Himabindu Lakkaraju

Contact Information	Harvard Business School Soldiers Field Road Boston, MA 02163 E-mail: hlakkaraju@hbs.edu; hlakkaraju@seas.harvard.edu Webpage: http://himalakkaraju.github.io		
Research Interests	Transparency, Fairness, and Safety in Artificial Intelligence (AI); Applications of AI to Law, Healthcare, Public Policy, and Education; AI for Decision-Making.		
Academic & Professional Experience	Harvard University Assistant Professor with appointments in Business School and Department of Computer Science	01/	2020 -
	Harvard University Postdoctoral Fellow at Harvard Business School	11/2018 - 1	2/2019
	Stanford University Research Assistant in the Department of Computer Science	9/2012 -	9/2018
	Microsoft Research, Redmond Visiting Researcher	5/2017 -	6/2017
	Microsoft Research, Redmond Research Intern	6/2016 -	9/2016
	University of Chicago Data Science for Social Good Fellow	6/2014 -	8/2014
	IBM Research - India, Bangalore Technical Staff Member	7/2010 -	7/2012
	SAP Research, Bangalore Visiting Researcher	7/2009 -	3/2010
	Adobe Systems Pvt. Ltd., Bangalore Software Engineer	7/2007 -	7/2008
Education	Stanford University Ph.D. in Computer Science Thesis: Enabling Machine Learning for High-Stakes Decision-Making	9/2012 -	9/2018
	Stanford University Master of Science (MS) in Computer Science	9/2012 -	9/2015
	Indian Institute of Science (IISc) Master of Engineering (MEng) in Computer Science & Automation Thesis: Exploring Topic Models for Understanding Sentiments Express Customer Reviews	8/2008 - sed in	7/2010
Selected Honors &	Amazon Research Award		2021
Achievements	National Science Foundation (NSF) Amazon Fairness in Al Grant		2021
	Google 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 by MIT Tech Review	2019
Named as one of the Innovators to Watch by Vanity Fair	2019
Selected for the prestigious Cowles Fellowship by Yale University (declined)	2018
INFORMS Data Mining Best Paper Award "Learning Cost-Effective and Interpretable Treatment Regimes"	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	2013-17
Eminence and Excellence Award for outstanding contributions to research IBM Research	2012
Research Division Award recognizing research contributions IBM Research	2012
Best Paper Award , SIAM International Conference on Data Mining (SDM) "Exploiting Coherence for the Simultaneous Discovery of Latent Facets and associated Sentiments"	2011
SPOT Award for outstanding product contributions Adobe Systems Pvt. Ltd.	2009
All India Rank 32 (99.82%ile) Graduate Aptitude Test in Engineering (GATE) Entrance examination for IISc & IITs in Computer Science & Engineering	2008
University Rank 10 , Bachelor of Engineering, Computer Science Out of 8000 students from 175 colleges	2007
As Faculty	
NSF-Amazon Fairness in AI (FAI) grant (US\$375,000) – co-PI Amazon Faculty Research Award (US\$70,000) – Sole PI Google Faculty Research Award (US\$600,000) – PI National Science Foundation (NSF) RI Small (US\$500,000) – Harvard PI HDSI & Bayer Trust in Science Award (US\$100,000) – Joint PI	2021-24 2021-2024 2020-23 2020-23 2020-21
As Student	
Microsoft Research Dissertation Grant (US\$20,000) Women in Machine Learning (WiML) Travel Grant for NIPS (US\$850) ICML Travel Grant (US\$1800) KDD Travel Grant (US\$1000 p.a.) Stanford Graduate Fellowship (tuition + US\$41,700 p.a.) NIPS Travel Grant (US\$1000) Google Anita Borg Scholarship (US\$10,000) Facebook Graduate Fellowship Finalist (US\$500) Indian Institute of Science Graduate Scholarship (tuition + Rs.96,000 p.a.) SAP India Research Grant (Rs.150,000) Undergraduate Merit scholarship (Rs.3000 p.a.)	2017 2017 2017 2014 - 2017 2013 - 2017 2016 2015 2013 2008 - 2010 2009 - 2010 2004 - 2007
O	

