# Himabindu Lakkaraju

| | |
|---|---|
| **Contact Information** | 491 Morgan Hall<br>15 Harvard Way<br>Boston, MA 02163<br>*E-mail:* hlakkaraju@hbs.edu; hlakkaraju@seas.harvard.edu<br>*Webpage:* http://himalakkaraju.github.io |
| **Research Interests** | Trustworthy Machine Learning (Interpretability, Fairness, Robustness, and Privacy); Large Language Models; Human-AI Interaction; Applications of AI/ML to Decision Making in Healthcare, Law, and Policy. |

**Academic & Professional Experience**

**Harvard University**  *01/2020 - Present*
*Assistant Professor* with appointments in the Business School and the Department of Computer Science (Affiliate)
Faculty Affiliate, Harvard Data Science Initiative

**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*

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

**Harvard University**  *11/2018 - 12/2019*
*Postdoctoral Fellow*, Business School & Department of Computer Science

**Stanford University**  *9/2012 - 9/2018*
*Research Assistant*, Department of Computer Science

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

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

**IBM Research**  *7/2010 - 7/2012*
*Research Engineer*

**Education**

**Stanford University**  *9/2012 - 9/2018*
Ph.D. in Computer Science

**Stanford University**  *9/2012 - 9/2015*
Master of Science (MS) in Computer Science

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

**Selected Honors & Achievements**

| | |
|---|---|
| **NSF CAREER Award** | 2023 |
| **AI2050 Early Career Fellowship** by Schmidt Futures | 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**<br>**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 |
| **Advisory Board Member**, **The Stanford Center for Legal Informatics** | 2020 - Present |
| **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 one of the **Rising Stars 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**

| | |
|---|---|
| NSF CAREER Award (US$550,664) – Sole PI | 2023 - 2028 |
| AI2050 Early Career Fellowship by Schmidt Futures (US$300,00) – Sole PI | 2023 - 2026 |
| Adobe Data Science Research Award (US$50,000) – 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**: **6634**          **h-index**: **34**          **i10-index**: **46**

(* below indicates equal contribution)

**Book Chapters**

[71] Analyzing Human Decisions and Machine Predictions in Bail Decision Making
Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan
(author names are ordered alphabetically)
The Inequality Reader: Contemporary and Foundational Readings in Race, Class, and Gender; Third Edition, 2022.

**Articles in Peer-Reviewed Journals**

[70] 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.**

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

[68] 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**

[67] 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**

[66] 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**

[65] 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.

[64] 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.

[63] 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%)**

[62] 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.

[61] 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.

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

[59] 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%)**

[58] 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.

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

[56] 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 in Healthcare, 2022.**

[55] 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.*

[54] 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.*

[53] 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%)**

[52] 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.*

[51] 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.*

[50] 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.*

[49] 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.*

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

[47] 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**.

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

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

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

[43] 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.

[42] 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%)**

[41] 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.*

[40] 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.*

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

[38] 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%)**

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

[36] 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%)**

[35] 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%)**

[34] 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%)**

[33] 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%)**

[32] 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**

[31] 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**

[30] 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.*

[29] 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.*

[28] 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%)**

[27] 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%)**

[26] 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.*

[25] 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%)**

[24] 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%)**

[23] Attention prediction on social media brand pages
Himabindu Lakkaraju, Jitendra Ajmera
*CIKM - ACM Conference on Information and Knowledge Management, 2011.*

[22] 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**

[21] TEM: A novel perspective to modeling content on microblogs
Himabindu Lakkaraju, Hyung-Il-Ahn
*WWW - International World Wide Web Conference, 2011.*

[20] 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**

[19] The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective [PDF] (under review)
Satyapriya Krishna*, Tessa Han*, Alex Gu, Shahin Jabbari, Steven Wu, Himabindu
Lakkaraju
Preliminary version presented at CHI Workshop on Trust and Reliance in Human-AI
Teams, 2022; **Featured in Fortune Magazine**.

[18] In-Context Unlearning: Language Models as Few Shot Unlearners [PDF]
(under review)
Martin Pawelczyk, Seth Neel, Himabindu Lakkaraju

[17] Are Large Language Models Post Hoc Explainers? [PDF] (under review)
Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju

[16] Certifying LLM Safety against Adversarial Prompting [PDF] (under review)
Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Li, Soheil Feizi, Himabindu
Lakkaraju

[15] Quantifying Uncertainty in Natural Language Explanations of Large Language Models
[PDF] (under review)
Sree Harsha Tanneru, Chirag Agarwal, Himabindu Lakkaraju

[14] Accurate, Explainable, and Private Models: Providing Recourse While Minimizing Training Data Leakage [PDF] (under review)
Catherine Huang, Chelsea Swoopes, Christina Xiao, Jiaqi Ma, Himabindu Lakkaraju

[13] Efficient Estimation of the Local Robustness of Machine Learning Models [PDF] (under review)
Tessa Han, Suraj Srinivas, Himabindu Lakkaraju

[12] Analyzing chain-of-thought prompting in Large language models via gradient-based feature Attributions [PDF] (under review)
Skyler Wu, Eric Shen, Charumathi Badrinath, Jiaqi Ma, Himabindu Lakkaraju

[11] Rethinking Explainability as a Dialogue: A Practitioner's Perspective [PDF] (under review)
Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, Sameer Singh
Preliminary version presented at NeurIPS Workshop on Human-Centered AI, 2022.

