ANAY MEHROTRA

anaymehrotra.com ⬦ anaymehrotra1@gmail.com

EDUCATION

Ph.D. in Computer Science, Yale University — Expected 2026
Advisors: Amin Karbasi and Manolis Zampetakis

M.S. and M.Phil. in Computer Science, Yale University — 2024

B.Tech. in Computer Science and Engineering, Indian Institute of Technology Kanpur (IIT Kanpur) — 2020
GPA: 9.9/10 (Top 5/850), Major GPA: 10/10

Exchange Semester, Computer Science, École Polytechnique Fédérale de Lausanne (EPFL) — Fall 2018

RESEARCH INTERNSHIPS

Student Researcher, Robust Intelligence (now *acquired* by Cisco), California, USA — November 2023 – May 2024
Designed a prompt-generator that makes leading LLMs output harmful content (results published at NeurIPS'24)
Results *featured in WIRED and TechCrunch*

Research Intern, Microsoft Research Lab, Karnataka, India — Summer 2022
Studied biases in ranking systems and designed ranking algorithms to mitigate biases (published at ACM AIES'23)

Visiting Student, Yale University, Connecticut, USA — March – Summer 2019
Researched benefits of interventions in presence of implicit biases (results published at ACM FAccT'20)

HONORS AND AWARDS

Sri Binay Kumar Sinha Award by IIT Kanpur for *industrially applicable research addressing societal problems* — 2020
Best Student Paper at the Workshop on Mechanism Design for Social Good (MD4SG) for Publication 1 — 2019
Academic Excellence Award in all eligible semesters at IIT Kanpur — 2016 – 2020
KVPY Scholar awarded by the Department of Science and Technology, India — 2015

SELECTED AWARDS FOR PROGRAMMING AND PROBLEM SOLVING

Finalist at the *ICPC World Final* – Rank 33 among 12,000+ teams — 2021
Winner of the *ICPC Asia West Continent Finals* — 2019
Winner of the ICPC Gwalior-Pune Regional Contest — 2019
Winner of the ICPC Kharagpur Regional Contest — 2019

PREPRINTS
In my field, authors are listed alphabetically; in exceptions, equal contributors are marked with "★"

23. *DDPM Score Matching and Distribution Learning*
    Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, Omar Montasser — 2025
    ▷ *Earlier version received an* **Oral Presentation** *at the ICLR 2025 Workshop on Deep Generative Model in ML*

22. *What Makes Treatment Effects Identifiable? Characterizations and Estimators Beyond Unconfoundedness*
    Yang Cai, Alkis Kalavasis, Katerina Mamali, Anay Mehrotra, Manolis Zampetakis — 2025

21. *Can SGD Select Good Fishermen? Local Convergence Under Self-Selection Biases and Beyond*
    Alkis Kalavasis, Anay Mehrotra, Felix Zhou — 2025

20. *Learning with Positive and Imperfect Unlabeled Data*
    Jane H. Lee, Anay Mehrotra, Manolis Zampetakis — 2025

19. *Characterizations of Language Generation with Breadth: The Interplay Between Hallucinations, Coverage, and Stability*
    Alkis Kalavasis, Anay Mehrotra, Grigoris Velegkas   2024

ALL PUBLICATIONS   *In my field, authors are listed alphabetically; in exceptions, equal contributors are marked with "★"*

18. *On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse*
    Alkis Kalavasis, Anay Mehrotra, Grigoris Velegkas
    ACM Symposium on Theory of Computing   STOC 2025

17. *Tree of Attacks: Jailbreaking Black-Box LLMs Automatically*
    Anay Mehrotra (★), Manolis Zampetakis, Paul K., B. Nelson, Hyrum Anderson, Yaron Singer, Amin Karbasi
    Conference on Advances in Neural Information Processing Systems   NeurIPS 2024
      ▷ *Also presented at the Workshop on the Next Generation of AI Safety at ICML 2024*
      ▷ *Featured in WIRED and TechCrunch*

16. *Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians*
    Jane H. Lee, Anay Mehrotra, Manolis Zampetakis
    IEEE Symposium on Foundations of Computer Science   FOCS 2024

