as an analogy, since interpretability has not yet been consensually formalized as a problem of statistical inference. Thus, it cannot produce precise diagnoses of non-identifiability, and only observes the empirical consequences. These include explanations that fail to generalize across domains (Ravichander et al., 2021; Heindrich et al., 2025; Kantamneni et al., 2025) and incompatible explanations for the same behavior (Dombrowski et al., 2019; Méloux et al., 2025). All are symptomatic of a deeper problem: a tendency of current methods to produce false positive discoveries. The most striking of such false discoveries are dead salmon types of artifacts, when methods generate plausible-looking explanations from clearly meaningless computations. If such a formalism were adopted, explanations could be treated as parameters of a statistical model to be inferred from computational traces produced by the network. The statistical model would encode the assumptions that a given method makes about how explanations manifest

in those traces. Then, an interpretability query is identifiable if all candidate explanations can, in principle, be distinguished by some available data. We argue that this is the path we should take to advance interpretability research.

3 The Statistical Inference Perspective

Let $\mathbf{X}$ denote the random input to a neural network (e.g. tokens in language models or pixels in vision models). We represent the network as a directed acyclic graph (DAG) $f : \mathcal{X} \to \mathcal{Y}$, whose nodes correspond to computational units (neurons, channels, or other units of computation) and whose directed edges encode information flow determined by parameterized maps (weights).

Formally, let $\mathbf{V}$ be the set of nodes in the computation graph $\mathcal{G}$, including the designated output node(s) $Y \in \mathbf{V}$. Each node $v \in \mathbf{V}$ computes a measurable function of its parents $\mathrm{pa}(v) \subset \mathbf{V}$, parameterized by weights W_v:

$$h_v \;=\; g_v\big(\{h_u :\; u \in \mathrm{pa}(v)\},\; W_v\big), \qquad v \in \mathbf{V},$$

where inputs are treated as source nodes without parents. The network output(s) are given by the activation(s) at the designated output node(s).

Behavior of interest. We study the behaviour of the network relative to an input distribution $P_{\mathbf{X}}$, which specifies the task or environment of interest. Different interpretability tasks naturally correspond to different choices of $P_{\mathbf{X}}$.

Structural causal model. The pair (f, P_X) induces a structural causal model (SCM) in the sense of Pearl (Pearl, 2009; Pearl & Mackenzie, 2018). Concretely, we may represent the computation as an SCM $\mathfrak{C} = (\mathcal{G},\; \mathbf{V},\; \mathbf{X},\; f,\; P_{\mathbf{X}})$, where:

- *Endogenous variables:* $\mathbf{V}$ are the network's computational variables (hidden activations, outputs);
- *Exogenous variables:* the inputs $\mathbf{X}$ to the model, and the input distribution $P_{\mathbf{X}}$, which specifies the marginal distribution of the corresponding source variable(s);
- *Structural assignments:* the deterministic/parameterized maps g_v define the structural equations of the model;
- *Causal graph:* $\mathcal{G}$ is the network's computation graph.

This SCM entails a unique observational distribution over all computational variables $\mathbf{V}$. Sampling from it corresponds to drawing inputs from $P_{\mathbf{X}}$, executing a forward pass, and recording intermediate activations. Crucially, an SCM encodes not only this observational distribution but also a family of interventional and counterfactual distributions. This SCM view of neural network corresponds to the standard perspective taken by mechanistic interpretability (Olah et al., 2018; Cammarata et al., 2020; Geiger et al., 2022b; 2025).

Interventions and counterfactuals. For a subset of variables $\mathbf{I} \subseteq \mathbf{V}$, a *hard intervention* $\mathrm{do}(\mathbf{I} = \mathbf{i})$ replaces the structural assignments for variables in $\mathbf{I}$ with constant functions taking the value $\mathbf{i}$. The intervened SCM, denoted $\mathfrak{C}; \mathrm{do}(\mathbf{I} = \mathbf{i})$, induces the interventional distribution $P_{\mathbf{V}}^{\mathfrak{C};\mathrm{do}(\mathbf{I}=\mathbf{i})}$, also defined over $\mathbf{V}$.