[10] On the Impact of Adversarially Robust Models on Algorithmic Recourse [PDF] (under review)
Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju
Preliminary version presented at NeurIPS Workshop on Trustworthy and Socially Responsible ML, 2022.

[9] Rethinking Stability for Attribution-Based Explanations [PDF] (under review)
Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, Himabindu Lakkaraju
Preliminary version presented at ICLR PAIR2Struct Workshop, 2022.

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

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

[6] Enforcing Right to Explanation: Bridging the Gaps between ML Research and Policy (working paper)
Himabindu Lakkaraju, Jiaqi Ma

[5] On the Incompatibility between AI Regulatory Guidelines (working paper)
Paul Hamilton, Davor Ljubenkov, Jiaqi Ma, Himabindu Lakkaraju

[4] An Empirical Study of the Trade-offs between Interpretability and Fairness [PDF]
Shahin Jabbari, Han-Ching Ou, Himabindu Lakkaraju, Milind Tambe
ICML Workshop on Human Interpretability in Machine Learning, 2020

[3] Aspect Specific Sentiment Analysis using Hierarchical Deep Learning [PDF]
Himabindu Lakkaraju, Richard Socher, Christopher Manning
NIPS Workshop on Deep Learning and Representation Learning, 2014

**Patents**

[2] Extraction and Grouping of Feature Words
Chiranjib Bhattacharyya, Himabindu Lakkaraju, Sunil Aravindam, Kaushik Nath
US8484228 B2

[1] Enhancing knowledge bases using rich social media
Jitendra Ajmera, Shantanu Godbole, Himabindu Lakkaraju, Ashish Verma
US20130224714 A1

| **Advising & Mentoring** | **Current Advisees**: | |
|---|---|---|
| | Suraj Srinivas, Postdoctoral Fellow, Harvard University | 2022 - Present |
| | Aounon Kumar, Postdoctoral Fellow, Harvard University | 2023 - Present |
| | Chirag Agarwal, Postdoctoral Fellow, Harvard University | 2020 - Present |
| | Martin Pawelczyk, Postdoctoral Fellow, Harvard University | 2023 - Present |
| | Tessa Han, PhD Student, Harvard University | 2020 - Present |
| | Satyapriya Krishna, PhD Student, Harvard University | 2021 - Present |
| | Dan Ley, PhD Student, Harvard University | 2022 - Present |
| | Alex Oesterling, PhD Student, Harvard University | 2022 - Present |
| | Usha Bhalla, PhD Student, Harvard University | 2022 - Present |
| | Paul Hamilton, PhD Student, Harvard Univeristy | 2023 - Present |
| | Elita Lobo, PhD Student, UMass Amherst | 2023 - Present |
| | Sree Harsha Tanneru, Masters Student, Harvard University | 2023 - Present |
| | Nikhil Nayak, Masters Student, Harvard University | 2023 - Present |
| | Aaron Li, Masters Student, Harvard University | 2023 - Present |
| | Yanchen Liu, Masters Student, Harvard University | 2023 - Present |
| | Charu Badrinath, Undergrad, Harvard University | 2023 - Present |
| | Eric Shen, Undergrad, Harvard University | 2023 - Present |
| | Catherine Huang, Undergrad, Harvard University | 2023 - Present |
| | Christina Xiao, Undergrad, Harvard University | 2023 - Present |

**Past Advisees and Interns**:
Jiaqi Ma (Postdoc, Harvard University => Assistant Professor, UIUC CS & IS)
Dylan Slack (PhD, UC Irvine => Research Scientist, Scale AI)
Sophie Hilgard (PhD, Harvard University => Research Scientist, Twitter)
Aditya Karan (MS, Harvard University => PhD Student, UIUC CS)
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)
Umang Bhatt (Research Intern, Harvard University => Assistant Professor, NYU CDS)
Harvineet Singh (Research Intern, Harvard University => Postdoc, UCSF/UC Berkeley)
Jessica Dai (Research Intern, Harvard University => PhD Student, UC Berkeley EECS)

