

Keyon Vafa

Postdoctoral Fellow
Harvard Data Science Initiative
Harvard University

kvafa@g.harvard.edu
www.keyonvafa.com

Education

Columbia University Ph.D. Computer Science Thesis: Interpretable Machine Learning for the Social Sciences: Applications in Political Science and Labor Economics Committee: David Blei (advisor), Susan Athey, Suresh Naidu, Zhou Yu, Richard Zemel	2017 - 2023
Columbia University M.S. Computer Science	2017 - 2018
Harvard University B.A. Computer Science and Statistics, <i>magna cum laude</i>	2012 - 2016

Awards and Fellowships

Morton B. Friedman Memorial Prize for Excellence in Engineering (<i>given to one graduate each year across engineering departments at Columbia</i>)	2024
Columbia University Nominee for ACM Doctoral Dissertation Award	2024
Harvard Data Science Initiative Postdoctoral Fellowship	2023 -
Cheung-Kong Innovation Doctoral Fellowship	2020 - 2022
Columbia University Nominee for Google PhD Fellowship	2019
National Science Foundation, Graduate Research Fellowship	2016 - 2019
Phi Beta Kappa Society	2016
Bok Center Certificate of Distinction in Teaching	2015
John Harvard Scholar (<i>grade point average in top 5% of class</i>)	2013 - 2015

Papers

- **K. Vafa**, S. Bentley, J. Kleinberg, S. Mullainathan. What's Producible May Not Be Reachable: Measuring the Steerability of Generative Models. *Neural Information Processing Systems (NeurIPS)*, 2025.
- **K. Vafa**, P. G. Chang, A. Rambachan, S. Mullainathan. What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. *International Conference on Machine Learning (ICML)*, 2025.

- M. Mancoridis, **K. Vafa**, B. Weeks, S. Mullainathan. Potemkin Understanding in Large Language Models. *International Conference on Machine Learning (ICML)*, 2025.
- **K. Vafa**, S. Athey, D. Blei. Estimating wage disparities using foundation models. *Proceedings of the National Academy of Science (PNAS)*, 2025.
- C. Lee, A. M. Rush, **K. Vafa**. Critical Thinking: Which Kinds of Complexity Govern Optimal Reasoning Length?. *International Joint Conference on Natural Language Processing & Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP & AACL)*, 2025.
- **K. Vafa**, J. Y. Chen, A. Rambachan, J. Kleinberg S. Mullainathan. Evaluating the World Model Implicit in a Generative Model. *Neural Information Processing Systems (NeurIPS)*, 2024 [spotlight].
- **K. Vafa**, A. Rambachan, S. Mullainathan. Do large language models perform the way people expect? Measuring the human generalization function. *International Conference on Machine Learning (ICML)*, 2024.
- **K. Vafa**. Is causal inference compatible with frictionless reproducibility? *Harvard Data Science Review*, 2024.
- E. Pierson, D. Shanmugam, R. Movva, J. Kleinberg, M. Agrawal, M. Dredze, K. Ferryman, J. W. Gichoya, D. Jurafsky, P. W. Koh, K. Levy, S. Mullainathan, Z. Obermeyer, H Suresh, **K. Vafa**. Using large language models to promote equity. *New England Journal of Medicine AI*, 2024.
- **K. Vafa**, E. Palikot, T. Du, A. Kanodia, S. Athey, D. Blei. CAREER: A foundation model for labor sequence data. *Transactions of Machine Learning Research (TMLR)* [also spotlight presentation at *NeurIPS Workshop on Distribution Shifts*], 2023.
- C. Zheng, **K. Vafa**, D. Blei. Revisiting topic-guided language models. *Transactions of Machine Learning Research (TMLR)*, 2023.
- C. Zheng, C. Shi, **K. Vafa**, A. Feder, D. Blei. An invariant learning characterization of controlled text generation. *Association for Computational Linguistics (ACL)*, 2023.
- **K. Vafa**. Interpretable machine learning for the social sciences: Applications in political science and labor economics. *Ph.D. Thesis*, 2023.
- **K. Vafa**, Y. Deng, D. Blei, A. Rush. Rationales for sequential predictions. *Empirical Methods in Natural Language Processing (EMNLP)*, 2021 [oral].
- A. Schein, **K. Vafa**, D. Sridhar, V. Veitch, J. Moffet, J. Quinn, N. Makiya, D. Blei, D. Green. A digital field experiment reveals large effects of friend-to-friend texting on voter turnout. *The Web Conference (WWW)*, 2021.
- **K. Vafa**, S. Naidu, D. Blei. Text-based ideal points. *Association for Computational Linguistics (ACL)*, 2020.

