David Reber

reber@uchicago.edu · davidpreber.com · Google Scholar
Department of Computer Science, University of Chicago

Research Interests

I am a second-year Ph.D. student in Computer Science at the University of Chicago, advised by Victor Veitch. My research is about understanding how Large Language Models (LLMs) reach decisions, and how this informs agentic safety evaluations for e.g. deception or collusion; hence my research lies at the intersection of causality, interpretability, and game theory.

Education

- **Ph.D. in Computer Science**, University of Chicago — 2023 – Present
  Advisor: Victor Veitch
- **Ph.D. in Applied Mathematics**, Columbia University — 2021 – 2022
  Advisor: Elias Bareinboim (transferred to University of Chicago)
- **M.S. in Mathematics**, Brigham Young University — 2017 – 2019
- **B.S. in Applied and Computational Mathematics**, Brigham Young University — 2012 – 2017

Publications

Under Review

- **Multiple Streams of Relation Extraction: Enriching and Recalling in Transformers**
  T. Nief, D. Reber, S. Richardson, A. Holtzman
  Under review at NeurIPS 2025

Conference Papers

- **RATE: Causal Explainability of Reward Models with Imperfect Counterfactuals**
  D. Reber, S. Richardson, T. Nief, C. Garbacea, V. Veitch
  International Conference on Machine Learning (ICML), 2025

Journal Articles

- **A simple stability criterion for dynamical systems with stochastic switching and/or stochastic time-delays**
  C. Carter, J. Murri, D. Reber, B. Webb
  Nonlinearity, 35(12), 6042, 2022
- **Intrinsic stability: stability of dynamical networks and switched systems with any type of time-delays**
  D. Reber, B. Webb
  Nonlinearity, 33(6), 2560, 2020

Funding

- **Long-Term Future Fund Grant**, Effective Ventures — 2023 – 2025
  Full PhD funding support for AI safety research

Teaching Experience

- **Instructor**, Brigham Young University
  - Quantitative Reasoning 2018
  - Competitive Coding 2016

Mentoring

- **XLab Summer Research Fellowship Mentor**, University of Chicago
  - Summer 2025: Office hours for all AI safety technical fellows
  - Summer 2023: Mentored Master's student on Othello GPT interpretability project
- **Research Group Leader**, Brigham Young University 2020
  - Led 4 undergraduate students in dynamical networks research
  - Research resulted in publication in *Nonlinearity*

Professional Experience

- **Machine Learning Engineer**, Medic.Life 2018 – 2020
  - Lead researcher on ML integration with health-monitoring systems
  - Contributed to 5 provisional patents
- **EarlyAlert Project Manager**, Brigham Young University 2018
  - Led team of 4 undergraduates developing predictive academic counseling tool

Service

Conference & Workshop Reviews

- **2025:** NeurIPS (Main Conference, MechInterp Workshop), ICLR, ICML, UAI (CAR Workshop), CLeaR
- **2024:** NeurIPS (Main Conference, SoLaR Workshop), ICLR, ICML (TIFA Workshop), CLeaR
- **2023:** NeurIPS (SoLaR Workshop, CRL Workshop), ICML (SCIS Workshop)

Professional Affiliations

- Member, Causal Incentives Working Group

Technical Skills

- **Deep Learning:** PyTorch, NNsight

Honors & Awards

- **National Merit Scholar**, Full-tuition award 2012 – 2016
- **Outstanding Student of Mathematics**, Brigham Young University 2016