Causal and Compositional Abstraction (Extended Abstract)

Robin Lorenz, Sean Tull
Quantinuum, Partnership House, London, United Kingdom

This submission is based on the arxiv preprint [23].

Abstracting away details in order to reason in terms of explanatory high-level concepts, and often while respecting *causal structure*, is a hallmark of both human cognition and scientific practice. This intuition has been made precise through notions of *causal abstraction*, pioneered over the past decade in a number of works [6, 29, 3, 2, 25, 11], based on the causal model framework [26]. Roughly, a 'high-level' causal model $\mathbb{H}$ stands in a causal abstraction relation to a 'low-level' one $\mathbb{L}$ when there is an abstraction map τ from low to high-level variables and a relation between interventions at each level such that they are are causally *consistent* with respect to each other.

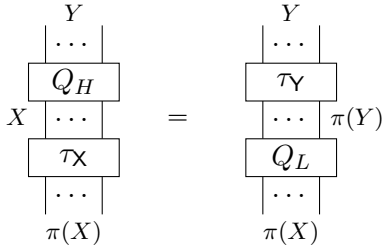
Causal abstraction is receiving increasing foundational and practical attention within a broad range of fields from AI to philosophy. It is key to *causal identifiability* problems in which causal hypotheses lie at a different level from empirical data (with examples ranging from brain data to weather phenomena) [5, 6, 34, 17]; it has been argued to be central to *mechanistic interpretability* [11] and to provide a gold standard form of *explainable AI* via abstractions from a low-level model such as a neural network to a high-level interpretable one [13, 12, 16, 14, 1, 33, 15, 32, 28]; and it is applied in a range of approaches to *causal representation learning* [30], in which AI models learn causal relations directly from low-level data [4, 36, 19, 35, 20, 21].

A language for abstraction. Despite forming a basic and essential notion for science, there is as yet no unified formalism for causal abstraction, or any consensus as to which of many closely related notions are appropriate in a given situation.

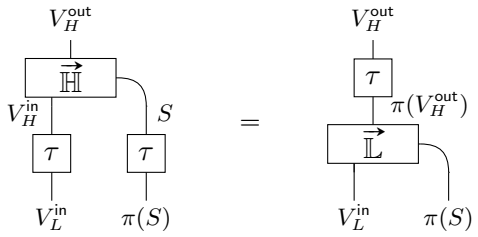
A categorical approach to formalising abstraction is ideal for both key aspects: one, capturing the structure of an individual model, through monoidal categories and string diagrams [27], and two, describing structural relationships between models, through functors and natural transformations [24]. For the first aspect, the ACT community has already developed a fully string diagrammatic account of both probability theory [9, 7] and the causal model framework [18, 22, 10]. More broadly, a wide range of AI models have been described as more general *compositional models* i.e. functors from structure to semantics [31].

In this work we provide such a general formalisation of abstraction, with causal abstractions as a special case. We find that in their essence abstractions are *natural transformations* between compositional models equipped with given sets of *queries*.

That is, we consider a model $\mathbb{M}$ to give semantics $[\![-]\!]_{\mathbb{M}}$ in a terminal SMC or Markov category $\mathbf{C}$ to a monoidal signature $\mathcal{Q}$ whose generating objects we call *types* and morphisms we call *queries*. An abstraction from a low-level model $\mathbb{L}$ to a high-level model $\mathbb{H}$, is then a family of morphisms τ_X into high-level types X from associated low-level types $\pi(X)$, and a relation sending each high-level query Q_H to at least one low-level query Q_L, where naturality amounts to the *consistency* condition, read bottom to top:



A classic case is a *constructive causal abstraction* from a low-level causal model $\mathbb{L}$ to a high-level causal model $\mathbb{H}$, which requires consistency between *Do-interventions* $\mathrm{Do}(\pi(S))$ and $\mathrm{Do}(S)$ at each level. We express this diagrammatically via a convention where bent wires denote intervened variables:[1]



Contributions. In this work we present this unified formalisation of (causal) abstractions as natural transformations, in string diagrammatic language. In more detail, we:

- Give a general categorical definition of abstraction, showing that this covers many existing forms of causal abstraction, such as constructive abstraction [3], exact transformations [29], and '$Q-\tau$-consistency' or 'counterfactual abstractions' [34], and leading us to a new precise definition of 'distributed' abstractions [11, 15], with each expressed in natural string diagrams (see Figure 1);
- Identify two basic related forms of abstraction, *upward* and *downward*, finding that while usually presented concretely as the former, the structural essence often lies in the latter, which are simply a functor $\pi\colon \mathbf{Q}_H \to \mathbf{Q}_L$ between queries and a natural transformation $\tau\colon [\![-]\!]_{\mathbb{L}} \circ \pi \implies [\![-]\!]_{\mathbb{H}}$;
- Introduce richer *component-level* (downward) abstractions, in which the functor π on queries extends to the level of model components (e.g. causal mechanisms), and use this to define a new stronger notion of *mechanism-level* constructive causal abstraction, which we characterise mathematically;
- By generalising from causal to compositional models, extend abstraction to *quantum* compositional models based on circuits, taking first steps in applying quantum-classical abstractions for interpretable quantum AI.

