Chirag Agarwal

Website: chirag-agarwall.github.io
Email: chiragagarwal@virginia.edu
Phone: (865)-406-2887

Academic & Professional Experience

University of Virginia
Assistant Professor — 2024 – Present
School of Data Science (primary appointment)
Department of Computer Science (by courtesy)
School of Medicine, Radiology and Medical Imaging (by courtesy)
Aikyam Lab

Harvard University
Postdoctoral Research Fellow — 2020–2022, 2023–2024
Host: Prof. Hima Lakkaraju and Prof. Marinka Zitnik

Adobe India
Research Scientist — 2022 – 2023

Auburn University
Research Assistant — Summer 2019

Robert Bosch LLC
Computer Vision/Augmented Reality Intern — Summer 2018

Tempus labs Inc.
Imaging Science Intern — Spring 2018

Kitware Inc.
Research and Development Intern — Summer 2017

Geisinger Health Systems
Research Intern — Summer 2016

Education

University of Illinois at Chicago
Ph.D. in Electrical and Computer Engineering — 2020

- Thesis: *Robustness and Explainability of Deep Neural Networks*
- Committee: Prof. Dan Schonfeld, Prof. Bharati Prasad, Prof. Mojtaba Soltanalian, Prof. Piotr Gmytrasiewicz, Prof. Anh Nguyen

University of Illinois at Chicago
M.S. in Electrical and Computer Engineering — 2018

Selected Honors & Achievements

OpenAI Researcher Access Program Award	2025
The Environmental Institute's Colab Award	2025
Cohere For AI Research Grant	2025
The LaCross AI Institute Fellowship in AI Research	2025
Dean's Strategic Fund	2024
Top Reviewer for NeurIPS	2023

Adobe Data Science Research Award	2023
Spotlight presentation, NeurIPS Ro-FoMo Workshop in Foundation Models	2023
Spotlight paper, ICML	2021
AINet Fellow by DAAD	2021
Spotlight presentation, ICML workshop on Human Interpretability in Machine Learning	2020
Spotlight paper, IEEE Conference on Image Processing (ICIP)	2019

Selected Grants & Awards

OpenAI Researcher Access Program Award (US $1,000) – Sole PI	2025
The Environmental Institute's Colab Award (US $100,000) – PI	2025
Cohere For AI Research Grant (US $2,000) – Sole PI	2025
The LaCross AI Institute Fellowship in AI Research (US $100,000) – PI	2025
Dean's Strategic Fund (US $7,440) – PI	2024
Adobe Data Science Research Award (US $50,000) – co-PI	2023
Harvard Data Science Initiative Microsoft Azure Credits (US $22,224) – co-PI	2023
AI for Social Good Google Workshop (US $10,000) – co-PI	2021
2 × Research Proposal accepted by Google Cloud Platform (US $1,000) – PI	2020

Research Articles

† denotes the author I co-mentored with the PI; * indicates an equal contribution.

Articles in Peer-Reviewed Journals

60. **C. Agarwal**, O. Queen†, H. Lakkaraju, M. Zitnik: Evaluating Explainability for Graph Neural Networks, *Nature Scientific Data, 2023.*
    189+ GitHub stars
    1,135+ Harvard Dataverse downloads

59. H. Honarvar, **C. Agarwal**, S. Somani, A. Vaid, J. Lampert, T. Wanyan, V. Y. Reddy, G. N. Nadkarni, R. Miotto1, M. Zitnik, F. Wang, B. S. Glicksberg: Enhancing convolutional neural network predictions of electrocardiograms with left ventricular dysfunction using a novel sub-waveform representation, *Cardiovascular Digital Health Journal, 2022.*

58. **C. Agarwal**, S. Gupta†, M. Y. Najjar, T. E. Weaver, X. J. Zhou, D. Schonfeld, B. Prasad: Deep Learning Analyses of Brain MRI to Identify Sleepiness in Treated Obstructive Sleep Apnea: A Pilot Study, *Journal of Sleep and Vigilance (JSV), 2022.*