- D. Tran, **K. Vafa**, K. Agrawal, L. Dinh, B. Poole. Discrete flows: Invertible generative models of discrete data. *Neural Information Processing Systems (NeurIPS)*, 2019.
- **K. Vafa**. Training deep Gaussian processes with sampling. *NeurIPS Workshop on Advances in Approximate Bayesian Inference Workshop*, 2016.
- **K. Vafa**, C. Haigh, A. Leung, N. Yonack. Price discrimination in the Princeton Review's online SAT tutoring service. *Journal of Technology Science*, 2015.

Preprints

- S. Sarkar, **K. Vafa**. Lookahead bias in pretrained language models. *Preprint*, 2024.
- T. Du, A. Kanodia, H. Brunborg, **K. Vafa**, S. Athey. LABOR-LLM: Language-based occupational representations with large language models. *Preprint*, 2024.

Talks

- **2025**
 - NeurIPS Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning, Invited Talk
 - University of Pennsylvania, ASSET Seminar
 - UC Berkeley, Special Seminar
 - University of Washington, ML Seminar
 - TTIC, Young Researcher Seminar Series
 - ETH Zurich, Economics and Data Science Online Seminar
 - Renaissance Technologies, Monthly Colloquium
 - AI For Science Workshop at Carnegie Mellon University
 - Carnegie Mellon University, Machine Learning Seminar
 - Oxford University, Metrics and Models Seminar Series
 - Microsoft Research New York
 - Cambridge University, Language Technology Lab Seminar Series
 - University of Chicago Booth, Machine Learning in Economics Summer Institute
 - Google, SPARK Seminar
- **2024**
 - MIT, Schwarzman College of Computing Workshop
 - Opportunity Insights, Interdisciplinary Lunch Seminar

- Princeton Language and Intelligence initiative (PLI) and Center for Information Technology Policy (CITP), Special Seminar
- NBER Summer Institute 2024, Session on Big Data and High-Performance Computing for Financial Economics
- MIT FutureTech, Workshop on the Role of AI in Science
- Seminar on Formal Languages and Neural Networks (FLaNN)
- JetBrains Research, ML for Code seminar
- Northwestern University, IDEAL Workshop on Theoretical Foundations of Human-AI Complementarity
- Econometric Society Interdisciplinary Frontiers (ESIF) conference on Economics and AI+ML
- University of Chicago Booth, Machine Learning in Economics Summer Institute
- Summer Institute in Computational Social Science (ODISSEI/Rotterdam), Guest Lecture
- Columbia Management, Analytics, and Data Conference
- Carnegie Mellon University (Heinz), Heinz Lunch Series
- **2023**
 - Google DeepMind
 - NBER Summer Institute 2023, Labor Studies
 - Zurich Workshop in AI + Economics
 - Copenhagen Center for Social Data Science at the University of Copenhagen, The Science of the Predicted Human Talk Series
 - Technical University of Denmark (DTU)
 - Harvard Institute for Quantitative Social Science, Workshop in Applied Statistics
 - University of Chicago Booth, Machine Learning in Economics Summer Institute
 - Stanford University, Golub Capital Social Impact Lab
 - ETH Zurich AI Center
 - Microsoft Research, Computational Social Science group
- **2022**
 - NeurIPS Workshop on Distribution Shifts, Spotlight Talk
 - Federal Committee on Statistical Methodology Conference
 - ETH Zurich
- **2021**
 - Google AI, NLP Reading Group
 - Hugging Face