Grants & Fellowships

Publications Total Citations: 2455

Articles in peer-reviewed journals

[36] 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

[35] Extracting Latent Personality Traits from Digital Footprints Michal Kosinski, Yilun Wang, **Himabindu Lakkaraju**, Jure Leskovec *Psychological Methods* - 2016

Articles in peer-reviewed conference proceedings

[34] 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

[33] Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses

Kaivalya Rawal, **Himabindu Lakkaraju**

NeurIPS - Advances in Neural Information Processing Systems, 2020

- [32] Incorporating Interpretable Output Constraints in Bayesian Neural Networks Wanqian Yang, Lars Lorch, Moritz Gaule, **Himabindu Lakkaraju**, Finale Doshi-Velez NeurlPS Advances in Neural Information Processing Systems, 2020
- [31] Robust and Stable Black Box Explanations
 Himabindu Lakkaraju, Nino Arsov, Osbert Bastani
 ICML International Conference on Machine Learning, 2020
 Invited Talk at INFORMS Annual Meeting, 2020
- [30] 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 **Invited Talk at INFORMS Annual Meeting, 2020**
- [29] 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 Invited Talk at INFORMS Annual Meeting, 2020
- [28] 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 Invited Talk at INFORMS Annual Meeting, 2017
- [27] 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

[26] 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 Invited Talk at INFORMS Annual Meeting, 2017 [25] 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

[24] 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

[23] 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

Invited Talk at INFORMS Annual Meeting 2016

[22] 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

[21] A Bayesian Framework for Modeling Human Evaluations **Himabindu Lakkaraju**, Jure Leskovec, Jon Kleinberg, Sendhil Mullainathan *SDM - SIAM International Conference on Data Mining*, 2015

[20] 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

[19] 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

[18] 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

[17] Attention prediction on social media brand pages

Himabindu Lakkaraju, Jitendra Ajmera

CIKM - ACM Conference on Information and Knowledge Management, 2011

[16] 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

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

[14] 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

Preprints

- [13] Towards Robust and Reliable Algorithmic Recourse Sohini Upadhyay*, Shalmali Joshi*, **Himabindu Lakkaraju** [PDF]
- [12] Towards a Unified Framework for Fair and Stable Graph Representation Learning Chirag Agarwal, **Himabindu Lakkaraju***, Marinka Zitnik*
 [PDF]
- [11] Towards the Unification and Robustness of Perturbation and Gradient Based Explanations

Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Steven Wu, **Himabindu Lakkaraju**[PDF]

[10] Can I Still Trust You?: Understanding the Impact of Distribution Shifts on Algorithmic Recourses

Kaivalya Rawal, Ece Kamar, **Himabindu Lakkaraju** [PDF]

- [9] Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring Tom Suhr, Sophie Hilgard, Himabindu Lakkaraju [PDF]
- [8] 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 Featured in VentureBeat [PDF]
- [7] Ensuring Actionable Recourse via Adversarial Training Alexis Ross, Himabindu Lakkaraju, Osbert Bastani [PDF]
- [6] How Much Should I Trust You? Modeling Uncertainty of Black Box Explanations Dylan Slack, Sophie Hilgard, Sameer Singh, Himabindu Lakkaraju [PDF]
- [5] Manipulated Counterfactuals: Risks of Hidden Biases in Algorithmic Recourse Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, Sameer Singh [PDF]
- [4] Towards Mixed Robustness: Learning Under Adversarial and Interventional Shifts Harvineet Singh, Shalmali Joshi, Finale Doshi-Velez, **Himabindu Lakkaraju** [PDF]
- [3] An Empirical Study of the Trade-Offs Between Interpretability and Fairness Shahin Jabbari, Han-Ching Ou, Milind Tambe, **Himabindu Lakkaraju** [PDF]

