Himabindu Lakkaraju

Contact Information

491 Morgan Hall
15 Harvard Way
Boston, MA 02163

Science and Engineering Complex
150 Western Ave, Suite 6.220
Boston, MA 02134

E-mail: hlakkaraju@seas.harvard.edu; himalakkaraju@google.com
Webpage: http://himalakkaraju.github.io

Research Interests

Trustworthy AI (Interpretability, Fairness, Robustness, Privacy, Safety, and Alignment); Machine Learning; Artificial Intelligence; Optimization; Statistics; Human-AI Interaction; Large Language Models; Applications of AI to Healthcare, Law, and Policy.

Academic & Professional Experience

Harvard University
Assistant Professor with appointments in the Business School and the Department of Computer Science — *01/2020 - Present*
Postdoctoral Fellow — *11/2018 - 12/2019*

Google Research, Cambridge
Senior Staff Research Scientist (Research Lead, Safety and Social Intelligence of Generative AI) — *08/2025 - Present*

Simons Institute for the Theory of Computing, UC Berkeley
Visiting Scientist, Summer Cluster on Interpretable Machine Learning — *06/2022 - 08/2022*
Visiting Graduate Student, Summer Cluster on Algorithmic Fairness — *07/2018 - 08/2018*

Microsoft Research, Redmond
Visiting Researcher — *5/2017 - 6/2017*
Research Intern — *6/2016 - 9/2016*

University of Chicago
Data Science for Social Good Fellow — *6/2014 - 8/2014*

IBM Research, Bengaluru and New York
Research Engineer — *7/2010 - 7/2012*

Advisory Roles

The Stanford Center for Legal Informatics, Stanford University
Advisory Board Member, Computational Antitrust Project — 01/2020 - Present

Fiddler AI
Chief AI Research Fellow and Advisor — *06/2021 - 11/2022*

Education

Stanford University
Doctor of Philosophy (PhD) in Computer Science — *9/2012 - 9/2018*
Master of Science (MS) in Computer Science — *9/2012 - 9/2015*

Indian Institute of Science (IISc)
Master of Engineering (MEng) in Computer Science & Automation — *8/2008 - 7/2010*

Selected Honors & Achievements

Alfred P. Sloan Research Fellowship in Computer Science — 2025
Outstanding Paper Award in New England NLP Symposium — 2025
AI2050 Early Career Fellowship by Schmidt Sciences — 2024
Distinguished Young Alumnus Award, Indian Institute of Science — 2024

NSF CAREER Award	2023
Named **Kavli Fellow** by the National Academy of Sciences	2023
Adobe Data Science Research Award	2023
Best Paper Award, **ICML Workshop on Interpretable ML in Healthcare**	2022
Outstanding Paper Award Honorable Mention **NeurIPS Workshop on Trustworthy and Socially Responsible Machine Learning**	2022
JP Morgan Faculty Research Award	2022
Selected as a member of the **National AI Advisory Committee** instituted by the US government (could not serve due to citizenship status)	2022
National Science Foundation (NSF) Amazon Fairness in AI Grant	2021
Google AI for Social Good Research Award	2021
Best Paper Runner Up, **ICML Workshop on Algorithmic Recourse**	2021
Google Research Award	2020
Amazon Research Award	2020
Co-founded **Trustworthy ML Initiative** with the goal of enabling easy access to resources on trustworthy ML & to build a community of researchers/practitioners	2020
Hoopes Prize for undergraduate thesis mentoring, Harvard University	2020
Named as one of the **35 Innovators Under 35** (Global) by MIT Tech Review	2019
Named as an **Innovator to Watch** by Vanity Fair	2019
Selected for the prestigious **Cowles Fellowship** by Yale University (declined)	2018
INFORMS Data Mining Best Paper Award	2017
Microsoft Research Dissertation Grant	2017
Named as a **Rising Star in Computer Science**	2016
Outstanding Reviewer Award International World Wide Web Conference (WWW)	2016
Google Anita Borg Fellowship in recognition of research and leadership	2015
Stanford Graduate Fellowship for exceptional academic performance Awarded to top 3% of Stanford Ph.D. students	2013-17
Eminence and Excellence Award for outstanding research contributions IBM Research	2012
Best Paper Award, **SIAM International Conference on Data Mining (SDM)**	2011
All India Rank 32 (99.82%ile) Graduate Aptitude Test in Engineering (GATE) Entrance examination for IISc & IITs in Computer Science & Engineering	2008

Selected Grants & Fellowships

As Faculty

Alfred P. Sloan Research Fellowship (US$75,000) – Sole PI	2025 - 2026
AI2050 Early Career Fellowship by Schmidt Sciences (US$300,00) – Sole PI	2024 - 2027
OpenAI Research Compute Award (US$10,000) – Sole PI	2024 - 2025
NSF CAREER Award (US$550,664) – Sole PI	2023 - 2028
Adobe Data Science Research Award (US$50,000) – PI	2023 - 2024
Microsoft Azure Compute Award (US$22,224) – Sole PI	2023 - 2024
D3 Institute at Harvard Grant (US$600,000) – Sole PI	2022 - 2025

