Three Desiderata for Faithfulness in Machine Learning Explanations: The Case for Causal Abstraction

Mette Friis Andersen, Maria Heuss, Ana Lucic
University of Amsterdam
mette.andersen@student.uva.nl, {m.c.heuss, a.lucic}@uva.nl

Abstract

Faithfulness is a broadly agreed-upon desideratum for explanations of machine learning model predictions. While many different methods have been adopted by the community, there is no agreed-upon definition of faithfulness. Here, we propose three desiderata for faithfulness beyond the standard intuition of accurately representing the reasoning process of the model, related to (1) enabling reverse-engineering of specific behaviors, (2) capturing interventionist causal relations, and (3) achieving an appropriate model decomposition. We argue that causal abstraction satisfies these, and provides a framework for evaluating faithfulness claims in the community.

1 Introduction

The field of explainable AI (XAI) aims to address the issue of making predictions from machine learning (ML) models more transparent. One of the main issues in XAI is that we need to make sure our explanations are *faithful*, broadly understood as "accurately representing the reasoning process of the model" [1; 2]. Previous work has surveyed XAI methods with respect to their faithfulness [3] without specifying exactly what we mean by faithfulness beyond this standard intuition. Moreover, Saphra and Wiegreffe [4] state that we need to "ground our empirical work in precise vocabulary", the lack of which creates "duplicated research efforts and limits shared knowledge". In their recent paper, Williams et al. [5] motivate the need for a philosophical grounding of mechanistic interpretability concepts. We answer their call in two ways. We first show how faithfulness is related to various desiderata of explanation, focusing on (1) enabling reverse-engineering, (2) capturing interventionist causal relationships and (3) decomposing the model at an appropriate level of granularity. While prior work has considered disambiguating such terms from faithfulness as "out of scope" [3], we contribute to initial efforts [6] on disambiguating such terms and show how these desiderata relate to faithfulness. We focus on these concepts because of their relevance in the context of model improvement and debugging, the most common use cases for XAI methods [7]. We also motivate an existing mechanistic interpretability framework, causal abstraction [8], as a method for measuring faithfulness, by showing that it satisfies our desiderata.

2 Desiderata for faithfulness

2.1 Desideratum 1: Enables reverse-engineering of specific model behavior(s)

Reverse-engineering is a desideratum for faithfulness because a faithful explanation of how a model produces an output should enable us to modify that behavior [9]. In particular, we are interested in ensuring we can modify *undesirable behaviors*. If an explanation does not provide this capability, then we lack the understanding for correcting undesirable behaviors. We note that by requiring

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

explanations that enable reverse-engineering, this does *not* require that it be intuitively understandable by humans. While human interpretability is important [3; 10; 11], it is distinct from faithfulness since a human-like reasoning process does not necessarily capture the reasoning process of the model [1; 12]. Nauta et al. [3] illustrates this as follows: "When the machine learning model is trained on flawed data, it learns nonsensical relations, which are in turn shown by the explanation. The explanation might then be perceived as being wrong, although it is truthfully reflecting the model's reasoning". Faithfulness requires accuracy about what the model actually does rather than conformity to human expectations of what it should do.

2.2 Desideratum 2: Captures (interventionist) causal relations

Having established reverse-engineering as a desideratum for faithfulness, it follows that our explanations must also support causal intervention. This is achieved via a causal explanation *in the interventionist sense* [13], as opposed to a causal explanation in the regularity-theorist sense [14; 15]. This strict demarcation is important, since what it means for an explanation to be *causal* is ambiguous in the literature. Saphra and Wiegreffe [4] define a cause using a regularity-theorist conception: "In a causal model, a causal mechanism is a function—governed by "lawlike regularities" (Little, 2004) — that transforms some subset of model variables (causes) into another subset (outcomes or effects)". Under this view, identifying stable correlations suffices for a causal explanation. In contrast, an interventionist conception of a cause C requires that C causes E if and only if intervening on C (*ceteris paribus*), produces a change in E [13]. This definition is counterfactual, manipulability-based and particularly suited for model improvement purposes, that is, in cases where we are interested in bringing about a change in E by exploiting the causal relation. Since a regularity theorist conception can be satisfied without yielding an insight into the reasoning process of the model, a faithful explanation should be a causal *interventionist* explanation.