- EMNLP Conference, Oral
- CFE-CMStatistics Conference
- **2020**
 - Cornell Tech University, Milstein Program Summer Speaker Series
- **2019**
 - Text as Data Conference at Stanford University
 - Caselaw Access Project Research Summit at Harvard Law School

Journal and Conference Reviewing

- **Reviewing**
 - Neural Information Processing Systems
 - 2017, 2018, 2019, 2020, 2021, 2022, 2024, 2025
 - International Conference on Machine Learning
 - 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025
 - International Conference on Learning Representations
 - 2018, 2019, 2020, 2021, 2025
 - Harvard Data Science Review
 - 2023 - present
 - Advances in Approximate Bayesian Inference
 - 2017, 2018, 2019, 2020, 2021, 2022, 2023
 - ACM Conference on Fairness, Accountability, and Transparency
 - 2023
 - I Can't Believe It's Not Better Workshop
 - 2020, 2021, 2022
 - Association for Computational Linguistics
 - 2021
 - ACL Rolling Review
 - 2021, 2025
- **Reviewing Recognition**
 - Reviewer Award, International Conference on Learning Representations (ICLR), 2021
 - Expert Reviewer, International Conference on Machine Learning (ICML), 2021
 - Top 33% Reviewer, International Conference on Machine Learning (ICML), 2020

- Top 10% Reviewer, Neural Information Processing Systems (NeurIPS), 2020

Work Experience

- Google Brain, Software Engineer Intern (2018-2019)
- Facebook Artificial Intelligence Research (FAIR), Research Intern (2017)
- Facebook, Data Science Intern (2015)
- Facebook, Software Engineer Intern (2014)

Teaching

- Foundations of Graphical Models, Teaching Assistant (taught by David Blei), 2018
- CS 281: Advanced Machine Learning, Teaching Fellow (taught by Finale-Doshi Velez), 2015
- CS 181: Introduction to Machine Learning, Teaching Fellow (taught by Ryan Adams), 2015

Organizing and Volunteering

- ICML 2025 Workshop on Assessing World Models (lead organizer)
- NeurIPS 2024 Workshop on Behavioral Machine Learning (lead organizer)
- Harvard Data Science Review, Early Career Board (2023 - present)
- Machine Learning in NYC, Organizer (2022-2023)
- GetUsPPE, Data Scientist (2020)

Popular Press

- “‘World Models,’ an Old Idea in AI, Mount a Comeback”. *Quanta Magazine*. September 2, 2025.
- “Can large language models figure out the real world?”. *MIT News*. August 25, 2025.
- “Science In Action”. *BBC News*. July 17, 2025.
- “Does AI understand?”. *Harvard Gazette*. July 16, 2025.

- “The Hidden Layer”. *Puck News*. July 15, 2025.
- “AI models just don’t understand what they’re talking about”. *The Register*. July 3, 2025.
- “We Now Know How AI ‘Thinks’—and It’s Barely Thinking at All”. *Wall Street Journal*. May 25, 2025.
- “How close is AI to human-level intelligence?” *Nature*. December 3, 2024.
- “Despite its impressive output, generative AI doesn’t have a coherent understanding of the world”. *MIT News*. November 4, 2024.
- “Large language models don’t behave like people, even though we may expect them to”. *MIT News*. July 23, 2024.
- “CAREER Prediction”. *Data Skeptic*. March 29, 2023.
- “SAT Prep Course Controversy: Why You May be Getting Overcharged.” *The Today Show*. March 11 2016.
- “Test Prep Is More Expensive—for Asian Students.” *The Atlantic*. September 3, 2015.
- “The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review”. *ProPublica*. September 1, 2015.