JP Morgan Faculty Research Award (US$110,000) – Sole PI 2022 - 2024
NSF-Amazon Fairness in AI (FAI) grant (US$375,000) – co-PI 2021 - 2024
Amazon Faculty Research Award (US$70,000) – Sole PI 2021 - 2024
Google AI for Social Good Research Award (US$10,000) – Sole PI 2021 - 2022
Google Research Award (US$600,000) – PI 2020 - 2024
NSF IIS: Robust Intelligence (RI) Small (US$450,000) – Harvard PI 2020 - 2023
Bayer Trust in Science Award (US$100,000) – PI 2020 - 2021

As Student

Microsoft Research Dissertation Grant (US$20,000) 2017
Stanford Graduate Fellowship (tuition + US$41,700 p.a.) 2013 - 2017
Google Anita Borg Scholarship (US$10,000) 2015
Facebook Graduate Fellowship Finalist (US$500) 2013
Indian Institute of Science Graduate Scholarship (tuition + Rs.96,000 p.a.) 2008 - 2010
SAP India Research Grant (Rs.150,000) 2009 - 2010

Research Articles

Total Citations: 12867 **h-index**: 46 **i10-index**: 85

(* below indicates equal contribution)

Articles in Peer-Reviewed Journals

[87] Advancing Science- and Evidence-Based AI Policy
Rishi Bommasani, Sanjeev Arora, Yejin Choi, Fei Fei Li, Daniel Ho, Dan Jurafsky, Sanmi Koyejo, Himabindu Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, Dawn Song
Science, 2025.

[86] Detecting LLM-Generated Peer Reviews
Vishisht Rao, Aounon Kumar, Himabindu Lakkaraju, Nihar B. Shah
PLOS ONE, 2025.

[85] The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective
Satyapriya Krishna*, Tessa Han*, Alex Gu, Steven Wu, Shahin Jabbari, Himabindu Lakkaraju
TMLR - Transactions on Machine Learning Research, 2024.
Best Paper Award, ICML Workshop on Interpretable ML, 2022.
Featured in Fortune Magazine

[84] TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations
Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju*, Sameer Singh*
Nature Machine Intelligence - 2023.
Outstanding Paper Award Honorable Mention, NeurIPS Workshop on Trustworthy and Socially Responsible ML, 2022.

[83] Evaluating Explainability for Graph Neural Networks
Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, Marinka Zitnik
Nature Scientific Data - 2023.

[82] When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making
Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, Himabindu Lakkaraju
TMLR - Transactions on Machine Learning Research, 2023.
Featured in VentureBeat

[81] Human Decisions and Machine Predictions
Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan

QJE - Quarterly Journal of Economics, 2018.
(author names are ordered alphabetically)
Featured in MIT Technology Review, Harvard Business Review, The New York Times, and as Research Spotlight on National Bureau of Economics front page

[80] Mining Digital Footprints to Extract Patterns and Predict Real-Life Outcomes
Michal Kosinski, Yilun Wang, Himabindu Lakkaraju, Jure Leskovec
Psychological Methods - 2016.

Articles in Peer-Reviewed Conference Proceedings

[79] EvoLM: In Search of Lost Language Model Training Dynamics
Zhenting Qi, Fan Nie, Alexandre Alahi, James Zou, Himabindu Lakkaraju, Yilun Du, Eric Xing, Sham Kakade, Hanlin Zhang
NeurIPS - Advances in Neural Information Processing Systems, 2025.
Oral Presentation (Top 0.4%)

[78] All Proxy Rewards are Bad, Can We Hedge to Make Some Useful?
Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, Flavio Calmon
NeurIPS - Advances in Neural Information Processing Systems, 2025.
Spotlight Presentation (Top 3.2%)

[77] Measuring the Faithfulness of Thinking Drafts in Large Reasoning Models
Zidi Xiong, Shan Chen, Zhenting Qi, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2025.

[76] How Post-Training Reshapes LLMs: A Mechanistic View on Knowledge, Truthfulness, Refusal, and Confidence
Hongzhe Du, Weikai Li, Min Cai, Karim Saraipour, Zimin Zhang, Himabindu Lakkaraju, Yizhou Sun, Shichang Zhang
COLM - Conference on Language Modeling, 2025.
Outstanding Paper Award, New England NLP Symposium 2025

[75] More RLHF, More Trust? On The Impact of Preference Alignment On Trustworthiness
Aaron Jiaxun Li, Satyapriya Krishna, Himabindu Lakkaraju
ICLR - International Conference on Learning Representations, 2025.
Oral Presentation (Top 1.8%)

[74] Follow My Instruction and Spill the Beans: Scalable Data Extraction from Retrieval-Augmented Generation Systems
Zhenting Qi, Hanlin Zhang, Eric P. Xing, Sham M. Kakade, Himabindu Lakkaraju
ICLR - International Conference on Learning Representations, 2025.

[73] Quantifying Generalization Complexity for Large Language Models
Zhenting Qi, Hongyin Luo, Xuliang Huang, Zhuokai Zhao, Yibo Jiang, Xiangjun Fan, Himabindu Lakkaraju, James Glass
ICLR - International Conference on Learning Representations, 2025.

[72] On the Impact of Fine-Tuning on Chain-of-Thought Reasoning
Elita Lobo, Chirag Agarwal, Himabindu Lakkaraju
NAACL - The North American Chapter of the Association for Computational Linguistics, 2025.

[71] Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE)
Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2024.

[70] MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models
Tessa Han, Aounon Kumar, Chirag Agarwal, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2024.