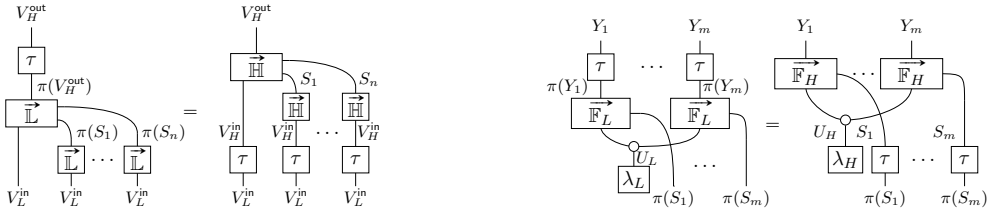


Figure 1: Abstraction conditions for (a) interchange interventions; (b) counterfactual queries.

Outlook. Overall, we hope to demonstrate that the ACT community has the potential to make powerful contributions to the growing field of (causal) abstraction and its applications across the sciences, thanks to the need for a formal perspective, the shared focus on structural relations, and the maturity of categorical approaches to causality.

[1] The independent work [8] has also recently formalised a sub-class of constructive causal abstractions in terms of naturality. We situate this within our framework in the full version of the article.

References

[1] Sander Beckers. Causal explanations and XAI. In *Conference on Causal Learning and Reasoning*, pages 90–109. PMLR, 2022.

[2] Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pages 606–615. PMLR, 2020.

[3] Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 2678–2685, 2019.

[4] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

[5] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. In *Artificial intelligence and statistics*, pages 361–369. PMLR, 2016.

[6] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

[7] Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

[8] Markus Englberger and Devendra Singh Dhami. Causal abstractions, categorically unified. *arXiv preprint arXiv:2510.05033*, 2025.

[9] Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

[10] Tobias Fritz and Andreas Klingler. The d-separation criterion in categorical probability. *J. Mach. Learn. Res*, 24(46):1–49, 2023.

[11] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *arXiv preprint arXiv:2301.04709*, 2023.

[12] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

[13] Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online, November 2020. Association for Computational Linguistics.

[14] 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 *International Conference on Machine Learning*, pages 7324–7338. PMLR, 2022.

[15] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR, 2024.

[16] Yaojie Hu and Jin Tian. Neuron dependency graphs: A causal abstraction of neural networks. In *International Conference on Machine Learning*, pages 9020–9040. PMLR, 2022.

[17] Jingzhou Huang, Jiuyao Lu, and Alexander Williams Tolbert. Causal feature learning in the social sciences. *arXiv preprint arXiv:2503.12784*, 2025.

[18] Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal inference by string diagram surgery. In *Foundations of Software Science and Computation Structures: 22nd International Conference, FOSSACS 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 22*, pages 313–329. Springer, 2019.

[19] Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. *arXiv preprint arXiv:2311.18639*, 2023.

[20] Avinash Kori, Ben Glocker, Bernhard Schölkopf, and Francesco Locatello. Unifying causal and object-centric representation learning allows causal composition. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

[21] Xiusi Li, Sékou-Oumar Kaba, and Siamak Ravanbakhsh. On the identifiability of causal abstractions. *arXiv preprint arXiv:2503.10834*, 2025.

[22] Robin Lorenz and Sean Tull. Causal models in string diagrams. *arXiv preprint arXiv:2304.07638.*, 2023.

[23] Robin Lorenz and Sean Tull. Causal and compositional abstraction. *arXiv preprint https://arxiv.org/abs/2602.16612.*, 2026.

[24] Saunders Mac Lane. *Categories for the working mathematician*, volume 5. Springer Science & Business Media, 2013.

[25] Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *Conference on Causal Learning and Reasoning*, pages 68–87. PMLR, 2023.

[26] Judea Pearl. *Causality*. Cambridge university press, 2009.

[27] Robin Piedeleu and Fabio Zanasi. *An Introduction to String Diagrams for Computer Scientists*. Elements in Applied Category Theory. Cambridge University Press, 2025.

[28] Theodora-Mara Pîslar, Sara Magliacane, and Atticus Geiger. Combining causal models for more accurate abstractions of neural networks. *arXiv preprint arXiv:2503.11429*, 2025.

[29] Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

[30] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[31] Sean Tull, Robin Lorenz, Stephen Clark, Ilyas Khan, and Bob Coecke. Towards compositional interpretability for xai. *arXiv preprint arXiv:2406.17583*, 2024.

[32] 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.

[33] Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Kevin Xia and Elias Bareinboim. Neural causal abstractions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20585–20595, 2024.

[35] Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*, 2024.

[36] Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.