Causal Abstraction as a Framework for Evaluating the Faithfulness of Machine Learning Explanations

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 [1]. Here, we propose desiderata for faithfulness beyond the standard intuition of "accurately representing the reasoning process of the model" [2; 3]. We argue that Causal Abstraction provides a framework capable of aligning 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" [2; 3]. Previous work has surveyed XAI methods with respect to their faithfulness [4] without specifying exactly what we mean by faithfulness beyond this standard intuition. Moreover, Saphra and Wiegreffe [5] 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. [6] motivate the need for a philosophical grounding of mechanistic interpretability (MI) concepts. We answer their call in two ways. We first show how faithfulness is related to various desiderata of explanation, focusing on reverse-engineering, causality and aptness of decomposition. While prior work has considered disambiguating such terms from faithfulness as "out of scope" [4], we contribute to initial efforts [7] on disambiguating such terms and show how these desiderata relate to faithfulness. Next, we show how an MI framework, Causal Abstraction, can be used as principled basis for comparing the extent to which different XAI methods generate faithful explanations. We motivate Causal Abstraction as a framework for measuring faithfulness by showing that it satisfies the desiderata we have identified.

2 Existing definitions of faithfulness and their limitations

We will review two prominent directions for measuring faithfulness: Causal Scrubbing [8] and Jacobian Matching [9], and highlight why these definitions are insufficiently capturing key desiderata of faithfulness.

Jacobian Matching. First, we consider Jacobian Matching [9], which measures the faithfulness of featurisers, in particular sparse autoencoders (SAEs). SAEs encode the input features into a sparse, higher-dimensional representation, which is used for prediction. In a restricted toy example, the authors find that the surrogate SAE, which has been trained on data with repeated samples, matches the behaviour of the underlying model. However, when matching Jacobians, the memorisation feature disappears, suggesting that the surrogate model was unfaithful to the underlying model.

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

We highlight two respects in which this method fails to capture a robust notion of faithfulness: First, **we need to choose the right hyperparameter k**, which is the dimension of our latent space in which we disentangle our low-level features. Without this, we might conclude that the model is employing some features, when in fact it might be employing *entangled* features. In this case, we do not achieve a faithful explanation. Secondly, the authors themselves mention that Jacobian matching is no guarantee for faithfulness, as **the model may superficially modify its gradients, and compensate with a higher bias**.

Causal Scrubbing. Causal Scrubbing [8] can be thought of as a feature ablation method that replaces activations hypothesized to be unnecessary for a prediction with the activations for a different in-distribution input sample. Once a circuit believed to drive a behavior is identified, its faithfulness is defined as how well it preserves the model's original behavior after the rest of the model is ablated, on a given dataset. Whilst causal scrubbing shares some similarities with the framework we will propose, it is more permissive and thus further from the stricter notion of faithfulness, which we are after. In particular, **the method only intervenes in the low-level model** and compares the *output* with the output of the high-level circuit on a class of data samples Jenner et al. [10]. Therefore, it could be the case that the higher-level circuit involves components that are not used for prediction. By not intervening in the intermediate activations of the high-level circuit, we might have an explanation that does not capture the reasoning process of the low-level model at the right level of grain.

3 Desiderata for faithfulness

Having shown that past approaches fail to capture important desiderata of faithfulness, we pin down key desiderata for faithfulness.

3.1 Reverse-engineering

Reverse-engineering is a desideratum for faithfulness, because the goal of describing how a model achieved a certain output should enable us to change that behavior [11]. If it does not, we are in the dark about how to handle undesirable model behaviors.

By requiring that our explanation enables reverse-engineering, we do *not* require that it be understandable by humans. This is non-trivial [4; 12; 13], yet a human-like reasoning process does not necessarily capture the reasoning process of the model [2; 14]. Nauta et al. [4] 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".

3.2 Interventionist causality

Since we have motivated the desideratum of reverse-engineering, our explanation should also support causal intervention. This is achieved via a causal explanation *in the interventionist sense* [15], as opposed to a causal explanation in the regularity-theorist sense [16; 17].

This strict demarcation is important, since what it means for an explanation to be *causal* is ambiguous in the literature. According to Saphra and Wiegreffe [5], *cause* is defined by 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)". However, on an interventionist account, C causes E if and only if intervening on C (*ceteris paribus*), produces a change in E [15]. This definition is counterfactual, manipulability-based and particularly suited for engineering 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.