57. B. Prasad*, **C. Agarwal***, E. Schonfeld, D. Schonfeld, B. Mokhlesi: Deep learning applied to polysomnography to predict blood pressure in obstructive sleep apnea and obesity hypoventilation: A proof-of-concept study, *Journal of Clinical Sleep Medicine (JCSM), 2020.*

56. **C. Agarwal**, J. Klobusicky, D. Schonfeld: Convergence of backpropagation with momentum for network architectures with skip connections, *Journal of Computational Mathematics (JCM), 2019.*

55. E. Cha, Y. Veturi, **C. Agarwal**, M. Arbabshirani, S. Pendergrass: Using Adipose Measures from Electronic Health Record Imaging Based Data for Discovery, *Journal of Obesity, 2018.*

Articles in Peer-Reviewed Conference Proceedings

54. T. Menta, S. Agrawal, **C. Agarwal**: Analyzing Memorization in Large Language Models through the Lens of Model Attribution, *NAACL, 2025.*
    Oral Presentation

53. A. Java, S. Shahid, **C. Agarwal**: Towards Operationalizing Right to Data Protection, *NAACL, 2025.*

52. E Lobo†, **C. Agarwal**, H Lakkaraju: On the Impact of Fine-Tuning on Chain-of-Thought Reasoning, *NAACL, 2025.*

51. T. Han, A. Kumar, **C. Agarwal**, H. Lakkaraju: Towards Safe and Aligned Large Language Models for Medicine, *NeurIPS Dataset and Benchmark Track, 2024.*
ICML Next Generation of AI Safety Workshop, 2024

50. A. Kumar, **C. Agarwal**, S. Srinivas, S. Feizi, H. Lakkaraju: Certifying LLM Safety against Adversarial Prompting, *COLM, 2024.*
Featured in Science News Magazine and D^3 Institute at Harvard

49. S. Krishna†, **C. Agarwal**, H. Lakkaraju: On the Impact of Adversarially Robust Models on Algorithmic Recourse, *AIES, 2024.*

48. S. Krishna†, **C. Agarwal**, H. Lakkaraju: Understanding the Effects of Iterative Prompting on Truthfulness, *International Conference on Machine Learning (ICML), 2024.*

47. S. H. Tanneru†, **C. Agarwal**, H. Lakkaraju: Uncertainty In Explanations Of Large Language Models, *International Conference on Artificial Intelligence and Statistics (AISTATS), 2024.*
Spotlight Presentation at the NeurIPS R0-FoMo Workshop, 2023

46. M. Llordes, D. Ganguly, S. Bhatia, **C. Agarwal**: Explain like I am BM25: Interpreting a Dense Model's Ranked-List with a Sparse Approximation, *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2023.*

45. A. Seth, M. Hemani, **C. Agarwal**: DeAR: Debiasing Vision-Language Models with Additive Residuals, *Conference on Computer Vision and Pattern Recognition (CVPR), 2023.*

44. S. Deshmukh†, A. Dasgupta, B. Krishnamurthy, N. Jiang, **C. Agarwal**, J. Subramanian, G. Theocharous: Trajectory-based Explainability Framework for Offline RL, *International Conference on Learning Representations (ICLR), 2023.*

43. J. Cheng†, G. Dasoulas, H. He, **C. Agarwal**, M. Zitnik: GNNDelete: A General Unlearning Strategy for Graph Neural Networks, *International Conference on Learning Representations (ICLR), 2023.*

42. V. Giunchiglia, C. V. Shukla, G, Gonzalez, **C. Agarwal**: Towards Training GNNs using Explanation Directed Message Passing, *Proceedings of the First Learning on Graphs Conference (LoG), 2022.*

41. **C. Agarwal**, E. Saxena†, S. Krishna†, M. Pawelczyk†, N. Johnson†, I. Puri†, M. Zitnik, H. Lakkaraju: OpenXAI: Towards a Transparent Evaluation of Model Explanations, *Conference on Neural Information Processing Systems (NeurIPS), 2022.*
247+ GitHub stars
15,854+ Harvard Dataverse downloads