[69] In-context Unlearning: Language Models as Few Shot Unlearners
Martin Pawelczyk, Seth Neel, Himabindu Lakkaraju
ICML - International Conference on Machine Learning, 2024.

[68] Understanding the Effects of Iterative Prompting on Truthfulness
Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju
ICML - International Conference on Machine Learning, 2024.

[67] Characterizing Data Point Vulnerability as Average-Case Robustness
Tessa Han, Suraj Srinivas, Himabindu Lakkaraju
UAI - International Conference on Uncertainty in Artificial Intelligence, 2024.

[66] Quantifying Uncertainty in Natural Language Explanations of Language Models
Sree Harsha Tanneru, Chirag Agarwal, Himabindu Lakkaraju
AISTATS - International Conference on Artificial Intelligence and Statistics, 2024. **Spotlight Presentation, NeurIPS Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models, 2023**.

[65] Fair Machine Unlearning: Data Removal while Mitigating Disparities
Alex Oesterling, Jiaqi Ma, Flavio Calmon, Himabindu Lakkaraju
AISTATS - International Conference on Artificial Intelligence and Statistics, 2024.

[64] Certifying LLM Safety against Adversarial Prompting
Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Li, Soheil Feizi, Himabindu Lakkaraju
COLM - Conference on Language Modeling, 2024
Featured in Science News

[63] Investigating the Fairness of Large Language Models for Predictions on Tabular Data
Yanchen Liu, Srishti Gautam, Jiaqi Ma, Himabindu Lakkaraju
NAACL - The North American Chapter of the Association for Computational Linguistics, 2024.

[62] A Study on the Calibration of In-context Learning
Hanlin Zhang, Yi-Fan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Himabindu Lakkaraju, Sham Kakade
NAACL - The North American Chapter of the Association for Computational Linguistics, 2024.

[61] On the Impact of Adversarially Robust Models on Algorithmic Recourse
Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju
AIES - AAAI/ACM Conference on AI, Ethics, and Society, 2024.

[60] Post hoc Explanations of Language Models can Improve Language Models
Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2023.

[59] Which Models have Perceptually-Aligned Gradients? An Explanation via Off-Manifold Robustness
Suraj Srinivas*, Sebastian Bordt*, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2023.
Spotlight Presentation (Top 3%)

[58] Verifiable Feature Attributions: A Bridge between Post Hoc Explainability and Inherent Interpretability
Usha Bhalla*, Suraj Srinivas*, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2023.

[57] M4: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities, and Models
Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, Haoyi Xiong
NeurIPS - Advances in Neural Information Processing Systems, 2023.

[56] Towards Bridging the Gaps between the Right to Explanation and the Right to be Forgotten
Satyapriya Krishna*, Jiaqi Ma*, Himabindu Lakkaraju
ICML - International Conference on Machine Learning, 2023.

[55] On the Impact of Actionable Explanations on Social Segregation
Ruijiang Gao, Himabindu Lakkaraju
ICML - International Conference on Machine Learning, 2023.

[54] On Minimizing the Impact of Dataset Shifts on Actionable Explanations
Anna Meyer*, Dan Ley*, Suraj Srinivas, Himabindu Lakkaraju
UAI - Conference on Uncertainty in Artificial Intelligence, 2023.
Oral Presentation (Top 5%)

[53] Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse
Martin Pawelczyk, Teresa Datta, Johannes van den Heuvel, Gjergji Kasneci, Himabindu Lakkaraju
ICLR - International Conference on Learning Representations, 2023.

[52] On the Privacy Risks of Algorithmic Recourse
Martin Pawelczyk, Himabindu Lakkaraju*, Seth Neel*
AISTATS - International Conference on Artificial Intelligence and Statistics, 2023.

[51] Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations
Tessa Han, Suraj Srinivas, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems (NeurIPS), 2022.
Best Paper Award, ICML Workshop on Interpretable ML, 2022.

[50] Flatten the Curve: Efficiently Training Low-Curvature Neural Networks
Suraj Srinivas, Kyle Matoba, Himabindu Lakkaraju, Francois Fleuret
NeurIPS - Advances in Neural Information Processing Systems (NeurIPS), 2022.

[49] OpenXAI: Towards a Transparent Evaluation of Model Explanations
Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems (NeurIPS), 2022.

[48] Data Poisoning Attacks on Off-Policy Evaluation Methods
Elita Lobo, Harvineet Singh, Marek Petrik, Cynthia Rudin, Himabindu Lakkaraju
UAI - Conference on Uncertainty in Artificial Intelligence, 2022.
Oral Presentation (Top 5%)

[47] Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis
Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, Himabindu Lakkaraju
AISTATS - International Conference on Artificial Intelligence and Statistics, 2022.

[46] Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods
Chirag Agarwal, Marinka Zitnik*, Himabindu Lakkaraju*
AISTATS - International Conference on Artificial Intelligence and Statistics, 2022.

[45] Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations
Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen Bach, Himabindu Lakkaraju
AIES - AAAI/ACM Conference on AI, Ethics, and Society, 2022.

[44] Towards Robust Off-Policy Evaluation via Human Inputs
Harvineet Singh, Shalmali Joshi, Finale Doshi-Velez, Himabindu Lakkaraju
AIES - AAAI/ACM Conference on AI, Ethics, and Society, 2022.

