

Natalie Collina

Cambridge, MA • people.csail.mit.edu/ncollina • nataliecollina@gmail.com

I study AI safety through the lens of learning and strategic interaction. My work develops game-theoretic and learning-theoretic frameworks for understanding when interacting AI and human agents are safe and aligned, and when strategic dynamics instead produce manipulation, collusion, or other failures. My work has been honored with a Joint Best Paper/Best Student Paper Award at the ACM Conference for Economics and Computation and has been featured in Quanta and Wired Magazines.

Education

MAPS/SERC Postdoctoral Fellow, MIT 2026 -

Supported by the Multi-agent AI Postdoctoral Fellowship and the Social and Ethical Responsibilities of Computing Fellowship

University of Pennsylvania, Ph.D. in Computer and Information Sciences 2021 - 2026

Advised by Aaron Roth and Michael Kearns. Supported by the IBM Research Fellowship.

Princeton University, B.S.E. 2019 Summa Cum Laude

Awarded the Outstanding Computer Science Senior Thesis Prize.

In Submission (*denotes author-contribution order; authorship is alphabetical otherwise)

Non-manipulability and Approachability Beyond Two-Player Games

Natalie Collina, Elijah Fullerton, Jingyi Liu, Matt Weinberg

[Delegating Authorization to Misaligned Agents: Coalitional Alignment and Safe Control](#)

Natalie Collina, Surbhi Goel, Aaron Roth, Sikata Sengupta

[Personalization Aids Pluralistic Alignment via Competition](#)

Natalie Collina, Surbhi Goel, Aaron Roth, Mirah Shi

Best Paper Award at ICLR AIMS Workshop

Failure Modes in AI Retraining Dynamics

Natalie Collina*, Kiarash Banihashem, Nicole Immorlica, Brendan Lucier, Aleksandrs Slivkins

Spotlight presentation at ICML NExT-Game Workshop

[Breaking Algorithmic Collusion in Human-AI Ecosystems](#)

Natalie Collina*, Eshwar Ram Arunachaleswaran, and Meena Jagadeesan

Accepted to the Algorithmic Collective Action Workshop, NeurIPS 2025

Accepted to 2026 ESIF Economics and AI+ML Meeting

[The Value of Ambiguous Commitments in Multi-Follower Games](#)

Natalie Collina, Rabanus Derr, and Aaron Roth

R&R, Games and Economic Behavior

Conference Publications

[The Sample Complexity of Multicalibration](#)

Natalie Collina, Jiuyao Lu, Georgy Noarov, Aaron Roth

SODA 2027

[Narrowing the Collaboration Gap, Probably](#)

Mirah Shi*, Marcel Hussing, Ira Globus-Harris, Natalie Collina, Aaron Roth, and Surbhi Goel

NeurIPS 2026

[Optimal Lower Bounds for Online Multicalibration](#)

Natalie Collina, Jiuyao Lu, Georgy Noarov, Aaron Roth

FOCS 2026

[Bandit Social Learning with Exploration Episodes](#)