40. **C. Agarwal**, D. D'Souza†, S. Hooker: Estimating Example Difficulty using Variance of Gradients, *Conference on Computer Vision and Pattern Recognition (CVPR), 2022.*
63+ GitHub stars

39. **C. Agarwal**, M. Zitnik, H. Lakkaraju: Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods, *International Conference on Artificial Intelligence and Statistics (AISTATS), 2022.*

38. M. Pawelczyk†, **C. Agarwal**, S. Joshi, S. Upadhyay, H. Lakkaraju: Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis, *International Conference on Artificial Intelligence and Statistics (AISTATS), 2022.*

37. **C. Agarwal**, H. Lakkaraju, M. Zitnik: Towards a Unified Framework for Fair and Stable Graph Representation Learning, *Conference on Uncertainty in Artificial Intelligence (UAI), 2021.*

36. S. Agarwal, S. Jabbari, **C. Agarwal**, S. Upadhyay, Z. S. Wu, H. Lakkaraju: Towards the Unification and Robustness of Perturbation and Gradient Based Explanations, *International Conference on Machine Learning (ICML), 2021.*
Spotlight Presentation

35. **C. Agarwal***, S. Khobahi*, D. Schonfeld, M. Soltanalian: CoroNet: A Deep Network Architecture for Semi-Supervised Task-Based Identification of COVID-19 from Chest X-ray Images, *SPIE Medical Imaging, 2021.*

34. **C. Agarwal**, A. Nguyen: Explaining image classifiers by removing input features using generative models, *Asian Conference on Computer Vision (ACCV), 2020.*

33. N. Bansal*, **C. Agarwal***, A. Nguyen*: SAM: The Sensitivity of Interpretability Methods to Hyperparameters, *Conference on Computer Vision and Pattern Recognition (CVPR), 2020.*
Oral presentation (Top 5%)

32. **C. Agarwal**, S. Khobahi, A. Bose, M. Soltanalian, D. Schonfeld: Deep-URL: A Model-Aware Approach To Blind Deconvolution Based On Deep Unfolded Richardson-Lucy Network, *IEEE Conference on Image Processing (ICIP), 2020.*

31. **C. Agarwal**, A. Nguyen, D. Schonfeld: Improving Robustness to Adversarial Examples by Encouraging Discriminative Features, *IEEE Conference on Image Processing (ICIP), 2019.*
Spotlight presentation (Top 10%)

30. M. Aloraini, M. Sharifzadeh, **C. Agarwal**, D. Schonfeld: Statistical Sequential Analysis for Object-based Video Forgery Detection, *Electronic Imaging, 2019.*

29. N. Khobragade*, **C. Agarwal***: Multi-class segmentation of neuronal electron microscopy images using deep learning, *SPIE Medical Imaging, 2018.*

28. **C. Agarwal**, M. Sharifzadeh, D. Schonfeld: CrossEncoders: A complex neural network compression framework, *IS&T International Symposium on Electronic Imaging, 2018.*

27. M. Sharifzadeh, **C. Agarwal**, M. Aloraini, D. Schonfeld: Convolutional neural network steganalysis's application to steganography, *IEEE Visual Communications and Image Processing (VCIP), 2017.*

26. **C. Agarwal**, A.H. Dallal, M.R. Arbabshirani, A. Patel, G. Moore: Unsupervised quantification of abdominal fat from CT images using Greedy Snakes, *SPIE Medical Imaging, 2017.*

25. A.H. Dallal, **C. Agarwal**, M.R. Arbabshirani, A. Patel, G. Moore: Automatic estimation of heart boundaries and cardiothoracic ratio from chest X-ray images, *SPIE Medical Imaging, 2017.*