3.3 Decomposition

Williams et al. [6] argue that achieving the right decomposition of model parts is a key open problem for XAI methods in mechanistic interpretability. In our case, achieving an apt decomposition is

required for reverse-engineering. To see why, we need to acknowledge that, trivially, any explication of all low-level details of the model decisions (e.g. the model parameters/activations as a whole) might be maximally faithful to the model, yet does not constitute an *interpretable* explanation of the model decision. As stated by Geiger et al. [18]: "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 do more than just equate the explanandum with the explanans; otherwise, our definition of faithfulness fails to enable reverse-engineering.

In addition, aptness of decomposition is needed to capture the causal relations leveraged by the model. To see this, we can consider SAEs. Here, dictionary size is a hyperparameter that influences the chosen level of grain [19]. If the dictionary size is too small, then the SAE will project the features into a small subspace, possibly not ensuring full disentanglement of the components leveraged by a transformer model. In turn, interventions on these features will not cleanly map to interventions in the base model, undercutting faithfulness. On the other hand, if the grain chosen is too fine, then features will track finer-grained details, and not meaningful semantic concepts. According to Yablo [20], the decomposition should carve up the model in a relevant way, not preserving such irrelevances.

To decompose the model internals, it has been documented in various studies that individual neurons are insufficient units for encoding disjunctive concepts [21; 22; 23; 24; 25; 26; 27]. In addition, some methods assume that features are linearly separable from the activations via linear transformations [28; 29; 30; 31]. However, as rightly remarked by Geiger et al. [26], for evaluating faithfulness, we ideally do not bake such assumptions into our method for analyzing the reasoning process of the model. Hence, optimising for the right decomposition should be integral to the objective of optimising for the faithfulness of explanation.

4 Causal Abstraction

In the previous section, we argued that reverse-engineering is a key reason why we desire faithfulness of explanation. We argued that in order to effectively reverse-engineer behaviors in a model, we need to understand its causal mechanisms in the interventionist sense. In order to achieve this, we need to decompose the model internals in such a way as to capture those causal relations. We show that one framework in the literature, Causal Abstraction, is able to generate model explanations that are faithful in the way we have specified.

4.1 Causal Abstraction: How to generate faithful explanations

Due to [32], when comparing different methods for generating an explanation of model behavior, one undergoes three steps: **(1)** Construct the low-level model $\mathcal{L}$ as a causal system in a given language. $\mathcal{L}$ is the explanandum: the thing we want to explain. **(2)** Construe 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 Appendix).

In practice, the high-level model $\mathcal{H}$ is obtained by either merging variables of the low-level model, merging output values, or marginalizing (that is, removing variables of the low-level model) [26]. Alternatively, we can obtain $\mathcal{H}$ by applying a rotation matrix to the model's hidden representations to disentangle polysemantic neurons (for instance, using SAEs). Generally, the hypothesis for the high-level model can be generated using various different XAI methods.

4.2 How Causal Abstraction satisfies the reverse-engineering desideratum

If a candidate faithfulness method satisfies the reverse-engineering desideratum, it is practically useful for engineers to debug undesirable behaviors. This requires the explanation to identify and *localise* features in the network. Under the Causal Abstraction framework, this is possible via the causal consistency-preserving mapping between higher-level and lower-level variables.

4.3 How Causal Abstraction ensures that the explanation is causal (in the interventionist sense)

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. In turn, this is measured by the degree to which interventions in the high-level model commute with interventions in the low-level model [26]. 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 (see Figure 2 in Appendix). Hence, this definition respects the interventionist definition of causality, and is formalized as follows:

$$\epsilon(\alpha) = \sup_{\iota} \|\alpha\left(\mathrm{do}_L(\iota)(M_L)\right) - \mathrm{do}_H(\iota)\left(\alpha(M_L)\right)\|$$

Where M_L is the low-level causal model, α is the abstraction map that transforms the low-level model into a high-level model, ι is an intervention applied at the low level, $\mathrm{do}_L(\iota)(M_L)$ is the low-level model's behavior under intervention, $\mathrm{do}_H(\iota)(\alpha(M_L))$ is the high-level model's behavior under the corresponding intervention, $\|\cdot\|$ is a norm measuring the distance between outcomes, and $\sup_\iota$ denotes the supremum (maximum) over all valid interventions. Under this definition, if $\epsilon = 0$, the abstraction (explanation) is exactly faithful. If ϵ is small, the abstraction is approximately faithful (the high-level interventions and low-level interventions approximately commute).

For example, consider again SAEs. An SAE 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.