[43] A Human-Centric Perspective on Model Monitoring
Murtuza N Shergadwala, Himabindu Lakkaraju, Krishnaram Kenthapadi
HCOMP - AAAI Conference on Human Computation and Crowdsourcing, 2022.

[42] Towards Robust and Reliable Algorithmic Recourse
Sohini Upadhyay*, Shalmali Joshi*, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems (NeurIPS), 2021.
Best Paper Runner Up, ICML Workshop on Algorithmic Recourse, 2021.

[41] Reliable Post hoc Explanations: Modeling Uncertainty in Explainability
Dylan Slack, Sophie Hilgard, Sameer Singh, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2021.

[40] Counterfactual Explanations Can Be Manipulated
Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, Sameer Singh
NeurIPS - Advances in Neural Information Processing Systems, 2021.

[39] Learning Models for Algorithmic Recourse
Alexis Ross, Himabindu Lakkaraju, Osbert Bastani
NeurIPS - Advances in Neural Information Processing Systems, 2021.

[38] Towards the Unification and Robustness of Perturbation and Gradient Based Explanations
Sushant Agarwal, Shahin Jabbari, Chirag Agarwal*, Sohini Upadhyay*, Steven Wu, Himabindu Lakkaraju
ICML - International Conference on Machine Learning, 2021.
Shorter version presented at Foundations of Responsible Computing (FORC), 2022.

[37] Towards a Unified Framework for Fair and Stable Graph Representation Learning
Chirag Agarwal, Himabindu Lakkaraju*, Marinka Zitnik*
UAI - Conference on Uncertainty in Artificial Intelligence, 2021.
Oral Presentation (Top 5%)

[36] Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring
Tom Suhr, Sophie Hilgard, Himabindu Lakkaraju
AIES - AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2021.

[35] Fair influence maximization: A welfare optimization approach
Aida Rahmattalabi, Shahin Jabbari, Himabindu Lakkaraju, Phebe Vayanos, Eric Rice, Milind Tambe
AAAI - AAAI International Conference on Artificial Intelligence, 2021.

[34] Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses
Kaivalya Rawal, Himabindu Lakkaraju
NeurIPS - Advances in Neural Information Processing Systems, 2020.

[33] Incorporating Interpretable Output Constraints in Bayesian Neural Networks
Wanqian Yang, Lars Lorch, Moritz Gaule, Himabindu Lakkaraju, Finale Doshi-Velez
NeurIPS - Advances in Neural Information Processing Systems, 2020.
Spotlight Presentation (Top 3%)

[32] Robust and Stable Black Box Explanations
Himabindu Lakkaraju, Nino Arsov, Osbert Bastani
ICML - International Conference on Machine Learning, 2020

[31] How do I fool you?: Manipulating User Trust via Misleading Black Box Explanations
Himabindu Lakkaraju, Osbert Bastani
AIES - AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2020.
Oral Presentation (Top 16.6%)

[30] Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods
Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, Himabindu Lakkaraju
AIES - AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2020.
Featured in Harvard Business Review and deeplearning.ai
Best Paper (Non-Archival) at AAAI Workshop on Safe AI, 2020
Oral Presentation (Top 16.6%)

[29] Faithful and Customizable Explanations of Black Box Models
Himabindu Lakkaraju, Ece Kamar, Rich Caruana, Jure Leskovec
AIES - AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2019.
Oral Presentation (Top 10%)

[28] The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables
Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan
KDD - ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2017.
Oral Presentation (Top 8.5%)

[27] Learning Cost-Effective and Interpretable Treatment Regimes
Himabindu Lakkaraju, Cynthia Rudin
AISTATS - International Conference on Artificial Intelligence and Statistics, 2017.
INFORMS Data Mining Best Paper Award, 2017

[26] Identifying Unknown-Unknowns in the Open World: Representations and Policies for Guided Exploration
Himabindu Lakkaraju, Ece Kamar, Rich Caruana, Eric Horvitz
AAAI - AAAI International Conference on Artificial Intelligence, 2017.
Featured in Bloomberg Technology

[25] Confusions over Time: An Interpretable Bayesian Model for Characterizing Trends in Decision Making
Himabindu Lakkaraju, Jure Leskovec
NIPS - Advances in Neural Information Processing Systems, 2016.

[24] Interpretable Decision Sets: A Joint Framework for Description and Prediction
Himabindu Lakkaraju, Stephen Bach, Jure Leskovec
KDD - ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[23] A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes
Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, Kecia Addison
KDD - ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.
Oral Presentation (Top 8.2%)

[22] A Bayesian Framework for Modeling Human Evaluations
Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, Sendhil Mullainathan
SDM - SIAM International Conference on Data Mining, 2015.
Oral Presentation (Top 5%)

[21] Who, When, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time
Everaldo Aguiar, Himabindu Lakkaraju, Nasir Bhanpuri, David Miller, Ben Yuhas, Kecia Addison, Shihching Liu, Marilyn Powell and Rayid Ghani
LAK - Learning Analytics and Knowledge Conference, 2015.