24. M.R. Arbabshirani, A.H. Dallal, **C. Agarwal**, A. Patel, G. Moore: Accurate segmentation of lung fields on chest radiographs using deep convolutional networks, *SPIE Medical Imaging, 2017.*

23. **C. Agarwal**, A. Bose, S. Maiti, N. Islam, S.K. Sarkar: Enhanced data hiding method using DWT based on Saliency model, *IEEE International Conference on Signal Processing, Computing and Control (ISPCC), 2013.*

22. S. Maiti, **C. Agarwal**, A. Bose, S.K. Sarkar: Robust data hiding technique in wavelet domain using saliency map, *International Journal of Advances in Engineering and Technology, 2013.*

21. N. Islam S. Maiti, A. Bose, **C. Agarwal**, S. K. Sarkar: An Improved Method of Pre-Filter Based Image Watermarking in DWT Domain, *International Journal of Computer Science and Technology, 2013.*

Preprints and Workshop Articles

20. **C. Agarwal**: Rethinking Explainability in the Era of Multimodal AI, arXiv, 2025

19. A Ghosh, D Datta, S Saha, **C. Agarwal**: The Multilingual Mind: A Survey of Multilingual Reasoning in Language Models, *arXiv, 2025.*

18. A Seth, D Manocha, **C. Agarwal**: Towards a Systematic Evaluation of Hallucinations in Large-Vision Language Models, *arXiv, 2025.*

17. D. Ley†, S. H. Tanneru†, **C. Agarwal**, H. Lakkaraju: On the Difficulty of Faithful Chain-of-Thought Reasoning in Large Language Models, *ICML TiFA Workshop, 2024.*
Featured in OpenAI o1 System Card Report

16. **C. Agarwal**, S. H. Tanneru, H. Lakkaraju: Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models, *arXiv, 2024.*
Featured in OpenAI o1 System Card Report

15. N. Kroeger†, D. Ley†, S. Krishna†, **C. Agarwal**, H. Lakkaraju: Are Large Language Models Post Hoc Explainers?, *Preliminary version presented at the NeurIPS XAIA Workshop, 2023.*

14. A. Java, S. Jandial, **C. Agarwal**: Towards Fair Knowledge Distillation using Student Feedback, *Preliminary version presented at the Efficient Systems for Foundation Models, ICML 2023.*

13. S.V. Deshmukh, Srivatsan R, S. Vijay, J. Subramanian, **C. Agarwal**: Counterfactual Explanation Policies in RL, *Preliminary version presented at "Could it have been different?" Counterfactuals in Minds and Machines Workshop, ICML 2023.*

12. T. R. Menta†, S. Jandial†, A. Patil, Vimal KB, S. Bachu, B. Krishnamurthy, V. N. Balasubramanian, **C. Agarwal**, M. Sarkar: Towards Estimating Transferability using Hard Subsets, *arXiv, 2023.*

11. **C. Agarwal**: Intriguing Properties of Visual-Language Model Explanations, *Preliminary version presented at RTML Workshop, ICLR 2023.*

10. S. Krishna†, **C. Agarwal**, H. Lakkaraju: On the Impact of Adversarially Robust Models on Algorithmic Recourse, *Preliminary version presented at Trustworthy and Socially Responsible ML Workshop, NeurIPS 2022.*

9. **C. Agarwal**, N. Johnson†, M. Pawelczyk†, S. Krishna†, E. Saxena†, M. Zitnik, H. Lakkaraju: Rethinking Stability for Attribution-based Explanations, *Preliminary version presented at PAIR^{2}Struct Workshop, ICLR, 2022.* **Oral Presentation**

8. D. D'Souza†, Z. Nussbaum†, **C. Agarwal**, S. Hooker: A Tale Of Two Long Tails, *Preliminary version presented at Uncertainty & Robustness in Deep Learning Workshop, ICML, 2021.*