2.3 Desideratum 3: Achieves appropriate decomposition

Achieving an appropriate level of decomposition of model parts is an important open problem for XAI methods in mechanistic interpretability [5]. A complete account of all low-level details of the model may be maximally faithful to the model's behavior but does not necessarily constitute an *interpretable* explanation of the model prediction. We define *appropriate decomposition* as a decomposition that captures the model's causal structure at a semantically meaningful level of abstraction. As Geiger et al. [16] argue: "For explanations that can engage with these questions ["Is the model robust to specific kinds of input", "Does it treat all groups fairly?", and "Is it safe to deploy?"], we need methods that are provably faithful to the low-level details but stated in higher-level conceptual terms". Therefore, a faithful explanation must go beyond equating the explanandum with the explanans.

Appropriate decomposition is also needed to capture the causal relations leveraged by the model. For example, in SAEs, the dictionary size is a hyperparameter that influences the chosen granularity level [17]. If the dictionary size is too small, then the SAE will project the features into an under-specified subspace, preventing full disentanglement of model components. As a result, interventions on these features will not cleanly map to interventions in the base model. If the dictionary size is too large, then this can result in features that capture finer-grained details rather semantically meaningful concepts.

Empirical studies have shown that individual neurons are insufficient units for encoding disjunctive concepts [18; 19; 20; 21; 22; 8; 23]. Some XAI methods assume that features are linearly separable from the activations via linear transformations [24; 25; 26; 27]. However, as Geiger et al. [8] emphasize, evaluating faithfulness should not depend on such structural assumptions about the model's internal reasoning process. Optimizing for an appropriate decomposition should therefore be treated as integral to optimizing for the faithfulness of explanation.

3 Existing definitions of faithfulness and their limitations

We will review two prominent directions for measuring faithfulness: causal scrubbing [28] and Jacobian matching [29], and highlight why these capture our desiderata insufficiently.

Jacobian matching. First, we consider Jacobian matching [29], which measures the faithfulness of transcoders: sparse autoencoders (SAEs) which take in a layer's input and predict its output. The main idea is to penalize differences between the Jacobian of the original model and the Jacobian of the

transcoder ($\|J_{\text{orig}} - J_{\text{transcoder}}\|_F^2$) [29] in order to force the transcoder to learn the actual mechanisms the original model is using. The authors devise a toy setup where they train an MLP on data with repeated examples, then train a transcoder to approximate the MLP. The transcoder is able to match the MLP's outputs with low error, but it achieves this through a different mechanism: it develops a "memorization" feature which activates on the repeated example, even though the MLP does not have such a feature. However, when they use Jacobian matching, this memorization feature disappears, indicating that Jacobian matching is a useful tool for deterring the transcoder from some unfaithful behavior, but it is still insufficient for ensuring faithful model mechanisms. This is because the method is vulnerable to gradient masking where the transcoder can learn to manipulate its gradients by creating features with large weights, and therefore large gradients, but with very small, negative biases such that the overall feature is barely active. In addition, Jacobian matching does not satisfy our three desiderata from Section 2, since encouraging matching gradients does not guarantee (i) correct internal mechanisms are identified by the transcoder, or (ii) causal equivalence under interventions. It also lacks appropriate decomposition, because aligning Jacobian matrices does not ensure the learned features correspond to semantically meaningful concepts.

