Faithfulness through Causal Abstraction: Aligning explanations of how models reason

Anonymous Author(s)
Affiliation
Address
email

Abstract

Faithfulness is a broadly agreed-upon desideratum for explanations of machine learning (ML) 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 highlight a recently introduced mechanistic interpretability (MI) framework, referred to as Causal Abstraction (CA), and argue that CA provides a framework capable of aligning faithfulness claims in the community.

1 Introduction

The field of explainable AI (XAI) tries 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 having specified exactly what we mean by faithfulness beyond the standard intuition positioned by Jacovi and Goldberg [2]. The need for this work is motivated by Saphra and Wiegreffe [5] stating 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 a common MI framework, Causal Abstraction (CA), can be used as principled basis for comparing the extent to which different XAI methods generate faithful explanations. We motivate this framework with reference to our desiderata.

2 Desiderata for faithfulness

2.1 Plausibility versus faithfulness for reverse-engineering

A central reason why we want a faithful explanation is to equip us to reverse-engineer undesirable model behaviors. We do not require that it be plausible to humans.

This is non-trivial: According to the survey by Nauta et al. [4], an explanation should be understandable to humans (see also [8; 9]). However, integrating the plausibility desideratum into the definition of *explanation* is unhelpful, because a human-like reasoning process does not always capture the reasoning process of the model (See Jacovi and Goldberg [2] and [10]). 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". The orthogonality of plausibility and faithfulness is supported empirically [11]. Here, we operate under a definition of *explanation* that defines explanation as being a causal claim, due to [12]. Whether it is explainable to humans is not necessary nor sufficient for faithfulness.

2.2 Interventionist causality

There are two main tenets in the causality literature: a regularity-theorist conception [13; 14] and an interventionist conception [15]. 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 we have motivated faithfulness by reverse-engineering aims, we therefore settle on the interventionist definition as being required for faithfulness, rather than the regularity theorist definition.

The elimination of the regularity theorist conception is independently motivated by the fact that a regularity theorist conception can be satisfied without yielding an insight into the reasoning process of the model. Anders et al. [16] support this by showing that an explanation can match outputs without reflecting the model's internal reasoning. An unbiased model can be trained to deceptively generate the same outputs as an underlying biased model, without this being detectable when using different attribution methods, such as Integrated Gradients or SHAP. For instance, an arbitrary unbiased feature (football club) can act as a proxy in the biased model, encoding the bias (woman), which the model is trained on.

However, the implementation of the interventionist conception is susceptible to error, as in the case of feature ablation. Modifying C through perturbation or zeroing out may produce a change in the effect E. However, this effect need not be attributable to the cause C, but to the fact that the perturbation produced an out-of-distribution sentence [17]. Hence, interventions on the model should be preceded by an apt decomposition of model features.

2.3 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 (1) reverse-engineering and (2) effectively capturing the causal relations of the model.

To see why it is required for reverse-engineering, 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.

To illustrate why aptness of decomposition is needed to capture the causal relations leveraged by the model, 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. This example shows that a failure to achieve an apt decomposition also leads to a failure of capturing the causal relations of the underlying model. Since we have argued that capturing the causal relation of the model is key for faithfulness, then aptness of decomposition is required for faithfulness.

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

3 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 suggest that one avenue of research is particularly apt for the purpose of measuring faithfulness in the interventionist sense we have defined: mechanistic interpretability, and within it, causal abstraction. We outline why this is the case, and what still needs uncovering to empirically validate this promise.

3.1 Mechanistic interpretability as a tool for faithfulness

As identified by Saphra and Wiegreffe [5], there are various ways in which *mechanistic interpretability* has been employed. The definition we will be employing here is narrow and causal. As argued by Geiger et al. [30], "the crucial question is, under what conditions a transparent algorithm constitutes a faithful interpretation of the known, but opaque, low-level details of a black box model [...] The question takes on particular significance for mechanistic interpretability, which, in contrast to behavioral interpretability [input-output alignment], is precisely aimed at reverse engineering the internals of a black box model in terms of a transparent algorithm". Within MI, we highlight Causal Abstraction [32; 30], and argue that it can provide a framework for evaluating explanation faithfulness by capturing the desiderata we have motivated.

3.2 Causal Abstraction: Using XAI methods for hypothesis testing

Due to [33], 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 explanans: 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* $\mathcal{L} \to \mathcal{H}$. We will specify the notion of causal consistency in the next section.

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) [30] (see Figure 1 in Appendix). Alternatively, we can obtain $\mathcal{H}$ by applying a rotation matrix to the input vectors to disentangle polysemantic neurons, (for instance, using Sparse Auto-Encoders). The key insight is that the hypothesis for the high-level model is generated using various different existing XAI methods.

3.3 Causal consistency and interventionist causality

According to the Causal Abstraction framework, faithfulness of an explanation (higher-level model) is measured by how well the explanation captures the causal mechanisms of the model, which in turn is captured by the *commutation of their interventions*. This means that 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.