15. *Smaller Confidence Intervals From IPW Estimators via Data-Dependent Coarsening*
    Alkis Kalavasis, Anay Mehrotra, Manolis Zampetakis
    Conference on Learning Theory   COLT 2024

14. *Fair Classification with Partial Feedback: An Exploration-Based Data-Collection Approach*
    Vijay Keswani (★), Anay Mehrotra (★), L. Elisa Celis
    International Conference on Machine Learning   ICML 2024
      ▷ *Also presented at ACM Conference on Equity and Access in Algorithms, Mechanisms, & Optimization'24*

13. *Bias in Evaluation Processes: An Optimization-Based Model*
    L. Elisa Celis, Amit Kumar, Anay Mehrotra, Nisheeth K. Vishnoi
    Conference on Advances in Neural Information Processing Systems   NeurIPS 2023

12. *Sampling Individually-Fair Rankings that are Always Group Fair*
    Sruthi Gorantla (★), Anay Mehrotra (★), Amit Deshpande, Anand Louis
    ACM Conference on Artificial Intelligence, Ethics, and Society   AIES 2023

11. *Subset Selection Based On Multiple Rankings in the Presence of Bias:*
    Effectiveness of Fairness Constraints for Multiwinner Voting Score Functions
    Niclas Boehmer, L. Elisa Celis, Lingxiao Huang, Anay Mehrotra, Nisheeth K. Vishnoi
    International Conference on Machine Learning   ICML 2023

10. *Maximizing Submodular Functions for Recommendation in the Presence of Biases*
    Anay Mehrotra and Nisheeth K. Vishnoi
    ACM Web Conference (formerly ACM WWW Conference)   WWW 2023

9. *Fair Ranking with Noisy Protected Attributes*
    Anay Mehrotra and Nisheeth K. Vishnoi
    36th Conference on Advances in Neural Information Processing Systems   NeurIPS 2022

8. *Revisiting Group Fairness Metrics: The Effect of Networks*
    Anay Mehrotra (★), Jeffrey Sachs (★), L. Elisa Celis
    ACM Conference on Computer-Supported Cooperative Work & Social Computing   CSCW 2022

7. *Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints*
   Anay Mehrotra, Bary S. R. Pradelski, Nisheeth K. Vishnoi
   ACM Conference on Fairness, Accountability, and Transparency   FAccT 2022

6. *Fairness for AUC via Feature Augmentation*
   Hortense Fong, Vineet Kumar, Anay Mehrotra, Nisheeth K. Vishnoi
   ACM Conference on Fairness, Accountability, and Transparency   FAccT 2022

5. *Fair Classification with Adversarial Perturbations*
   L. Elisa Celis, Anay Mehrotra, Nisheeth K. Vishnoi
   Conference on Advances in Neural Information Processing Systems   NeurIPS 2021

4. *Mitigating Bias in Set Selection with Noisy Protected Attributes*
   Anay Mehrotra (⋆) and L. Elisa Celis
   ACM Conference on Fairness, Accountability, and Transparency   FAccT 2021
   ▷ *Also presented at the Workshop on Co-Development of Computer Science and Law, 2020*

3. *The Effect of the Rooney Rule on Implicit Bias in the Long Term*
   L. Elisa Celis, Chris Hays, Anay Mehrotra, Nisheeth K. Vishnoi
   ACM Conference on Fairness, Accountability, and Transparency   FAccT 2021

2. *Interventions for Ranking in the Presence of Implicit Bias*
   L. Elisa Celis, Anay Mehrotra, Nisheeth K. Vishnoi
   ACM Conference on Fairness, Accountability, and Transparency   FAccT 2020

1. *Toward Controlling Discrimination in Online Ad Auctions*
   L. Elisa Celis, Anay Mehrotra, Nisheeth K. Vishnoi
   International Conference on Machine Learning   ICML 2019
   ▷ *Oral presentation* and *Best Student Paper* at *Mechanism Design for Social Good (MD4SG), 2019*
   ▷ *Accepted at the Stony Brook International Conference on Game Theory (GT Center), 2019*

TEACHING AND MENTORING EXPERIENCE

Lecturer at the Yale ICPC Club	2025 – Present
Problem setter at the Yale ICPC Club	2024 – Present