[20] What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media
Himabindu Lakkaraju, Julian McAuley, Jure Leskovec
ICWSM - International AAAI Conference on Weblogs and Social Media, 2013.
Featured in Time, Forbes, Phys.Org, Business Insider, New Scientist
Oral Presentation (Top 3%)

[19] Dynamic Multi-Relational Chinese Restaurant Process for Analyzing Influences on Users in Social Media
Himabindu Lakkaraju, Indrajit Bhattacharya, Chiranjib Bhattacharyya
ICDM - IEEE International Conference on Data Mining, 2012.
Oral Presentation (Top 8.6%)

[18] Attention Prediction on Social Media Brand Pages
Himabindu Lakkaraju, Jitendra Ajmera
CIKM - ACM Conference on Information and Knowledge Management, 2011.

[17] Exploiting Coherence for the Simultaneous Discovery of Latent Facets and Associated Sentiments
Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya, Srujana Merugu
SDM - SIAM International Conference on Data Mining, 2011.
Best Paper Award

[16] TEM: A Novel Perspective to Modeling Content on Microblogs
Himabindu Lakkaraju, Hyung-Il-Ahn
WWW - International World Wide Web Conference, 2011.

[15] Smart News Feeds for Social Networks Using Scalable Joint Latent Factor Models
Himabindu Lakkaraju, Angshu Rai, Srujana Merugu
WWW - International World Wide Web Conference, 2011.

Selected Preprints, Working Papers, and Workshop Articles

[14] Who Gets Credit or Blame? Attributing Accountability in Modern AI Systems [PDF] (under review)
Shichang Zhang, Hongzhe Du, Jiaqi Ma, Himabindu Lakkaraju

[13] Generalized Group Data Attribution: A Novel Framework for Efficient Data Valuation [PDF] (under review)
Dan Ley, Shichang Zhang, Suraj Srinivas, Gili Rusak, Himabindu Lakkaraju

[12] On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models [PDF] (under review)
Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, Himabindu Lakkaraju
Featured in OpenAI o1 System Card Report

[11] Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models [PDF] (under review)
Chirag Agarwal, Sree Harsha Tanneru, Himabindu Lakkaraju
Featured in OpenAI o1 System Card Report

[10] Manipulating Large Language Models to Increase Product Visibility [PDF] (under review)
Aounon Kumar, Himabindu Lakkaraju
Featured in The New York Times, The Guardian, Communications of the ACM, and Towards Data Science

[9] Towards Unifying Interpretability and Control: Evaluation via Intervention [PDF] (under review)
Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, Himabindu Lakkaraju

[8] OpenHEXAI: An Open-Source Framework for Human-Centered Evaluation of Explainable Machine Learning [PDF] (under review)
Jiaqi Ma, Vivian Lai, Yiming Zhang, Chacha Chen, Paul Hamilton, Davor Ljubenkov, Himabindu Lakkaraju, Chenhao Tan

[7] Generalizing Trust: Weak-to-Strong Trustworthiness in Language Models [PDF] (working paper)
Martin Pawelczyk, Lillian Sun, Zhenting Qi, Aounon Kumar, Himabindu Lakkaraju

[6] Operationalizing the Blueprint for an AI Bill of Rights: Recommendations for Practitioners, Researchers, and Policy Makers [PDF] (working paper)
Alex Oesterling, Usha Bhalla, Suresh Venkatasubramanian, Himabindu Lakkaraju

[5] When Algorithms Explain Themselves: AI Adoption and Accuracy of Experts' Decisions [PDF] (working paper)
Himabindu Lakkaraju, Chiara Farronato

[4] Human vs. LLM Evaluators: A Comparative Study of Knowledge Work Assessment (working paper)
David Zollikofer, Chirag Agarwal, Aounon Kumar, Fabrizio Dell'Acqua, Karim Lakhani, Himabindu Lakkaraju

[3] Can Model Explanations Help Reduce Biases in Real-World Decision Making? (working paper)
Himabindu Lakkaraju, Paul Hamilton, Sarah Tan

Patents

[2] Extraction and grouping of feature words
Chiranjib Bhattacharyya, Himabindu Lakkaraju, Kaushik Nath, Sunil Arvindam
US8484228 B2

[1] Enhancing knowledge bases using rich social media
Jitendra Ajmera, Shantanu Ravindra Godbole, Himabindu Lakkaraju, Bernard Andrew Roden, Ashish Verma
US10192458 B2

Advising & Mentoring

Current Advisees:

Martin Pawelczyk, Postdoctoral Fellow, Harvard University	2023 - Present
Shichang Zhang, Postdoctoral Fellow, Harvard University	2024 - Present
Dan Ley, PhD Student, Harvard CS	2022 - Present
Alex Oesterling, PhD Student, Harvard CS	2022 - Present
Usha Bhalla, PhD Student, Harvard CS	2022 - Present
Zidi Xiong, PhD Student, Harvard CS	2024 - Present
Jenny Wang, PhD Student, Harvard Business School	2024 - Present
Zhenting Qi, PhD Student, Harvard University	2024 - Present