Causal Scrubbing. Causal scrubbing [28] is a method for testing weather a hypothesized mechanistic explanation, represented as a computational graph, i.e., a *circuit*, faithfully describes how a model works. It is based on first identifying a potential circuit: which components are involved, how information flows between them, and which information is causally relevant at each step. Then, we systematically ablate parts of the model that are irrelevant to the circuit, while preserving information the circuit deems causally important. If the irrelevant parts were indeed unnecessary, then there should be no impact on the resulting prediction. The faithfulness of a circuit is defined as how well the model preserves its original behavior on a dataset after ablating everything except the hypothesized circuit. While causal scrubbing shares some similarities with causal abstraction, there are some important differences. In particular, causal scrubbing tests circuits at a single level of abstraction: it validates whether low-level components implement a proposed mechanism, but does not enforce correspondence or similarity between high-level algorithmic descriptions and low-level circuit implementations. Therefore, it is possible to have a circuit verified by causal scrubbing, which does not align with higher-level interpretations of model behavior. However, causal scrubbing does not satisfy our three desiderata from Section 2 because (i) it only supports reverse-engineering if the hypothesized circuit is correct, (ii) it operates at a fixed level of abstraction and does not guarantee causal equivalence across different levels.

4 Causal abstraction as a framework for evaluating faithfulness

In the previous section, we argued that reverse-engineering motivates our definition of faithfulness: understanding a model's behavior requires capturing its causal mechanisms via interventionist analysis, which in turn requires a decomposition that reflects those causal relations at the appropriate level of granularity. In this section, we argue that causal abstraction [8] provides a framework for generating model explanations that satisfy the desiderata we have outlined.

According to Icard [30], evaluating explanations of model behavior involves three main steps:

1. Construct the low-level model $\mathcal{L}$ as a causal system in a given language. $\mathcal{L}$ is the explanandum.
2. Create the candidate high-level model $\mathcal{H}$ obtained using one of our explainability methods captured in the language. $\mathcal{H}$ is the explanation of the low-level model, and is referred to as an *abstraction*.
3. Specify the relation between them, and whether that relation has such characteristics that it can be described as a *causal consistency-preserving relation* $\mathcal{L} \to \mathcal{H}$. A relation is causal consistency-preserving if and only if interventions in the low-level model $\mathcal{L}$ commute with interventions in the abstraction $\mathcal{H}$ (see Figure 1).

In practice, the high-level model $\mathcal{H}$ can be obtained by either merging or marginalizing variables of the low-level model [8], or by applying linear transformations to disentagle polysemantic neurons (for instance, using SAEs). The hypothesis for $\mathcal{H}$ can also be generated using various XAI methods.

A method satisfying the reverse-engineering desideratum allows us to pinpoint and debug undesirable model behaviors. In causal abstraction, this is achieved by constructing a high-level model $\mathcal{H}$ that is causally consistent with the low-level model $\mathcal{L}$, thereby allowing us to identify which low-level

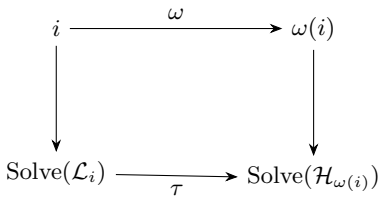

Figure 1: The causal relations of the low-level model $\mathcal{L}_i$ are captured by the high-level model $\mathcal{H}_{\omega(i)}$. τ is the mapping of total configurations of the low-level system, and its correspondence in the high-level system, given its nodes and relations. ω is the mapping of interventions on the low-level model to the high-level model. This is formalized as: $\tau\big(\mathrm{Solve}(\mathcal{L}_i)\big) = \mathrm{Solve}\big(\mathcal{H}_{\omega(i)}\big)$ [31].

attributes correspond to which high-level concepts, effectively localizing the mechanisms responsible for specific behaviors. This causal consistency-preserving mapping allows us to identify which behaviors we want to understand from $\mathcal{H}$, and understand how they are implemented in $\mathcal{L}$.