| **Teaching Experience** | Instructor, Explainable Artificial Intelligence<br>Department of Computer Science, Harvard University<br>**First ever course on this emerging topic** | 2020 - 2023 |
|---|---|---|
| | Instructor, Introduction to Data Science and Machine Learning<br>Harvard Business School | 2020 - 2023 |
| | Instructor, A Short Course on Explainable Machine Learning<br>Stanford Center for AI Safety | 2022 |
| | Instructor, Introduction to ML for Social Scientists<br>Harvard Business School | Spring 2020 |
| | Instructor, Explainable and Accurate AI for High-Stakes Decision Making<br>Harvard Online Analytics Program | 2020 - 2022 |
| | Guest Lecture, User Evaluations in Explainable Machine Learning<br>UC Berkeley: Human-Centered AI Course | Spring 2023 |
| | Guest Lecture, Explainable ML in the Era of Foundation Models<br>Carnegie Mellon University: Trustworthy AI Course | Spring 2023 |
| | Guest Lecture, Evaluating ML Models in the Presence of Unobservables<br>Stanford University: Counterfactuals: The Science of What Ifs? | Spring 2021 |
| | Guest Lecture, Explainable Machine Learning<br>Harvard University: AI for Social Impact Course | Spring 2021 |

Guest Lecture, Explainable Machine Learning　　　　　　　　　　Autumn 2020
Carnegie Mellon University: Advanced Introduction to Machine Learning Course

Guest Lecture, Explainable Machine Learning in Practice　　　　　Autumn 2020
Carnegie Mellon University: Human-AI Interaction Course

Guest Lecture, Introduction to Data Science, Stanford Law School　　Spring 2016

Guest Lecture, Algorithms for Submodular Optimization　　　　　　Winter 2016
Stanford University: Mining Massive Data Sets Course

Co-instructor, Introduction to Python Programming　　　　　　　　Spring 2015
Stanford University: Girls Teaching Girls to Code (GTGTC) Initiative

Teaching Assistant for
Stanford University: Mining Massive Data Sets Course　　　　　　　Winter 2016
Stanford University: Social & Information Network Analysis Course　　Autumn 2014
Indian Institute of Science: Machine Learning Course　　　　　　　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　CHIL 2021
　　　　　　　(Invited Tutorial)

　　　　　　　Explaining Machine Learning Predictions:　　　　　　　　　NeurIPS 2020
　　　　　　　State-of-the-art, Challenges, and Opportunities　　　　　　　AAAI 2021

**Invited Talks**　MIT Data Science Laboratory　　　　　　　　　　　　　　　　2023
**& Panel Discussions**　US Securities and Exchange Commission　　　　　　　　　　　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　　　2023
　　　　　　　Responsible Human-Centric AI
　　　　　　　**Keynote** at AAAI Workshop on Deployable AI　　　　　　　　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,　2022
　　　　　　　Reasoning on Structured Data
　　　　　　　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　　2022
　　　　　　　U.S. Department of Commerce, NIST, Stanford HAI, and the FinRegLab
　　　　　　　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
　　　　　　　Harvard Data Science Initiative's Annual Conference　　　　　　2022
　　　　　　　Berkman Klein Center, Harvard University　　　　　　　　　　2022
　　　　　　　Amazon Alexa Rising Star Speaker Series　　　　　　　　　　2022
　　　　　　　University of Southern California　　　　　　　　　　　　　　2022

| | |
|---|---|
| Fireside Chat on Explainability, Fiddler AI | 2022 |
| INFORMS Annual Meeting | 2016 - 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 |
| Harvard CRCS Seminar, Cambridge | 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 ML Initiative | 2020 - Present |

We launched this initiative to enable easy access to resources on trustworthy ML, to showcase and promote the work of researchers from underrepresented groups, and to build a community of researchers and practitioners working on the topic.

**Panelist and Reviewer:**                                                    2020 - Present
4 National Science Foundation (NSF) Review Panels,
Directorate for Computer and Information Science and Engineering (CISE)

**Co-Chair:**

| | |
|---|---|
| NeurIPS Workshop on Regulatable Machine Learning | 2023 |
| NeurIPS Workshop on Explainable Artificial Intelligence | 2023 |
| KDD Trustworthy AI Day | 2022 |
| ICML Workshop on New Frontiers in Adversarial Machine Learning | 2022 |
| KDD Deep Learning Day | 2021 |
| 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 |
| Grace Hopper India Conference | 2011 |

**Sponsorship Chair:**

| | |
|---|---|
| FAccT - ACM Conference on Fairness, Accountability, and Transparency | 2023 |

**Tutorial Chair:**

| | |
|---|---|
| WSDM - ACM Conference on Web Search and Data Mining | 2024 |

**Area Chair:**

| | |
|---|---|
| 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 |
| ICML - *International Conference on Machine Learning* | 2019 - 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 |
| JMLR - *Journal of Machine Learning Research* | 2020 - 2023 |
| MS - *Management Science* | 2021 - 2023 |
| OR - *Operations Research* | 2021 - 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 University | 2020 - 2023 |
| Member, Ph.D. Student Selection Committee, Harvard University | 2020 - 2023 |
| Member, Ph.D. Student Selection Committee, Stanford University | 2016 |

**Selected Media Coverage**

TIME: Chuck Schumer wants AI to be explainable. It's harder than it sounds
Fortune: Explainable AI & The Disagreement Problem
Harvard Business Review: The AI transparency paradox
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
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