Past Advisees and Interns:
Jiaqi Ma (Postdoc, Harvard University => Assistant Professor, UIUC)
Chirag Agarwal (Postdoc, Harvard University => Assistant Professor, University of Virginia)
Suraj Srinivas (Postdoc, Harvard University => Research Scientist, Robert Bosch AI)
Aounon Kumar (Postdoc, Harvard University => Research Scientist, Samsung Research)
Dylan Slack (PhD, UC Irvine => Research Scientist, Google DeepMind)
Satyapriya Krishna (PhD, Harvard University => Research Scientist, Amazon)
Tessa Han (PhD, Harvard University => Postdoc, Harvard Medical School)
Aaron Li (MS, Harvard University => PhD Student, UC Berkeley EECS)
Yanchen Liu (MS, Harvard University => PhD Student, MIT IDSS)
Aditya Karan (MS, Harvard University => PhD Student, UIUC CS)
Sree Harsha Tanneru (MS, Harvard University => Research Engineer, Google DeepMind)
Kaivalya Rawal (MS, Harvard University => Research Fellow, Oxford University)
Alexis Ross (Undergraduate, Harvard University => PhD Student, MIT EECS)
Isha Puri (Undergraduate, Harvard University => PhD Student, MIT EECS)
Emily Jia (Undergraduate, Harvard University => Data Scientist, Figma)
Eshika Saxena (Undergraduate, Harvard University => AI Research Engineer, Meta)
Catherine Huang (Undergraduate, Harvard University => Quant Trader, IMC Trading)
Charu Badrinath (Undergraduate, Harvard University => Engineer, Palantir Technologies)
Christina Xiao (Undergraduate, Harvard University => Engineer, Bloomberg)
Umang Bhatt (Research Intern, Harvard University => Faculty Fellow, NYU CDS)
Ruijiang Gao (Research Intern, Harvard University => Assistant Professor, UT Dallas)
Harvineet Singh (Research Intern, Harvard University => Postdoc, UCSF/UC Berkeley)
Jessica Dai (Research Intern, Harvard University => PhD Student, UC Berkeley EECS)
Tom Suhr (Research Intern, Harvard University => PhD Student, Max Planck Institute)

Teaching Experience

Instructor, Explainable Artificial Intelligence — 2019 - Present
Department of Computer Science, Harvard University
(First ever full-fledged course on this topic)

	Instructor, Introduction to Data Science and Machine Learning Harvard Business School	2020 - Present
	Instructor, A Short Course on Explainable Machine Learning Stanford Center for AI Safety	2022
	Instructor, Introduction to ML for Social Scientists Harvard Business School	Spring 2020
	Instructor, Explainable and Accurate AI for High-Stakes Decision Making Harvard Online Analytics Program	2020 - 2023
	Guest Lecture, Explainable ML in the Era of Foundation Models Cornell University: Algorithmic Fairness Course	Spring 2024
	Guest Lecture, User Evaluations in Explainable Machine Learning UC Berkeley: Human-Centered AI Course	Spring 2023
	Guest Lecture, Explainable ML in the Era of Foundation Models Carnegie Mellon University: Trustworthy AI Course	Spring 2023
	Guest Lecture, Evaluating ML Models in the Presence of Unobservables Stanford University: Counterfactuals: The Science of What Ifs?	Spring 2021
	Guest Lecture, An Overview of Explainable Machine Learning Harvard University: AI for Social Impact Course	Spring 2021
	Guest Lecture, Algorithms for Explainable Machine Learning Carnegie Mellon University: Advanced Introduction to Machine Learning Course	Autumn 2020
	Guest Lecture, Explainable Machine Learning in Practice Carnegie Mellon University: Human-AI Interaction Course	Autumn 2020
	Guest Lecture, Introduction to Data Science, Stanford Law School	Spring 2016
	Guest Lecture, Algorithms for Submodular Optimization Stanford University: Mining Massive Data Sets Course	Winter 2016
	Co-instructor, Introduction to Python Programming Stanford University: Girls Teaching Girls to Code (GTGTC) Initiative	Spring 2015
	Teaching Assistant for Stanford University: Mining Massive Data Sets Course Stanford University: Social & Information Network Analysis Course Indian Institute of Science: Machine Learning Course	 Winter 2016 Autumn 2014 Autumn 2010
Tutorials	Trustworthy Machine Learning in the Era of Foundation Models	ICML, FAccT, KDD 2023
	Model Monitoring in Practice: Lessons Learned and Open Challenges	KDD, FAccT 2022
	Explainable ML in the Wild: When Not to Trust Your Explanations	FAccT 2021
	Explainable ML: Understanding the Limits and Pushing the Boundaries (Invited Tutorial)	CHIL 2021
	Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities	NeurIPS 2020, AAAI 2021
Selected Keynote & Invited Talks	**Keynote** at NeurIPS Workshop on Trustworthy Foundation Models	2025
	Keynote at Google Agentic AI Safety Workshop	2025
	Pinterest **Distinguished Lecture**	2025
	UC Berkeley School of Information	2025
	Stanford Management Science and Engineering	2025
	University of Washington School of Information	2025
	New York University Computer Science Department and Center for Data Science	2025
	Frontier AI Safety and Policy Panel by MIT and UK AI Safety Institute	2024