7. H. Honarvar, **C. Agarwal**, S. Somani, A. Vaid, J. Lampert, T. Wanyan, V. Y. Reddy, G. N. Nadkarni, R. Miotto1, M. Zitnik, F. Wang, B. S. Glicksberg: A novel representation of electrocardiogram waveforms for enhancing deep learning predictions, *Preliminary version presented at Interpretable Machine Learning in Healthcare Workshop, ICML, 2021.*

6. **C. Agarwal***, P. Chen*, A. Nguyen: Intriguing generalization and simplicity of adversarially trained neural networks, *Preliminary version presented at Human Interpretability in Machine Learning Workshop, ICML, 2020.* **Spotlight Presentation**

5. **C. Agarwal**, B. Dong, D. Schonfeld, A. Hoogs: An explainable adversarial robustness metric for deep learning neural networks, 2018.

4. M. Sharifzadeh, **C. Agarwal**, M. Salarian, D. Schonfeld: A new parallel message-distribution technique for cost-based steganography, 2017.

Patents

3. T. Menta, A. Patil, S. Jandial, Balaji K, **C. Agarwal**, M. Sarkar: Systems and methods for machine learning transferability. Application number: 18178225, 2024

2. M. Hemani, A. Seth, **C. Agarwal**: Debiasing vision-language models with additive residuals. Application number: 18322253, 2024

1. S. Deshmukh, A. Dasgupta, **C. Agarwal**, B. Krishnamurthy, G. Theocharous, J. Subramanian.: Novel Trajectory-based Explainability Framework for RL-based Decision Making. Internal Reference: P11853-US.

Teaching Experience

Instructor, Decoding Large Language Models Spring 2025
School of Data Science, UVA

Guest Lecture on Explainable Artificial Intelligence Fall 2024, Spring 2025
Course on Foundation of Data Science, UVA

Guest Lecture on Explainable Artificial Intelligence Spring 2021, 2023
Course on Interpretability and Explainability in ML, Harvard University

Teaching Assistant
University of Illinois at Chicago Spring, Fall 2014 - 2020
Pattern Recognition, Image Analysis & Computer Vision,
Digital Signal Processing, Neural Networks.

Tutorials

Explainability in Graph Deep Learning for Biomedicine ISMB 2024

Training the Next-Generation of AI Students Excellence School 2023

Explainable ML in the Wild: When Not to Trust Your Explanations FAccT 2021

Workshop

Generative AI meets Explainable AI	XAI 2025
Workshop on Regulatable ML	NeurIPS 2023-2024

Invited Talks

Social Contagions, Artificial Intelligence, and Democracy Workshop	2025
FAR.AI	2024
UVA Conference on Leadership in Business, Data and Intelligence – LaCross AI Institute	2024
Workshop on Privacy and Interpretability in Generative AI: Peering into the Black Box	2024
Computer Vision Talks	2023
TrustML Young Scientists Seminars at RIKEN-AIP, Japan	2022
Adobe Research: XAI: Challenges and Solutions	2022
CAI Summer School at IIIT-Delhi	2022
LOGML Summer School	2022
2d3d.ai	2021
W&B - Weights & Biases Salon	2020

Mentorship

Current Advisee

Eileanor LaRocco, Ph.D. Student, UVA	2025-Present
Eric Onyame, Ph.D. Student, UVA	2025-Present
Kefan Song, Ph.D. Student, UVA	2025-Present
Madelyn Mathai, Ph.D. Student, UVA	2025-Present
Runtao Zhou, Ph.D. Student, UVA	2025-Present
Akash Ghosh, Ph.D. Student, IIT Patna	2024-Present
Ashish Seth, Ph.D. Student, UMD	2024-Present
Elita Lobo, Ph.D. Student, UMass Amherst	2023-Present