Counterfactual distributions in an SCM are carried out in two standard steps (abduction, intervention):

1. *Abduction:* condition on observed evidence $\mathbf{E} = \mathbf{e}$ and update beliefs about the exogenous variables via the posterior $P_{\mathbf{X}|\mathbf{E}=\mathbf{e}}$;
2. *Intervention:* modify the model by applying an intervention $\mathrm{do}(\mathbf{I} = \mathbf{i})$;

The abduced and intervened model $\mathfrak{C} \mid \mathbf{E} = \mathbf{e}; \mathrm{do}(\mathbf{I} = \mathbf{i})$ induces the counterfactual distribution over $\mathbf{V}$: $P_{\mathbf{V}}^{\mathfrak{C}|\mathbf{E}=\mathbf{e};\mathrm{do}(\mathbf{I}=\mathbf{i})}$. In many neural-network settings, when we observe some evidence, we typically know exactly which single input instance $\mathbf{x}$ is observed, so the abduction step is trivial (conditioning on $\mathbf{X} = \mathbf{x}$ fixes the corresponding source node). A counterfactual query also reduces to an interventional query analyzing the model's behavior on a single input instance.

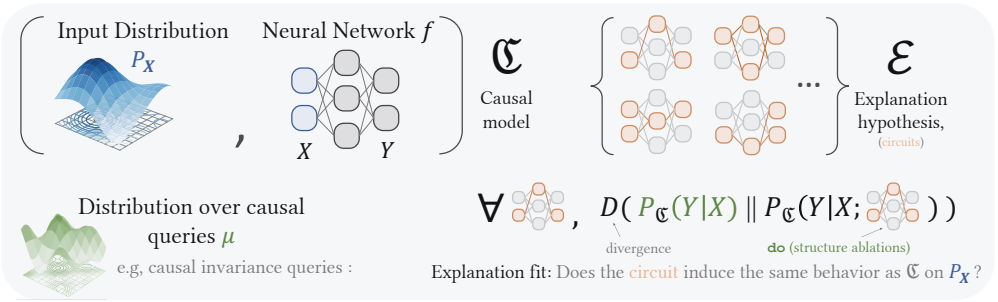

Figure 1: **Illustration of the statistical-causal perspective.** The input distribution and the neural network form an SCM $\mathfrak{C}$ from which it is possible to draw causal queries (properties of its observational, interventional, and counterfactual distributions). An interpretability task is a choice of causal queries to explain (represented by a distribution μ over queries), a hypothesis space $\mathcal{E}$ of surrogate models, the candidate explanation, and a measure of error of candidate explanations on selected causal queries.

Causal query. A *causal query* $q(\mathfrak{C})$ is any measurable functional or distribution that can be evaluated from the SCM (possibly after conditioning or intervention). Examples include: the marginal distribution of $\mathbf{V}$; the effect on the output distribution of ablating a set of neurons; or the counterfactual output for a specific input under a surgical change to the computation. This definition aligns with the standard notion of estimands in causal inference: classical causal estimands (average treatment effects, conditional causal effects, etc.) are causal queries.

3.1 Explanations as surrogate models for query answering

We formalise explanations as *surrogate models* designed to answer a chosen collection of causal queries. Suppose we have a measurable space of queries $(Q, \mathcal{F})$ together with a probability measure μ on Q (the *query distribution*) that encodes which queries are important or likely to be asked for the interpretability task at hand. Classical causal inference corresponds to concentrating μ on a single query; interpretability addresses the task of answering many queries drawn from a nontrivial μ by a surrogate model. The framework is visually illustrated for circuit discovery in Fig. 1.

Let $\mathcal{E}$ be a hypothesis class of candidate surrogates (circuits, probes, feature-maps, small causal models, ...). Each surrogate $e \in \mathcal{E}$ defines a *query-answering map* S_e, so that $S_e(q)$ is the surrogate's predicted answer for query q.

