

Causal Foundations for Interpretability: What Your Methods Actually Identify

Shruti Joshi, Aaron Mueller, Patrik Reizinger, Dhanya Sridhar

<https://causality4mechinterp.github.io/>

Interpretability findings on large language models often fail to generalise, and causal claims can outrun the evidence. A recurring source, we suggest, is a borrowed vocabulary: interpretability inherited the words of causality — "mechanisms", "circuits", "features" — without always inheriting the bookkeeping that keeps causal claims precise. This tutorial teaches that bookkeeping for practitioners. Participants leave able to take any interpretability result and answer three questions: what claim does this method license, what would strengthen the claim, and when will it generalise?

Shruti Joshi

email: shruti.joshi@mila.quebec

website: <https://shrutij01.github.io/>

Shruti Joshi is a PhD student at Mila and Université de Montréal, advised by Dhanya Sridhar. Her research asks how models can perform unsupervised discovery of physical mechanisms by collecting data to falsify their own hypotheses, and how to evaluate the representations they learn for the invariances they recover, drawing on identifiability theory and causal inference. She has extended this work to the interpretability and control of large language models, with a focus on safety through generalizable guarantees and on understanding models' reasoning processes.

Aaron Mueller

email: amueller@bu.edu

website: <https://aaronmueller.github.io/>

Aaron Mueller is an Assistant Professor of Computer Science, with a courtesy appointment in Data Science, at Boston University. His research addresses the robustness, efficiency, and interpretability of language models, spanning causal and mechanistic interpretability methods, linguistically informed evaluations of language models, and the development of more sample-efficient models. He completed his PhD at Johns Hopkins University's Center for Language and Speech Processing, advised by Tal Linzen and Mark Dredze, and held a Zuckerman postdoctoral fellowship at Northeastern University and the Technion.

Patrik Reizinger

email: patrik.reizinger@tuebingen.mpg.de

website: <https://rpatrik96.github.io>

Patrik Reizinger is an IVADO postdoctoral fellow in causal representation learning at Mila and Université de Montréal. His research works toward understanding deep learning through identifiable causal and self-supervised representation learning, out-of-distribution generalization, and compositionality in language models, extending machine learning theory to account for inductive biases grounded in physical and causal structure. He completed his doctoral research at the Max Planck Institute for Intelligent Systems in Tübingen and holds degrees in electrical engineering from the Budapest University of Technology and Economics, and is part of the ELLIS and IMPRS-IS programs.

Dhanya Sridhar

email: dhanya.sridhar@mila.quebec

website: <https://www.dsridhar.com/>

Dhanya Sridhar is an Assistant Professor in the Department of Informatics and Operations Research (DIRO) at Université de Montréal and a core academic member of Mila - Quebec Artificial Intelligence Institute, where she holds a Canada CIFAR AI Chair and co-leads the IVADO R3AI working group on safe and aligned AI. Her research combines causality and machine learning to build AI systems that are robust to distribution shifts, adapt efficiently to new tasks, and support scientific discovery — spanning causal representation learning under minimal supervision, causally grounded predictors, causal abstraction for interpreting large models, in-context learning, and responsible AI.