According to causal abstraction, the faithfulness of a high-level explanation with respect to the low-level model is measured by how well the explanation captures the causal mechanisms of the model, measured by the degree to which *interventions in the high-level model commute with interventions in the low-level model* [8]. Ideally, intervening in the low-level model and then abstracting should produce the same result as abstracting first and then intervening in the high-level model. This definition respects the interventionist definition of causality and is formalized as:

$$\epsilon(\tau) = \sup_{i} \left\| \tau\big(\mathrm{Solve}(\mathcal{L}_i)\big) - \mathrm{Solve}\big(\mathcal{H}_{\omega(i)}\big) \right\|, \tag{1}$$

where $\mathcal{L}$ is the low-level model, $\mathcal{H}$ is the high-level model, τ is the abstraction map that transforms the low-level model into a high-level model, i is an intervention applied to the low-level model, and $\omega(i)$ is the corresponding intervention in the high-level model. Solve is the output behavior of the low-level or high-level model under these interventions. $\|\cdot\|$ is a norm measuring the distance between outcomes, and $\sup_\iota$ denotes the supremum (maximum) over all valid interventions.

For example, consider an SAE which learns a set of sparse latent features that can be treated as candidate high-level variables used to hypothesise a causal model $\mathcal{H}$. The aligned features in the low-level network are then taken to be the neurons most strongly associated with that SAE latent variable. If interventions on the variables in the high-level model $\mathcal{H}$ fail to commute with the variables in the low-level model $\mathcal{L}$ under interchange interventions, then the explanation is unfaithful.

Due to Geiger et al. [16], we show another example in Figure 2. Imagine we have the low-level model $\mathcal{L}$ and hypothesize the higher-level model $\mathcal{H}$. $\mathcal{L}$ adds three numbers together, and we hypothesize that it does so by first adding two number together, resulting in one sum, and adding the final number to this sum. For each variable, we hypothesize a mapping τ: $\mathcal{L} \rightarrow \mathcal{H}$. Assume we want to test whether the high-level variable $\mathcal{H}_1$ abstracts the low-level variable $\mathcal{L}_1$ on the toy-data set consisting of {[1,3,5],[4,5,6]}. We can repeat this process for all variables as specified by τ . We first run the high-level model on our data, and save the activations. We get $\mathcal{H}_1 = 4$ and output 9 for our input [1,3,5], and $\mathcal{H}_1 = 9$ and output 15 for our second input [4,5,6]. Imagine we patch $\mathcal{H}_1$ (intervene on variable $\mathcal{H}_1$), such that $\mathcal{H}_1 = 9$. Given input [1,3,5], we get 14 as expected. We hypothesize that $\mathcal{L}_1$ is captured by $\mathcal{H}_1$. We test this by first running the full low-level model on the two inputs, and get the same output as the non-patched high-level model, 9 and 15. Then we patch the activation at $\mathcal{L}_1$ by the same value as the corresponding high-level variable, so $\mathcal{L}_1 = 9$. If we get the same output as a result of the patching (14), the we have a piece of evidence that the variables are performing the same causal function. We do this for all variables, across all inputs, and achieve a final faithfulness score measuring the extent to which the model $\mathcal{H}$ respects the causal structure of model $\mathcal{L}$. We note that, in principle, this allows for multiple explanation models whose interventions commute well with the underlying model, but carve the model up in different ways. Causal abstraction does not claim that there is one true explanation.