Past Advisee and Interns

Susmit Agrawal, MS Student, IIT-H → PhD Student @ Tubingen (IMPRS-IS)	2024-2025
Tarun R Menta, Research Engineer @ Adobe → DataLab	2024-2025
Abhinav Java, Research Engineer @ Adobe → Microsoft	2023-2025
Simra Sahid, Research Engineer @ Adobe → Microsoft	2024-2025
Dan Ley, Ph.D. Student, Harvard University	2023-2024
Nicholas Kroeger, Ph.D. Student, University of Florida	2023-2024
Sree Harsha Tanneru, MS Student @ Harvard → Research Engineer, DeepMind	2023-2024
Satyapriya Krishna, Ph.D. Student, Harvard University → Amazon	2020-2024
Martin Pawelczyk, Ph.D. Student, University of Tübingen	2021-2022
Valentina Giunchiglia, Ph.D. Student, Imperial College London	2022-2023
Chirag Varun Shukla, Ph.D. Student, LMU Munich	2022-2023
Jiali Cheng, Ph.D. Student, University of Massachusetts Lowell	2022-2023
Surgan Jandial, Research Engineer @ Adobe → MS Student @ CMU	2022-2023
Shripad V Deshmukh, Research Engineer @ Adobe → Ph.D. student at UMass Amherst	2022-2023
Nari Johnson, Undergrad @ Harvard University → Ph.D. student @ CMU	2022
Eshika Saxena, Undergrad, Harvard University → Meta	2022
Isha Puri, Undergrad, Harvard University → Ph.D. student @ MIT	2022
Owen Queen, Undergrad, University of Tennessee, Knoxville → Ph.D. student @ Stanford	2021-2022
Daniel D'souza, Data Scientist, Proquest → Cohere	2021-2022

Community Service

Founder: Agyeya Artificial IQ Foundation	2023-Present
Open Collaboration Initiatives: TrustworthyML Initiative and MLCollective	2021-2023

External Ph.D. Examiner:

Sichao Li - Australian National University	2025
Jessica Rumbelow - University of St. Andrews	2023

Reviewer for Grants and Proposals:

Deutsche Forschungsgemeinschaft (DFG) German Research Foundation	2025

Program Committee for Workshops:

ACL-SRW - Student Research Workshop (SRW)	ACL, 2025
RegML - Regulatable Machine Learning (RegML)	NeurIPS, 2023-2024
WHI - Workshop for Women in Machine Learning (WiML),	NeurIPS, 2024
XAI4CV - Explainable AI for Computer Vision (XAI4CV) Workshop	CVPR, 2023
SRML - Workshop on Socially Responsible Machine Learning	ICLR, 2022
AdvML - New Frontiers in Adversarial Machine Learning	2022, 2024
SRML - Workshop on Socially Responsible Machine Learning	ICML,2021
SeSML - Workshop on Security and Safety in Machine Learning Systems	ICLR, 2021
AROW - Workshop on Adversarial Robustness in the Real World	ECCV, 2020-2021
WHI - Workshop on Human Interpretability in Machine Learning	ICML, 2020

Program Committee for Conferences:

NeurIPS - Advances in Neural Information Processing Systems	2021-2025
NeurIPS - Datasets and Benchmark Track	2022-2025
KDD - ACM SIGKDD Conference on Knowledge Discovery and Data Mining	2021-2023
ICML - International Conference on Machine Learning	2021-2025
FAccT - ACM Conference on Fairness, Accountability, and Transparency	2022-2023
ICLR - International Conference on Learning Representations	2022-2025
AAAI - AAAI International Conference on Artificial Intelligence	2023-2025
AIES - AAAI Conference on AI, Ethics, and Society	2024-2025
XAI World Conference	2024
AISTATS - International Conference on Artificial Intelligence and Statistics	2023
WACV - IEEE/CVF Winter Conference on Applications of Computer Vision	2023
CVPR - IEEE/CVF Conference on Computer Vision and Pattern Recognition	2023
ICCV - IEEE/CVF International Conference on Computer Vision	2023
ACL - ACL Rolling Review	2023
LOG - Learning on Graphs Conference	2022

Journal Reviewing:

Nature Communications	2024
TMLR - The Transactions on Machine Learning Research	2022-2024
TMI - IEEE Transactions on Medical Imaging	2022