To measure fidelity of the explanation (the surrogate model), we can, for example, choose a divergence D between probability measures (total-variation, Kullback–Leibler where defined, Wasserstein, or a suitable MMD) and define the population risk by

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

The interpretability task is defined by the tuple $(\mu, \mathcal{E}, D)$ which specifies the query distribution, the hypothesis space of surrogate models, and the error measurement D. We call the explanatory task *identifiable* if L_μ admits a unique minimiser in $\mathcal{E}$ (uniqueness may be relaxed to hold only up to *allowed* symmetries of the surrogate class). Identifiability captures whether the chosen surrogate class can, in principle, be distinguished using the queries deemed relevant by μ.

Estimation. In practice, we only observe a finite set of queries $q_1, \ldots, q_n \overset{\text{i.i.d.}}{\sim} \mu$, and for each query we collect a finite number of computational traces obtained by sampling inputs and executing the model. Denote by T_n the resulting finite dataset of traces. An interpretability method M acts as a statistical estimator that maps T_n to an estimated surrogate $\hat{e} := M(T_n)$. A natural approach with the above framing is to search for the explanation $e \in \mathcal{E}$ that minimizes the empirical risk $\widehat{L_\mu}(e)$. Finite-sample uncertainty arises both from sampling queries and from the limited number of traces per query. The intuitive notion of *consistency* of such estimators (convergence of $\hat{e}$ in $\mathcal{E}$ as the amount of data grows) is a question of both statistical consistency and computational tractability: for instance, exhaustively testing all possible interventions is infeasible because the intervention space

is (in general) infinite. A relevant question is then how many intervention queries are needed to be confident that a faithful circuit has been discovered (Hanna et al., 2024a).

Interpretation of common methods. We provide in Table 1 a high-level description of how common methods could be reframed within the statistical-causal inference framework.

Method	Hypothesis space $\mathcal{E}$	Causal queries and error measurement D
Benchmarking	Single numerical statistic (e.g. accuracy) representing performance.	Observational query: output distribution under $P_{\mathbf{X}}$; error measured by difference between estimated accuracy and true accuracy (standard parameter inference in statistical inference).
Probes	Simple classifiers (often linear models) mapping hidden activations to labels.	Observational queries: conditional distribution of labels given hidden activations; error measured by classification loss or divergence between predicted and true conditional distributions.
Feature attribution / saliency	Attribution scores or saliency maps over inputs.	Counterfactual queries: effect of perturbing/intervening on input features on outputs; error measured by deviation of surrogate scores from actual effect of interventions. We note that Senetaire et al. (2023) already propose a framing of feature attributions as statistical inference.
Circuit discovery	Subgraphs of the original computation graph connecting inputs to outputs.	Interventional invariance queries: behavior under ablations of parts of the computation; error measured by discrepancy between outputs of surrogate circuit and full model (e.g. predictive accuracy or divergence).
Causal tracing	Numerical score of importance for the target computational unit.	Counterfactual queries of the type of mediation analysis: natural direct effect of a component when mediating the causal link between inputs to outputs; error measured by the difference between true and estimated direct effects.
Causal abstraction	High-level causal models (abstract variables and mappings from low-level variables $\mathbf{V}$).	Interventional invariance queries: preservation of distributions that are valid under abstraction mapping; error measured by divergence between abstract model predictions and true interventional distributions.

Table 1: Common interpretability methods expressed in the framework of surrogate models: hypothesis space, causal queries, and error measurement.

Takeaway. Viewing interpretability as the construction of a surrogate that approximates a distribution of causal queries makes explicit the (often implicit) choices practitioners must make: which queries matter (the measure μ), how closely they must be matched (the divergence D), and what should the surrogate be (the hypothesis class $\mathcal{E}$ and any potential regulariser). This perspective

brings the standard apparatus of statistical and causal inference (estimands, identifiability, risk, finite-sample error, ...) into interpretability research. Conceptually, it highlights that explanation is a form of *computational compression*: the surrogate compresses some aspects of the network's computation of interests, the ones encoded by the query distribution μ.