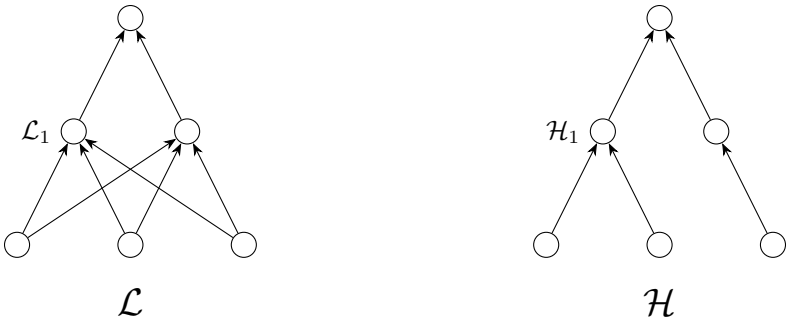

Figure 2: Illustration of causal abstraction: a low-level model $\mathcal{L}$ adding three numbers is mapped to a hypothesized high-level model $\mathcal{H}$ that sums two numbers first, then adds the third. For each variable, a mapping τ is defined. Activations are recorded for toy inputs, and patching interventions test whether high-level variables replicate the causal effects of their low-level counterparts. Matching outputs after patching provides evidence of causal alignment, and repeating this for all variables yields a faithfulness score for $\mathcal{H}$ relative to $\mathcal{L}$.

5 Critiques of causal abstraction as a framework for faithfulness

Casual abstraction does not identify the one true model explanation. Méloux et al. [32] argue that causal abstraction can be too permissive, as it permits multiple explanations for the same model behavior. The authors pose that this is problematic, particularly because it permits two *conflicting* explanations. Using the authors' definition, two explanations can be deemed conflicting even though they would be considered compatible according to the standard criteria in the philosophy of science (see [33]), where two theories can explain the same body of evidence, diverging only with respect to epistemic virtues, such as parsimony. In Méloux et al.'s definition, these epistemic virtues are not kept constant, so the explanations need not be conflicting. Therefore, we do not require the existence of a single, unique explanation for a particular behavior.

Causal abstraction may be too permissive and not sufficient to measure faithfulness. Sutter et al. [34] claim that without assumptions on how models encode information, causal abstraction cannot reliably produce faithful explanations. Complex non-linear abstractions can overfit on a dataset, achieving high scores while being overly complex. Geiger et al. [35] makes a practical assumption that the mapping τ is a linear combination of the activation of the underlying neurons. This is not part of the causal abstraction framework [36], but this assumption is important for adhering to our desiderata. Without this assumption, causal abstraction can be indeed too permissive, resulting in a *non-linear representation dilemma* where any neural network can be mapped to any algorithm. We acknowledge this critique, and capture this concern by the desideratum that the explanation should capture the underlying model at an appropriate decomposition.

Causal abstraction leads to "interpretability illusions". Makelov et al. [37] show that subspace activation patching might lead to a causal effect in the output because it activates a dormant causal pathway that contributes causally to the output. This is similar to the idea that the act of intervening on a variable might also alter other variables. The claim is that this constitutes an "interpretability illusion". Since DAS leverages interchange interventions through activation patching, the concern is that DAS gives rise to potential illusions. However, Wu et al. [38] responds to this critique, arguing that what Makelov et al. call an illusion is not always an illusion.

6 Conclusion

We propose three desiderata for faithfulness: (1) enabling reverse-engineering of specific behaviors, (2) capturing interventionist causal relations, and (3) achieving (1) and (2) at the appropriate level of decomposition. We have argued that causal abstraction is an example of an XAI method that adheres to these desiderata. The framework respects the reverse-engineering objective by integrating the interventionist definition of causality in the faithfulness objective, and it is aimed at carving the low-level model into an appropriate higher-level abstraction.

References

[1] Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, 2020. Association for Computational Linguistics.

[2] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, CA, USA, August 2016. ACM.

[3] Meike Nauta, Jan Trienes, Shreyasi Pathak, Michelle Peters, Elisa Nguyen, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023. doi: 10.1145/3583558. URL `https://arxiv.org/abs/2201.08164`.

[4] Naomi Saphra and Sarah Wiegreffe. Mechanistic? In *Proceedings of the BlackBoxNLP Workshop at EMNLP 2024*, 2024. doi: 10.48550/arXiv.2410.09087. URL `https://arxiv.org/abs/2410.09087`.