4.4 How Causal Abstraction ensures that the explanation captures the model at the right level of grain

We argued previously that faithfulness requires more than just equating the explanandum with the explanans. Instead, it requires aptness of decomposition. Due to Geiger et al. [26], what is desired is "a constructive causal abstraction", which is "a 'lossy' exact transformation that merges microvariables into macrovariables, while maintaining a precise and accurate description of the original model mechanisms". Thus, the art is to capture into macrovariables an approximation that captures the mechanisms of the model sufficiently well.

Note that there can in principle be multiple models, whose interventions commute well with the underlying model, but carves the model up in different ways. The claim is not that we should find the *truest* explanation. As we will later argue in our response to [33], there may be multiple competing explanations relative to the dataset at hand. As argued by Hewitt et al. [34], "there is no one right level of abstraction at which to tackle the understanding problem, but it is key to hit a good balance". For SAEs or other featurisers, where the right level of grain can be adjusted via a hyperparameter *k*, we can plot the $\epsilon(\alpha)$ score against *k*, and choose the value of *k* that optimizes the score.

5 Critiques of Causal Abstraction as a framework for faithfulness

Casual abstraction does not identify the *true* model explanation. Méloux et al. [33] shows that causal abstraction is potentially *too permissive*, as it permits multiple explanations that explain the same model behaviour. The authors pose that this is problematic, particularly because it permits two **conflicting** explanations. However, two explanations can be conflicting according to the definition employed by the authors, when they would be compatible according to the standard literature in the philosophy of science (see [35]). According to this literature, 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, but could be compatible. So, we conclude that the fact that Causal Abstraction cannot identify one correct explanation is no flaw on the method, as we should not assume that one uniquely better explanation exists. **So, faithfulness does not require a uniqueness desideratum.**

***Any* neural network can be mapped to *any* algorithm under Causal Abstraction, suggesting the framework may be too permissive and not sufficient to measure faithfulness.** Sutter et al.

[36] show that Causal Abstraction is not enough for faithfulness, as it becomes vacuous without assumptions about how models encode information. The idea is that abstractions that implement complex non-linearities will be able to achieve a higher accuracy on the benchmark employed by Causal Abstraction at the expense of being overly complex. We acknowledge this critique, and capture this concern by the desideratum that the explanation should capture the underlying model at the right level of grain. Since overfitting the model clearly does not capture the model at the right level of grain, this desideratum should eliminate this concern. We note that one version of Causal Abstraction, Distributed Alignment Search (DAS) [37], captures this requirement by assuming the linearity hypothesis.

Causal abstraction leads to "interpretability illusions". Makelov et al. [38] 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 very 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 worry is that DAS gives rise to potential illusions. Wu et al. [39] responds to this critique, showing that what Makelov et al. name an illusion, need not be an illusion. Makelov et al's definition of illusion allows perfectly good causal pathways to be dubbed "illusions". So, Wu et al. [39] bites the bullet: illusions will be common, "except it is not an "illusion" in any useful sense". Hence, it does not follow from this critique that Causal Abstraction does not generate faithful model explanations.

6 Further work

Sampling interventions. To make sure the high-level model captures the causal relations of the low-level model, we would ideally exhaust all possible interventions. However, this is not feasible in practice: as the model scales, we will have more possible hypotheses (high-level models), and for each one we would have to test all possible interventions. Still, we are able to capture a notion of faithfulness by using a sample of interventions, thereby capturing the intuition by Barez et al. [40] that faithfulness requires "partial alignment with the model's reasoning". It remains an open empirical question whether causal consistency in *this partial sense* and benchmarks measuring faithfulness via ground truth explanations [41] are compatible.

Hypothesis generation. Generating hypotheses is expensive: for current deep learning models, the number of abstractions to test can be very large [42]. One solution to this problem is to train the model to be more like the hypothesised higher-level model. The idea is that we can use the higher-level model to generate counterfactual examples and use this as ground truths against which we optimize our low-level model [43]. Due to Mueller et al. [44], this method (DAS) based on Causal Abstraction ranked highest on the faithfulness metric, and is therefore promising for overcoming this problem.

7 Conclusion

We have motivated three desiderata for faithfulness: (1) **Reverse-engineering:** A definition of faithfulness should enable reverse-engineering. (2) **Interventions (not regularities):** A faithful explanation should capture the causal relations in the interventionist sense such that reverse-engineering can be effectively achieved. (3) **Decomposition:** An explanation that captures the causal relations and aims for reverse-engineering of the model is carved up at the apt level of grain.

Furthermore, we have positioned a framework, Causal Abstraction, that allows us to compare the faithfulness of candidate explanations generated via existing XAI methods. 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 apt higher-order abstraction. However, open empirical problems remain, including how to sample for interventions when exhausting the entire set of possible interventions might be intractable, and how to effectively generate hypotheses for high-level models.