4 Toy Illustration: Probing as Hypothesis Testing

The formalism above naturally yields an empirical risk objective

$$\arg\min_{e} L_\mu(e) + \Omega(e),$$

as an explanation discovery strategy, where $\Omega(e)$ is a regulariser that can favor simpler explanations. Empirical risk minimization (ERM) may identify the best-fitting surrogate within a class $\mathcal{E}$, but by itself it does not establish whether the fit of the discovered explanation is meaningful. Hypothesis testing provides a principled tool for assessing this question, offering a way to control false discoveries by comparing against relevant null hypotheses.

A small empirical loss may arise spuriously, for example, because the architecture or trivial correlations allow many surrogates to perform well, creating an illusion of explanatory validity. In striking examples of dead salmon artifacts, explanations with a good fit are found on random networks.

This is analogous to the well-known "dead salmon" problem, where statistical procedures spuriously detect meaningful signals in random data. To guard against such false discoveries, one can test explanations against null models that randomize the network computation. This measures not only whether the explanation produces a good fit, but if it fits better than what would be expected from attempting to explain random computation. Concretely, we can test whether the discovered explanation $e^\star = M(\mathfrak{C})$ predicts the *true network's* answers to a chosen set of queries Q more reliably than explanations derived from randomized counterparts of $\mathfrak{C}$.

Illustration: probing as hypothesis testing. Consider the task of assessing whether a neural network has **learned** to detect loops when classifying handwritten digits. A common probing approach is to train a classifier on the network's activations and declare that the concept is learned if the probe predicts the presence of loops above random guessing on test images. However, even a randomly initialized network, implementing an arbitrary but fixed transformation of the input, can contain enough information to support above-chance probe accuracy.

The formalism developed above allows us to recast probing as a hypothesis test against a null distribution of random computations. Let V_ℓ denote the representation at layer ℓ. A probe trained on $\mathcal{D}_{\text{train}}$ yields a classifier $\widehat{e} = \mathsf{Probe}(V_\ell; \mathcal{D}_{\text{train}})$. Define the population accuracy

$$S(\widehat{e}; \mathfrak{C}) \;=\; \Pr_{(X,Y)\sim P_{XY}} \left[\widehat{e}(V_\ell(X)) = Y\right],$$

and its empirical estimate on a test set $\mathcal{D}_{\text{test}}$,

$$\widehat{S}(\widehat{e}; \mathfrak{C}) \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\widehat{e}(V_\ell(x_i)) = y_i\}.$$

For each randomized network $\widetilde{\mathfrak{C}}^{(b)}$, we repeat the probing procedure to obtain $\widehat{e}^{(b)}$ and evaluate $\widehat{S}^{(b)} = \widehat{S}(\widehat{e}^{(b)}; \widetilde{\mathfrak{C}}^{(b)})$ on the same $\mathcal{D}_{\text{test}}$. The one-sided Monte Carlo p-value (alternative: the true representation yields higher accuracy than randomized computations) is

$$\widehat{p}_{\text{probe}} \;=\; \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left\{\widehat{S}^{(b)} \geq \widehat{S}(\widehat{e}; \mathfrak{C})\right\}}{B+1}.$$

This quantifies the likelihood of achieving the observed probe accuracy under randomized computation. A low p-value provides evidence that the network representation genuinely encodes the probed concept. Many different randomization procedures can be considered (full model random initialization, layer-specific randomization, weight shuffling, ...). By design, explaining random networks will fall within the null distribution and yield a high p-value, eliminating dead salmon issues.

Toy example. We illustrate with a three-layer random MLP (128 neurons per layer) on MNIST digits. The explanatory task is to classify whether digits contain a loop using probe classifiers trained and tested on balanced subsets of data. Standard probes achieve above-chance accuracy at all layers, which is an instance of a dead salmon artifact. Importantly, reporting only the variability due to finite-sample uncertainty of the test set (e.g., 95% confidence intervals via bootstrap resampling of the test set) is insufficient: the probe accuracy exceeds random chance with high confidence. In contrast, the hypothesis test against randomized computation correctly fails to reject the null, yielding large p-values and eliminating false positives.