[5] Iwan Williams, Ninell Oldenburg, Ruchira Dhar, Joshua Hatherley, Constanza Fierro, Nina Rajcic, Sandrine R. Schiller, Filippos Stamatiou, and Anders Søgaard. Mechanistic interpretability needs philosophy. *arXiv preprint arXiv:2506.18852*, 2025. doi: 10.48550/arXiv.2506.18852. URL `https://arxiv.org/abs/2506.18852`.

[6] Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in RAG attributions. In *ICTIR 2025: The 15th International Conference on the Theory of Information Retrieval*. ACM, July 2025.

[7] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20)*, pages 648–657, Barcelona, Spain, 2020. Association for Computing Machinery. doi: 10.1145/3351095.3375624. URL `https://doi.org/10.1145/3351095.3375624`.

[8] 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. *Journal of Machine Learning Research*, 26:1–63, May 2025.

[9] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 2019.

[10] Afshin F. Markus, Jan A. Kors, and Peter R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655, January 2021. ISSN 1532-0464. doi: 10.1016/j.jbi.2020.103655.

[11] Emre Beyazit, Duygu Tuncel, Xiaoning Yuan, Nian-Feng Tzeng, and Xindong Wu. Learning interpretable representations with informative entanglements. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1649–1655. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/228.

[12] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL `https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf`.

[13] James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

[14] David Hume. *A Treatise of Human Nature*. Oxford University Press, 2nd edition, 1739.

[15] John Stuart Mill. *A System of Logic, Ratiocinative and Inductive*. Harper & Brothers, 8th edition, 1874.

[16] 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*, October 2022. URL https://ai.stanford.edu/blog/causal-abstraction/.

[17] Constantin Venhoff, Anisoara Calinescu, Philip Torr, and Christian Schroeder de Witt. Sage: Scalable ground truth evaluations for large sparse autoencoders. *arXiv preprint arXiv:2410.07456*, 2024.

[18] Matt Harradon, Abhishek Leiderer, et al. Causal learning and explanation of deep neural networks via autoencoded activations. *Proceedings of the 2018 AAAI Conference on Artificial Intelligence*, 2018.

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

[20] Gabriel Goh, Nick Cammarata, Chelsea Voss, et al. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. Distill.pub.

[21] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[22] Tolga Bolukbasi, Adam Pearce, Ann Yuan, et al. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

[23] Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. Ravel: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. doi: 10.48550/arXiv.2402.17700. URL https://arxiv.org/abs/2402.17700.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. doi: 10.48550/arXiv.1301.3781. URL https://arxiv.org/abs/1301.3781.

[25] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://transformer-circuits.pub/2022/toy_models_of_superposition. Transformer Circuits Thread (blog post).

[26] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

[27] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2402.03855*, 2024.

[28] LawrenceC, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny, Ansh Radhakrishnan, Buck, and Nate Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses. https://www.alignmentforum.org/posts/ZpHq3eQDFkN7gDtuF/causal-scrubbing-a-method-for-rigorously-testing, December 2022. Redwood Research.

[29] Chris Olah. A toy model of mechanistic (un)faithfulness. https://transformer-circuits.pub/2025/toy-model-of-mechanistic-unfaithfulness, August 2025. Transformer Circuits Thread.

[30] Thomas Icard. Causal abstraction and computational explanation. Invited talk, Center for Philosophy of Science, streamed live on YouTube, 2024. URL https://www.youtube.com/watch?v=sbOb6ReLKs0. April 19, 2024.

[31] Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017. doi: 10.48550/arXiv.1707.00819. URL http://auai.org/uai2017/proceedings/papers/11.pdf.

[32] Maxime Méloux, François Portet, Silviu Maniu, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *Proceedings of the International Conference on Learning Representations (ICLR)*, Grenoble, France, 2025. OpenReview.

[33] Kyle Stanford. Underdetermination of scientific theory. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2023 edition, 2023. URL https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination/. Summer 2023 Edition.

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

[35] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *arXiv*, 2024.

[36] Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

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

[38] Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to makelov et al. (2023)'s "interpretability illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024. URL https://arxiv.org/abs/2401.12631.