References

[1] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 2024. doi: 10.48550/arXiv.2209.11326. URL https://arxiv.org/abs/2209.11326. Published in Computational Linguistics, June 2024.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] Erik Jenner, Adrià Garriga-alonso, and Egor Zverev. A comparison of causal scrubbing, causal abstractions, and related methods. *Alignment Forum*, Jun 2023. URL https://www.alignmentforum.org/posts/uLMWMeBG3ruoBRhMW/a-comparison-of-causal-scrubbing-causal-abstractions-and.

[11] Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[12] 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.

[13] 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.

[14] 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.

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

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

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

[18] 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/.

[19] 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.

[20] Stephen Yablo. Mental causation. *The Philosophical Review*, 101(2):245–280, 1992.

[21] 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.

[22] 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/.

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

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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).

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] John Hewitt, Robert Geirhos, and Been Kim. We can't understand ai using our existing vocabulary. *arXiv preprint arXiv:2502.07586*, 2025. URL https://arxiv.org/abs/2502.07586. Accessed: 2025-10-26.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. Chain-of-thought is not explainability. Working paper / preprint (alphaXiv), 2025.

[41] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *arXiv preprint arXiv:1903.03894*, 2019. doi: 10.48550/arXiv.1903.03894. URL https://arxiv.org/abs/1903.03894.

[42] Atticus Geiger. Causal abstractions: Understanding high-level causes in neural networks. *Stanford AI Blog*, Stanford University, 2023. URL https://ai.stanford.edu/blog/causal-abstraction/. Accessed: 2025-08-23.

[43] 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 Szepesvári, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR, July 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

[44] Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun

Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. Mib: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR, 2025. doi: 10.48550/arXiv.2504.13151. URL https://arxiv.org/abs/2504.13151. To appear.

[45] 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.

A Appendix / supplemental material

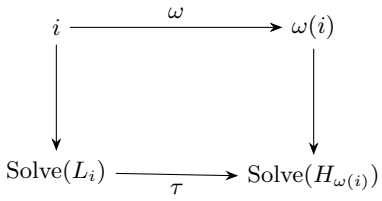

Figure 1: This commutative diagram captures causal consistency: the causal relations of the low-level model L_i are captured by the high-level model $H_{\omega(i)}$. Intuitively, "if we first perform part of the computation using the neural network and then apply the mapping to get the state of a high-level variable, the outcome should be identical to applying the mapping and then performing the computation of the high-level algorithm" [33]. We can characterize this relation in terms of the submappings τ and ω. τ is defined as 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: $\tau\big(\mathrm{Solve}(\mathcal{L}_i)\big) = \mathrm{Solve}\big(\mathcal{H}_{\omega(i)}\big)$ [45].

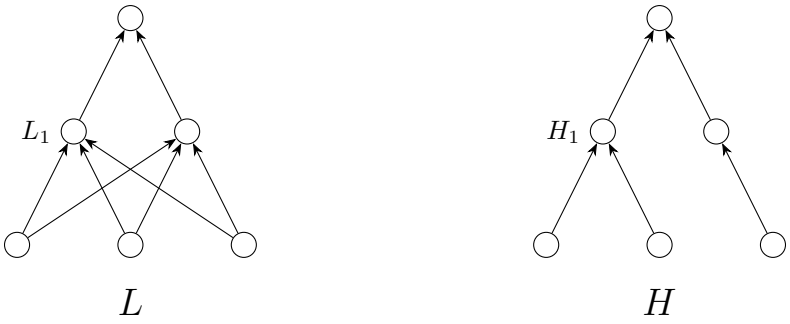

Figure 2: For illustration, we include an example of causal abstraction in use due to Geiger et al. [18]: Imagine we have the low-level model $\mathcal{L}$ and hypothesise the higher-level model $\mathcal{H}$. $\mathcal{L}$ adds three numbers together, and we hypothesise that it does so by first adding two number together, resulting in one sum, and adding the final number to this sum. For each variable, we hypothesise a mapping τ from low-level to high level model. Assume we want to test whether the high-level variable H_1 abstracts the low-level variable 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 $H_1 = 4$ and output 9 for our input [1,3,5], and $H_1 = 9$ and output 15 for our second input [4,5,6]. Imagine we patch H_1 (intervene on variable H_1), such that $H_1 = 9$. Given input [1,3,5], we get 14 as expected. We hypothesise that L_1 is captured by 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 L_1 by the same value as the corresponding high-level variable, so $L_1 = 9$. If we get the same output as a result of the patching (14), the we have a piece of evidence that the variables is 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 $\mathcal{L}$

.