5 Discussion

We have argued for a statistical–causal perspective on post-hoc interpretability and introduced one preliminary instantiation of such a framework. While alternative formalizations are possible, the present approach illustrates the potential of grounding interpretability in rigorous statistical inference. This perspective can shift interpretability research from a set of heuristics toward one of principled hypothesis testing, thereby opening a broad research agenda. A natural next step is to reformulate other interpretability methods within this statistical–causal framework, clarify the theoretical conditions under which particular explanations are identifiable and understand their consistency properties. The simple hypothesis tests against randomized computations demonstrate how to eliminate embarrassing "dead salmon" artifacts and can, more generally, reduce the false discoveries so prevalent in our field.

Layer	Probe acc.	p-value (ours)
1	0.85 (0.81, 0.89)	0.47
2	0.82 (0.77, 0.87)	0.52
3	0.79 (0.71, 0.86)	0.55

Table 2: Comparison of standard probes versus our hypothesis test on a 3-layer random MLP. Standard probes falsely detect the target concept, while our test eliminates false discoveries.

Several challenges and opportunities arise from this perspective Senetaire et al. (2023). First, there is a need to design interpretability tasks with improved identifiability guarantees. Second, estimation strategies must be developed that are computationally tractable while offering satisfactory statistical properties (e.g., bias, variance, and consistency). Third, the specification of null distributions is particularly important to target different scientific questions and forces us to state the alternative hypotheses explicitly. Finally, the framework admits a natural Bayesian extension: priors over explanations can be informed by previously extracted knowledge, enabling interpretability to accumulate results in a systematic and principled manner rather than through isolated case studies. Interestingly, Bordt et al. (2025) argued in favor of viewing explanations as a descriptive statistical task of function (e.g., neural networks).

More broadly, interpretability research can benefit from methodological lessons in adjacent sciences. Neuroscience, for example, has confronted analogous challenges in the study of complex, distributed systems, and its methodological hardships (Poldrack, 2006; Mehler & Kording, 2020; Jonas & Kording, 2017), like the anecdote of the dead salmon, can continue to inform our research. However, we must also acknowledge that identifiability *might* be fundamentally out of reach for many interpretability queries. Complex systems often resist simplification, and some systems can only be *understood* by step-by-step simulations (Arthur, 2009; Israeli & Goldenfeld, 2004; Strogatz, 2001). Even if we cannot reduce the causal complexity of AI systems, adopting a rigorous statistical framing enables us to chart the boundary between what can and cannot be explained, and to study interpretability queries systematically under these constraints.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJh6Ztuxl.

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

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. URL https://arxiv.org/abs/1910.10045.

W Brian Arthur. Complexity and the economy. In *Handbook of Research on Complexity*. Edward Elgar Publishing, 2009.

Emily Barnes and James Hutson. 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.

Sander Beckers and Joseph Y. Halpern. 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.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 606–615. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/beckers20a.html.

Yonatan Belinkov. 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.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, et al. International scientific report on the safety of advanced ai (interim report). *arXiv preprint arXiv:2412.05282*, 2024.

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

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

Sebastian Bordt, Eric Raidl, and Ulrike von Luxburg. Position: Rethinking explainable machine learning as applied statistics. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=b2gM1HyAgE.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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.

George Casella and Roger Berger. *Statistical inference*. CRC press, 2024.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 361–369, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/chalupka16.html.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 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.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(1), January 2022. ISSN 1532-4435.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 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.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, June 2021. doi: 10.1162/coli_a_00404. URL https://aclanthology.org/2021.cl-2.13.

Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and engineering ethics*, 26(6):3333–3361, 2020.

Organisation for Economic Co-operation and Development. *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*. OECD Publishing, 2023.

Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F. Huber, and Christian Horz. How should ai decisions be explained? requirements for explanations from the perspective of european law. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:438–450, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31648. URL http://dx.doi.org/10.1609/aies.v7i1.31648.