Keynote at EMNLP Workshop on BlackboxNLP	2024
Keynote at CIKM Workshop on Generative AI for E-Commerce	2024
Learning Machines Seminar, Cornell University	2024
First Annual Summit on Responsible Computing, AI, and Society, Georgia Tech	2024
Princeton University Workshop on Understanding in Natural and Artificial Minds	2024
Johns Hopkins CS Seminar Series	2024
US Securities and Exchange Commission	2024
MIT Data Science Seminar Series	2024
UPenn Center for Safe, Explainable, and Trustworthy AI Seminar Series	2024
AAAI Workshop on Privacy-Preserving Artificial Intelligence	2024
NSF Workshop on Advanced Automated Systems, Contestability, and the Law	2024
Google, Stanford, and UW Madison Workshop on Securing the Future of GenAI	2023
Yale and Google Joint Workshop on Theory and Practice of Foundation Models	2023
ICML Workshop on Interpretable ML in Healthcare	2023
ICML Workshop on Counterfactuals in Minds and Machines	2023
ICLR Workshop on Trustworthy & Reliable Large-Scale Machine Learning Models	2023
RSS Workshop on Safe Autonomy	2023
Mind and Machine Intelligence Summit, UC Santa Barbara	2023
Cornell University and Weill Cornell Medicine	2023
Kavli Frontiers of Science Symposium	2023
Cohere AI	2023
Keynote at AAAI Workshop on Representation Learning for Responsible Human-Centric AI	2023
Keynote at AAAI Workshop on Deployable AI	2023
INFORMS Annual Meeting	2016 - 2023
NeurIPS Workshop on Women in Machine Learning (WiML)	2022
NeurIPS Workshop on Machine Learning for Health (ML4H)	2022
ICLR Workshop on Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data	2022
CVPR Workshop on Explainable AI for Computer Vision	2022
Keynote at WWW Workshop on Explainable AI in Health	2022
ECCV Workshop on Adversarial Robustness in the Real World	2022
Panel Discussion on AI and the Economy, Jointly Organized by U.S. Department of Commerce, NIST, Stanford HAI, and the FinRegLab	2022
Simons Institute (Berkeley) Workshop on Societal Considerations and Applications	2022
Stanford Center for AI Safety Workshop on Explainable AI	2022
Stanford Human-Centered Artificial Intelligence (HAI) Conference	2022
Stanford Digital Econ Seminar	2022
MIT Initiative on the Digital Economy (IDE) Seminar Series	2022
Amazon Alexa Rising Star Speaker Series	2022
Keynote at ACM CIKM Conference	2021
NIST AI Risk Management Framework Workshop	2021
Pinterest **Distinguished Lecture**	2021
NeurIPS Workshop on Algorithmic Fairness through the Lens of Causality and Robustness	2021
NeurIPS Workshop on Explainable AI Approaches for Debugging and Diagnosis	2021
NeurIPS Workshop on Human and Machine Decisions	2021
Keynote at ICML Workshop on Interpretable ML in Healthcare	2021
Keynote at KDD Workshop on ML in finance	2021
AI for Good Summit organized by International Telecommunications Union & the United Nations	2021
Keynote at CVPR Workshop on Responsible Computer Vision	2021
Keynote at ICLR Workshop on Responsible AI	2021
Keynote at ASPLOS Workshop on Systems Architecture for Robust, Safe, and Resilient Software	2021
Keynote at MLSys Workshop on Personalized Recommender Systems & Algorithms	2021
University of Cambridge	2021
Neurosym Webinar Series, Jointly Organized by UPenn, MIT, Caltech, and Stanford	2021

Voices of Data Science, UMass Amherst	2021
Max Planck Symposium on Computing and Society	2021
Keynote at CVPR Workshop on Fair, Data-Efficient and Trusted Computer Vision	2020
Keynote at MICCAI Workshop on Interpretability in Medical Imaging	2020
ETH - Center for Law and Economics, Zurich	2020
University of Michigan, Ann Arbor	2019
AI World Conference & Expo, Cambridge	2019
EmTech MIT Conference, Cambridge	2019
Google DeepMind Annual Summit, Cambridge	2019
Women in Machine Learning Workshop, Boston	2019
ICLR Workshop on Safe Machine Learning, New Orleans	2019
Harvard Data Science Conference, Cambridge	2018
South Park Commons, San Francisco	2018
Computer Science Departmental Seminars at Carnegie Mellon University, UIUC, Harvard University, Georgia Tech, Yale University, UC San Diego, USC, UCLA, UC Irvine, Duke University, Brown University, University of Michigan, University of Maryland	2018
Machine Learning Departmental Seminar at Carnegie Mellon University	2018
Operations Research Departmental Seminars at Columbia University, Cornell University, Princeton University	2018
NYU Stern School of Business, New York	2018
MIT Sloan School of Management, Cambridge	2018
Harvard Business School, Boston	2018
UC Berkeley School of Public Health, San Francisco	2018
Microsoft Research, Redmond	2017, 2018
IBM Thomas J. Watson Research Center, New York	2017
Machine Learning Seminar at Duke University, Durham	2017
Keynote at ICML Workshop on Automatic Machine Learning, Sydney, Australia	2017
Stanford Biomedical Data Science Lecture Series, Palo Alto	2017
Stanford Symbolic Systems Coffee Chat Series, Palo Alto	2017
Stanford Data Science Workshop, Palo Alto	2017
Rising Stars Workshop in EECS, Pittsburgh	2016
CodeX Center, Stanford Law School, Palo Alto	2016
KDD Workshop on Data Science for Social Good, New York	2014
University of Chicago Computation Institute, Chicago	2014
Grace Hopper India Chapter, Bangalore, India	2011

Community Service

Co-Founder & Chair: Trustworthy & Regulatable ML Initiatives — 2020 - Present

We launched these initiatives to enable easy access to resources on these topics, to showcase and promote the work of researchers from underrepresented groups, and to build a community of researchers and practitioners working on these topics.