Research Mentor, Yale Univeristy
▷ Ziyu Zhu (→ IMC)   *"On Learning in Presence of Friction"* (Senior Bachelors Thesis Project)   Spring 2024

Teaching Fellow, Computer Science Department, Yale University
▷ For the undergraduate *algorithms course* (CPSC 365 & ECON 365)   Sep – Dec 2023
▷ For the graduate-level course *"Algorithms via Continuous Optimization"* (CPSC 368 & 516)   Jan – May 2023
▷ For the graduate-level course *"Algorithms via Continuous Optimization"* (CPSC 463 & 563)   Sep – Dec 2021

Academic Mentor, Counseling Service, IIT Kanpur   Jun 2017 – Jun 2018
▷ Conducted institute-wide remedial lectures for the *introductory course on electromagnetism*

RECENT TALKS

"Learning with Positive and Imperfect Unlabeled Data" at MIT A&C Seminar	Cambridge, April 2025
"On the Limits of Language Generation:..." at Invited Talk at Nanjing University	Online, April 2025
"On the Limits of Language Generation:..." at CU Theory Student Seminar	New York City, March 2025
"Learning with Positive and Imperfect Unlabeled Data" at Yale Theory Student	New Haven, March 2025
"Can SGD Select Good Fishermen? Local Convergence..." at Yale Theory Student Seminar	New Haven, 2024

"Efficient Statistics With Unknown Truncation, ..." at IEEE FOCS   Chicago, 2024

"Efficient Statistics With Unknown Truncation, ..." in Bangalore Theory Seminars at IISc   Bengaluru, 2024

"Tree of Attacks: Jailbreaking Black-Box LLMs Automatically" in Lunch Seminar at Robust Intelligence   Online, 2024

"Efficient Statistics With Unknown Truncation, ..." at Yale Theory Student Seminar   New Haven, 2024

"Selection with Implicit Bias: Evaluating the Efficacy of..." at Yale Theory Student Seminar   New Haven, 2023

"Maximizing Submodular Functions for Recommendation in the..." at the ACM Web Conference   Austin, 2023

"Fairness for AUC via Feature Augmentation" at ACM FAccT   Seoul, 2022

"Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints" at ACM FAccT   Seoul, 2022

"Fair Classification with Adversarial Perturbations" at NeurIPS   Online, 2021

"The Effect of the Rooney Rule on Implicit Bias in the Long Term" at ACM FAccT   Online, 2021

"Mitigating Bias in Set Selection with Noisy Protected Attributes" at ACM FAccT   Online, 2021

"Interventions for Ranking in the Presence of Implicit Bias" at ACM FAccT   Barcelona, 2020

"Applications of Communication Complexity in Lower Bounds for Streaming Algorithms" at SIGTACS   Kanpur, 2020

"Toward Controlling Discrimination in Online Ad..." at International Conf. on Game Theory   Stony Brook, 2019

"Toward Controlling Discrimination in Online Ad..." at Mechanism Design for Social Good   Phoenix, 2019

INDUSTRY INTERNSHIPS

Quantitative Researcher Intern, Two Sigma, New York, USA   Summer 2024

Algorithm-Developer Intern, Hudson River Trading, New York, USA   Summer 2023

SERVICE

Program committee member at the following conferences
COLT (2025), ACM FAccT (2025), ACM FAccT (2024), ACM FAccT (2023), ACM FAccT (2022)

Reviewer or sub-reviewer for the following conferences and journals
JAIR (2025), ICML (2025), AISTATS (2025), STACS (2025), NeurIPS (2024), EAAMO (2024), FOCS (2024), ITCS (2024), NeurIPS (2023), ICML (2022), NeurIPS (2021)

Volunteer at the following events and societies
▷ The Symposium on the Foundations of Responsible Computing (FORC)   2022
▷ The First Yale Workshop on AI, Ethics, and Society   2019
▷ Student guide at the Counseling Service, IIT Kanpur   2017

President of the Robotics Club, Delhi Public School, RK Puram   2015

EXTRACURRICULAR ACTIVITY

Running   Completed in the Lausanne half-marathon   2018
Fine Arts   Showcased paintings at the All India Fine Art and Craft Society, Delhi   2016
Fine Arts   Awarded a diploma in Fine Arts from Pracheen Kala Kendra   2015