Geiger et al. [30] formalize the degree to which the abstraction respects the causal structure of the target model under interventions by the following formula:

$$\epsilon(\alpha) = \sup_{\iota} \left\| \alpha\left(do_L(\iota)(M_L)\right) - do_H(\iota)\left(\alpha(M_L)\right) \right\| .$$

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 and low-level models approximately commute, see Figure 2 in Appendix).

For example, consider again Sparse Autoencoders (SAEs). An SAE learns a set of sparse latent features that can be treated as candidate high-level variables used to hypothesize a causal model $\mathcal{H}$. The aligned features Π_X in the low-level network are then taken to be the neurons most strongly associated with that SAE latent variable. If the high-level model $\mathcal{H}$ fails to match $\mathcal{L}$ under interchange interventions, then the method is unfaithful.

Future research: 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. [34] 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 [35] are compatible.

3.4 Decomposition

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. [30], 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.

Future research: However, generating hypotheses is expensive: For current deep learning models, the number of abstractions to test can be very large [32]. One solution to this problem is to train the model to be more like the hypothesized higher-level causal 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 [36]. Due to Mueller et al. [37], this method (Distributed Alignment Search) ranked highest on the faithfulness metric based on Causal Abstraction, and is therefore promising for overcoming this problem.

4 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 that allows us to compare already existing XAI methods in terms of their faithfulness. 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] 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.

[9] 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.

[10] 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`.

[11] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Why is plausibility surprisingly problematic as an xai criterion? *arXiv preprint arXiv:2303.17707*, 2023. v3 updated 18 Jun 2024.

[12] Nancy Cartwright. *How the Laws of Physics Lie*. Oxford University Press, Oxford, 1983.

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

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

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

[16] Christoph J. Anders, Pavel Pasliev, Anne Dombrowski, Klaus-Robert Müller, and Patrick Kessel. Fairwashing explanations with off-manifold detergent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119. PMLR, 2020.

[17] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, pages 9737–9748, 2019.

[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] 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`.

[22] 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).

[23] 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.

[24] 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.

[25] 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.

[26] 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/`.

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

[28] 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.

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

[30] 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.

[31] 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`.

[32] 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.

[33] 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.

[34] 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.

[35] 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`.

[36] 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`.

[37] 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.

[38] 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

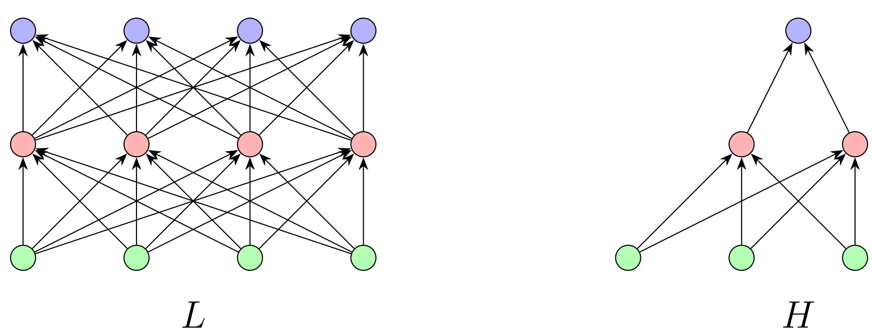

Figure 1: Low-level model $\mathcal{L}$ (left) and high-level model $\mathcal{H}$ (right) [33]

$$
\begin{tikzcd}
i \arrow[r, "\omega"] \arrow[d] & \omega(i) \arrow[d] \\
\mathrm{Solve}(L_i) \arrow[r, "\tau"'] & \mathrm{Solve}(H_{\omega(i)})
\end{tikzcd}
$$

Figure 2: This commutative diagram captures causal consistency: the causal relations of the low-level model L_i are captured by the high-level model $H_{\omega(i)}$. We can characterize this relation in terms of the submappings τ and ω, where 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, and ω is the mapping of interventions on the low-level model to the high-level model. This is formalized as: $\tau\big(\mathrm{Solve}(\mathcal{L}_i)\big) = \mathrm{Solve}\big(\mathcal{H}_{\omega(i)}\big)$ [38].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: answer: [Yes]

   Justification: We state in both the abstract and the introduction that our contribution is to specify desiderata for faithfulness, and to show how these can be implemented using a pre-existing framework (Causal Abstraction).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We acknowledge that there are limits to implementing the framework of causal abstraction, including the (1) computational efficiency of hypothesis-generation for the higher-order model and (2) the computational efficiency of exhausting the set of possible interventions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: The paper does not include theoretical results (formal proofs), only arguments. Premises of those arguments are spelled out and logical inference rules are properly employed.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case

of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA] .

   Justification: The answer NA means that paper does not include experiments requiring code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: The paper does not constitute a risk of harm.

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] .

    Justification: Our definition of faithfulness promotes positive societal impacts by enabling reverse-engineering of model behavior, which can improve safety and trust in model predictions. We highlight potential negative societal impacts of employing a definition of faithfulness, which assumes a regularity-theorist view of causality: Explanations generated by some XAI methods may appear trustworthy while being misleading.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA] .

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA] .

    Justification: This research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.