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, Ben Roden

US20130224714 A1

Advising & Mentoring	Shalmali Joshi, Postdoctoral Fellow, Harvard University Chirag Agarwal, Postdoctoral Fellow, Harvard University Shahin Jabbari, Postdoctoral Fellow, Harvard University Haipeng Chen, Postdoctoral Fellow, Harvard University Sohini Upadhyay, PhD Student, Harvard University Sophie Hilgard, PhD Student, Harvard University Maya Balakrishnan, PhD Student, Harvard University Dylan Slack, PhD Student, UC Irvine Aida Rahmttalabi, PhD Student, USC Kaivalya Rawal, MS Student, Harvard University Alexis Ross, Undergrad, Harvard University Aditya Karan, MS Student, Harvard University Jorma Gorns, Undergrad, Harvard University Emily Jia, Undergrad, Harvard University Wanqian Yang, Undergrad, Harvard University	2020 - Present 2020 - Present 2019 - Present 2020 - Present 2020 - Present 2019 - 2020 2020 - Present 2019 - 2020 2019 - Present 2019 - Present 2019 - Present 2019 - 2020 2019 - 2020 2019 - 2020 2019 - 2020
	Nino Arsov, Visiting Researcher, Stanford University Rishabh Bhargava, MS Student, Stanford University	2016, 2019 - 2020 2015
	Yilun Wang, MS Student, Stanford University	2014 - 2015
	Mrinal Kanti Das, Ph.D. Student, Indian Institute of Science Hemant Purohit, Ph.D. Student, Wright State University	2011 2011
Teaching	Instructor, Interpretability and Explainability in ML Harvard CS & Harvard Business School	Fall 2019 &
Experience	First ever course on this emerging topic	Spring 2020
	Instructor, Technology and Operations Management Harvard Business School	Fall 2020
	Instructor, Explainable and Accurate AI for High-Stakes Decision Makin Harvard Business Analytics Program (HBAP)	ng Summer & Spring 2020
	Instructor, Introduction to ML for Social Scientists, Harvard Business Scientists Doctoral course on Empirical Technology and Operations Management	
	Guest Lecture, Introduction to Data Science, Stanford Law School	Spring 2016
	Co-instructor, Probability with Mathemagics, Stanford: Splash Initiative for High School Students	Spring 2016
	Teaching Assistant, Stanford: Mining Massive Data Sets (CS 246)	Winter 2016
	Guest Lecture, Algorithms for Submodular Optimization Stanford: Mining Massive Data Sets (CS 246)	Winter 2016
	Co-instructor, Introduction to Python Programming Stanford: Girls Teaching Girls to Code (GTGTC) for High School Stude	Spring 2015 ents
	Mathematics and Science Tutor DreamCatchers Nonprofit Organization, Palo Alto	Winter 2015
	Head Teaching Assistant, Stanford: Social & Information Network Analysis (CS 224W)	Autumn 2014
	Head Teaching Assistant, Indian Institute of Science: Machine Learning	Autumn 2010
Tutorials	Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities	AAAI 2021
	Explainable ML in the Wild: When Not to Trust Your Explanations	FAccT 2021
	Explainable ML: Understanding the Limits and Pushing the Boundaries Invited Tutorial	S CHIL 2021

Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities

NeurlPS 2020

Invited Talks	Keynote at CVPR Workshop on Responsible Computer Vision	2021
& Panel Discussions	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
	Voices of Data Science, UMass Amherst	2021
	Max Planck Symposium on Computing and Society	2021
	Machine Learning Department and Institute of Software Research at	2020
	Carnegie Mellon University	
	Keynote at CVPR Workshop on Fair, Data-Efficient and Trusted Computer Vision	2020
	Keynote at MICCAI Workshop on Interpretability in Medical Imaging	2020
	3 Invited Talks at INFORMS Annual Meeting	2020
	ETH - Center for Law and Economics, Zurich	2020
	University of Michigan, Ann Arbor	2019
	Harvard CRCS Seminar, Cambridge	2019
	INFORMS Annual Meeting, Seattle	2019
	Al 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
	Microsoft Research, Redmond	2018
	Computer Science Department at UCSD, San Diego	2018
	Computer Science Department at University of Michigan, Ann Arbor	2018
	Computer Science Department at Brown University, Providence	2018
	Computer Science Department at UIUC, Urbana Champaign	2018
	Computer Science Department at USC, Los Angeles	2018
	Machine Learning and Computer Science Departments at	
	Carnegie Mellon University, Pittsburgh	2018
	Computer Science Department at UCLA, Los Angeles	2018
	Computer Science Department at UCI, Irvine	2018
	Computer Science Department at Duke University, Durham	2018
	Computer Science Department at University of Maryland, College Park	2018
	NYU Stern School of Business, New York	2018
	Operations Research and Information Engineering Department at	
	Cornell University, Ithaca	2018
	Industrial Engineering and Operations Research Department at	
	Columbia University, New York	2018
	College of Computing at Georgia Tech, Atlanta	2018
	Computer Science Department at Harvard University, Cambridge	2018
	Computer Science Department at Yale University, New Haven	2018
	MIT Sloan School of Management, Cambridge	2018
	Harvard Business School, Boston	2018
	Operations Research and Financial Engineering Department at	
	Princeton University, Princeton	2018
	UC Berkeley School of Public Health, San Francisco	2018
	Microsoft Research, Redmond, USA	2017
	IBM Thomas J. Watson Research Center, New York	2017
	Machine Learning Seminar at Duke University, Durham	2017
	INFORMS Annual Meeting, Houston	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 Stanford Data Science Retreat, Palo Alto Workshop on Demystifying Artificial Intelligence, San Francisco Disruptive Innovation in Law Conference, Sydney, Australia Rising Stars Workshop, Pittsburgh Robert Bosch Research, Palo Alto INFORMS Annual Meeting, Nashville Stanford Data Science Retreat, Palo Alto Future Law: Watson and Beyond (Panel Discussion), Stanford Law School CodeX Center, Stanford Law School, Palo Alto KDD Workshop on Data Science for Social Good, New York University of Chicago Computation Institute, Chicago Stanford HCI Retreat, San Francisco Yahoo IR Summer School, Bangalore, India Indian Institute of Science Talk Series, Bangalore, India Grace Hopper India Chapter, Bangalore, India	2017 2017 2017 2017 2016 2016 2016 2016 2016 2014 2014 2014 2013 2011 2011
Co-Founder & Organizer: Trustworthy ML Initiative We launched this initiative to enable easy access to resources on trustworthy ML and to build a community of researchers and practitioners working on the topic.	
Organizer: ELLIS Human-Centric Machine Learning Workshop Session on Trustworthy Machine Learning at INFORMS Session on Fairness in Machine Learning at INFORMS Workshop on Debugging Machine Learning Models at International Conference on Learning Representations (ICLR) Workshop for spreading awareness about STEM fields among middle school girls Stanford's Girls Teaching Girls To Code (GTGTC) Women in Data Science for Social Good Group, UChicago Grace Hopper India Conference	2021 2020 2019 2019 2016 2015 2014 2011
· ·	9 - 2021 9 - 2021 2020
AAAI - AAAI International Conference on Artificial Intelligence ICML - International Conference on Machine Learning ICLR - International Conference on Learning Representations IJCAI - International Joint Conference on Artificial Intelligence WWW - International World Wide Web Conference NIPS - Advances in Neural Information Processing Systems KDD - ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2015 CIKM - ACM Conference on Information and Knowledge Management 201	9 - 2020 2019 2018 3 - 2019 7 - 2018 5 - 2017 5 - 2017 1, 2017 2016 2015 2011
Journal Reviewer: TWEB - ACM Transactions on the Web PLOS ONE - Public Library of Science ONE EJOR - European Journal of Operational Research TKDD - ACM Transactions on Knowledge Discovery from Data	2017 2017 2017 2016

Community Service

Other:

Mentor, Stanford Science Penpals

Member, Ph.D. Student Selection Committee, Stanford Computer Science

Mentor and Sponsor, Children International

Member, Stanford AI Women Group

2013 - Present
2014 - Present

Selected Media Coverage

Harvard Business Review: The Al 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

Bloomberg Technology: Researchers combat gender and racial bias in Al

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 Gizmodo: This equation can tell you how successful a reddit post can be

GigaOm: How to maximize your reddit upvotes, by the numbers

References

Jure Leskovec

Associate Professor

Department of Computer Science

Stanford University

Jon Kleinberg

Tisch University Professor Department of Computer Science Department of Information Science Cornell University

Cynthia Rudin

Professor

Department of Computer Science

Department of Electrical and Computer Engineering

Department of Statistical Science

Duke University

Milind Tambe

Gordon McKay Professor Department of Computer Science Harvard University

Ece Kamar

Senior Principal Area Research Manager Adaptive Systems and Interaction Group Microsoft Research

Eric Horvitz

Chief Scientific Officer Microsoft