Panelist and Reviewer: — 2020 - Present

4 National Science Foundation (NSF) Review Panels, Directorate for Computer and Information Science and Engineering (CISE)

Conference Organization:

NeurIPS Conference (Ethics Chair)	2024 - 2025
WSDM Conference (Tutorial Chair)	2024
FAccT Conference (Sponsorship Chair)	2023
KDD Trustworthy AI Day (Program Chair)	2022
KDD Deep Learning Day (Program Chair)	2021
Grace Hopper India Conference (Program Chair)	2011

Workshop Chair:

NeurIPS Workshop on Regulatable Machine Learning	2023 - 2024
NeurIPS Workshop on Explainable Artificial Intelligence	2023 - 2024
ICML Workshop on New Frontiers in Adversarial Machine Learning	2022
ICML Workshop on Algorithmic Recourse	2021
ELLIS Human-Centric Machine Learning Workshop	2021

Session on Trustworthy Machine Learning at INFORMS 2020
Session on Fairness in Machine Learning at INFORMS 2019
ICLR Workshop on Debugging Machine Learning Models 2019
Workshop for spreading awareness about STEM fields among middle school girls 2016
Stanford's Girls Teaching Girls To Code (GTGTC) 2015

Area Chair:

ICML - *International Conference on Machine Learning* 2019 - 2024
NeurIPS - *Advances in Neural Information Processing Systems* 2019 - 2023
ICLR - *International Conference on Learning Representations* 2020 - 2023
AISTATS - *International Conference on Artificial Intelligence and Statistics* 2021 - 2023

Program Committee:

AISTATS - *International Conference on Artificial Intelligence and Statistics* 2019 - 2020
FAccT - *ACM Conference on Fairness, Accountability, and Transparency* 2019 - 2020
AAAI - *AAAI International Conference on Artificial Intelligence* 2019
ICML - *International Conference on Machine Learning* 2018
ICLR - *International Conference on Learning Representations* 2018 - 2019
IJCAI - *International Joint Conference on Artificial Intelligence* 2018 - 2019
WWW - *International World Wide Web Conference* 2017 - 2018
NIPS - *Advances in Neural Information Processing Systems* 2016 - 2017
KDD - *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2015 - 2017
CIKM - *ACM Conference on Information and Knowledge Management* 2011, 2017
SDM - *SIAM International Conference on Data Mining* 2015
UAI - *Conference on Uncertainty in Artificial Intelligence* 2011
AAAI - *AAAI conference on Artificial Intelligence* 2011

Journal Reviewing and Editing:

Frontiers in Big Data (Associate Editor) 2021 - 2023
MS - *Management Science* 2021 - 2025
OR - *Operations Research* 2021 - 2025
JMLR - *Journal of Machine Learning Research* 2020 - 2023
TWEB - *ACM Transactions on the Web* 2017
PLOS ONE - *Public Library of Science ONE* 2017
TKDD - *ACM Transactions on Knowledge Discovery from Data* 2016
TKDE - *IEEE Transactions on Knowledge and Data Engineering* 2015

Other:

Member, Faculty Hiring Committee, Harvard Business School 2020 - 2025
Member, Ph.D. Selection Committee, Harvard CS & Business School 2020 - 2025
Member, Ph.D. Selection Committee, Stanford CS 2016

Media Coverage

The New York Times: How Do You Change a Chatbot's Mind?
TIME: Chuck Schumer wants AI to be explainable. It's harder than it sounds
The Guardian: The chatbot optimisation game: can we trust AI web searches?
Communications of the ACM: When LLMs Learn to Lie
Towards Data Science: Can Recommendations from LLMs Be Manipulated?
Science News: AI chatbots can be tricked into misbehaving. Can scientists stop it?
Boston Globe: A technophobe's guide to AI chatbots
HBS Working Knowledge: How Some 'Gibberish' Code Can Give Products an Edge
Fortune: What's wrong with explainable A.I.
HBS Working Knowledge: Why Technology Alone Can't Solve AI's Bias Problem
Harvard Business Review: The AI transparency paradox
Forbes: Information Technology Powers (Almost) All Innovation
PR Newswire: MIT Technology Review Announces 2019 Innovators Under 35
deeplearning.ai: Bias Goes Undercover: Adversarial attacks can fool explainable AI
MIT Technology Review: How to upgrade judges with machine learning
Harvard Business Review: Solving social problems with machine learning
The New York Times: Even Imperfect Algorithms Can Improve the Criminal Justice System

VentureBeat: Confidence, uncertainty, and trust in AI affect how humans make decisions
Wired: This Agency Wants to Figure Out Exactly How Much You Trust AI
Bloomberg Technology: Researchers combat gender and racial bias in AI
Forbes: How to craft the perfect Reddit posting
TIME: How to succeed on Reddit
Business Insider: How to execute the perfect Reddit submission
Business Insider: How a title can sink or float a piece of content
Phys.org: Stanford Trio explore success formula for Reddit posts
International Business Times: The secret to what makes something go viral
New Scientist: Things that make a meme explode
The Verge: The math behind successful Reddit submissions
ACM TechNews: Stanford trio explore success formula for Reddit posts
Gizmodo: This equation can tell you how successful a reddit post can be
GigaOm: How to maximize your reddit upvotes, by the numbers