Atticus Geiger, Zhengxuan Wu, Karel D'Oosterlinck, Elisa Kreiss, Noah D. Goodman, Thomas Icard, and Christopher Potts. Faithful, interpretable model explanations via causal abstraction. Stanford AI Lab Blog, 2022a. URL https://ai.stanford.edu/blog/causal-abstraction/.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (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.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2025. URL https://arxiv.org/abs/2301.04709.

Amirata Ghorbani, Abubakar Abid, and James Zou. 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.

Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *CoRR*, abs/1907.07165, 2019. URL http://arxiv.org/abs/1907.07165.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms, 2024a. URL https://arxiv.org/abs/2403.17806.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*, 2024b.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Sparse autoencoders can interpret randomly initialized transformers, 2025. URL https://arxiv.org/abs/2501.17727.

Lovis Heindrich, Philip Torr, Fazl Barez, and Veronika Thost. Do sparse autoencoders generalize? a case study of answerability, 2025. URL https://arxiv.org/abs/2502.19964.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark, 2023. URL https://arxiv.org/abs/2305.17553.

Navot Israeli and Nigel Goldenfeld. Computational irreducibility and the predictability of complex physical systems. *Physical review letters*, 92(7):074105, 2004.

Eric Jonas and Konrad Paul Kording. Could a neuroscientist understand a microprocessor? *PLOS Computational Biology*, 13(1):1–24, 01 2017. doi: 10.1371/journal.pcbi.1005268. URL https://doi.org/10.1371/journal.pcbi.1005268.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing, 2025. URL https://arxiv.org/abs/2502.16681.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo P. Monti, and Aapo Hyvärinen. 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.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

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

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *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.

Anastasiya Kiseleva, Dimitris Kotzinos, and Paul De Hert. Transparency of ai in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. *Frontiers in artificial intelligence*, 5:879603, 2022.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

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

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

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Ratsch, Sylvain Gelly, and Olivier Bachem. 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.

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

Filippo Lorè, Pierpaolo Basile, Annalisa Appice, Marco de Gemmis, Donato Malerba, and Giovanni Semeraro. An ai framework to support decisions on gdpr compliance. *Journal of Intelligent Information Systems*, 61(2):541–568, 2023.

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

David Marc Anton Mehler and Konrad Paul Kording. The lure of misleading causal statements in functional connectivity research, 2020. URL https://arxiv.org/abs/1812.03363.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. A glitch in the matrix? locating and detecting language model grounding with fakepedia, 2024. URL https://arxiv.org/abs/2312.02073.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability, 2024. URL https://arxiv.org/abs/2408.01416.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26422–26436. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/de043a5e421240eb846da8effe472ff1-Paper.pdf.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

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

Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

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

Angela Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, Chicago, 2017.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–52, 2021.

Dhanesh Ramachandram, Himanshu Joshi, Judy Zhu, Dhari Gandhi, Lucas Hartman, and Ananya Raval. Transparent ai: The case for interpretability and explainability, 2025. URL https://arxiv.org/abs/2507.23535.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 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.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. ID 11, August 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf. *equal contribution.

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

Hugo Henri Joseph Senetaire, Damien Garreau, Jes Frellsen, and Pierre-Alexandre Mattei. Explainability as statistical inference. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (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.

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

Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (studies in nonlinearity)*, volume 1. Westview press, 2001.

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

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery, 2023. URL https://arxiv.org/abs/2310.10348.

Damien Teney, Maxime Peyrard, and Ehsan Abbasnejad. Predicting is not understanding: Recognizing and addressing underspecification in machine learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, 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.

G Tononi, O Sporns, and G M Edelman. 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.

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

Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. Ms., Stanford University, 2023. URL https://arxiv.org/abs/2305.08809.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 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.

Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cedric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. 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.

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

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Hanwei Zhang, Felipe Torres Figueroa, and Holger Hermanns. Saliency Maps Give a False Sense of Explanability to Image Classifiers: An empirical evaluation across methods and metrics. In Vu Nguyen and Hsuan-Tien Lin (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.