Kiarash Banihashem, Natalie Collina, Aleksandrs Slivkins

ICML 2026

Emergent Alignment via Competition	ICML 2026
Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi <i>Accepted to the Algorithmic Collective Action Workshop, NeurIPS 2025</i> <i>Accepted to 2026 ESIF Economics and AI+ML Meeting</i>	
Collaborative Prediction: Tractable Information Aggregation via Agreement	SODA 2026
Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi <i>Spotlight presentation at the EC 2025 Workshop on Human-AI Collaboration</i> <i>Spotlight presentation at the 2025 TTIC Workshop on Incentives for Collaborative Learning and Data Sharing</i>	
Swap Regret and Correlated Equilibria Beyond Normal-Form Games	EC 2025
Eshwar Ram Arunachaleswaran, Natalie Collina, Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan <i>Best Paper and Best Paper with a Student Lead Author</i> <i>Invited to Highlights of Algorithms 2026</i>	
Learning to Play Against Unknown Opponents	EC 2025
Eshwar Ram Arunachaleswaran, Natalie Collina, and Jon Schneider	
Tractable Agreement Protocols	STOC 2025
Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth <i>Also accepted to the Pluralistic Alignment Workshop, NeurIPS 2024</i>	
Algorithmic Collusion Without Threats	ITCS 2025
Eshwar Ram Arunachaleswaran, Natalie Collina, Sampath Kannan, Aaron Roth, and Juba Ziani <i>Also accepted to CSLAW 2025, non-archival track</i> <i>Featured in Quanta and Wired</i>	
An Elementary Predictor Obtaining $2\sqrt{T}$ Distance to Calibration	SODA 2025
Eshwar Ram Arunachaleswaran, Natalie Collina, Aaron Roth, and Mirah Shi <i>Also accepted to the OPT24 Workshop, NeurIPS 2024</i>	
Repeated Contracting with Multiple Non-Myopic Agents: Policy Regret and Limited Liability	EC 2024
Natalie Collina, Varun Gupta, and Aaron Roth. <i>Also accepted to the 2024 ESIF Economics and AI+ML Meeting.</i>	
Pareto-Optimal Algorithms for Learning in Repeated Games	EC 2024
Eshwar Ram Arunachaleswaran, Natalie Collina, and Jon Schneider. <i>Also accepted to the 2024 ESIF Economics and AI+ML Meeting.</i>	
Efficient Prior-Free Mechanisms for No-Regret Agents	EC 2024
Natalie Collina, Aaron Roth, and Han Shao.	
Efficient Stackelberg Strategies for Finitely Repeated Games	AAMAS 2023
Natalie Collina*, Eshwar Ram Arunachaleswaran, and Michael Kearns.	
Dynamic Weighted Matching with Heterogenous Arrival and Departure Rates	WINE 2020
Natalie Collina, Nicole Immorlica, Kevin Leyton-Brown, Brendan Lucier, and Neil Newman.	
On the (in)-approximability of Bayesian Mechanism Design for a Combinatorial Buyer	EC 2020
Natalie Collina and Matt Weinberg.	

Journal Publications

[How To Find Fantastic AI Papers: Self-Rankings as a Predictor of Scientific Impact Beyond Peer-Review](#)

Buxin Su*, Natalie Collina, Garrett Wen, Didong Li, Kyunghyun Cho, Jianqing Fan, Bingxin Zhao, Weijie Su
Nature Computational Science

[The ICML 2023 Ranking Experiment: Examining Author Self-Assessment in ML/AI Peer Review](#)

Buxin Su*, Jiayao Zhang, Natalie Collina, Yuling Yan, Didong Li, Jianqing Fan, Aaron Roth, and Weijie J. Su
Journal of the American Statistical Association, Discussion Paper
Also accepted to the Workshop on Incentives in Academia, EC 2024.

Awards & Honors

Best Paper Award, ICLR AIMS 2026
Best Paper with a Student Lead Author, EC 2025
Best Paper Award, EC 2025
IBM PhD Fellowship (2024-2025)
Rising Star in EECS, MIT (2024)
Gift from Amazon AWS in Trustworthy AI (2023)
NSF-GRFP Honorable Mention (2020)
Outstanding Computer Science Senior Thesis Prize (2019)
Sigma Xi Book Award for Outstanding Undergraduate Research (2019)

Work Experience

Microsoft Research Intern, Economics and Computation Group Summer 2025
Google Software Engineer Fall 2019 - Spring 2021
Microsoft Research Intern, Economics and Computation Group Summer 2019
Amazon Web Services Software Development Intern Summer 2018

Selected Talks

Calibration, Decisions and Collaboration in Learning

ICML 2026 Tutorial, with Aaron Roth and Ira Globus-Harris
2026 NSF AI-SDM Human-AI Complementarity Workshop, two-hour tutorial with Ira Globus-Harris

Failure Modes in AI Retraining Dynamics

2026 ICML NExT-Game Workshop, Spotlight

Emergent Alignment via Competition

2026 ESIF Economics and AI+ML Meeting
2026 Simons Institute, Agency in Collaborative Learning Workshop
2026 Simons Collaboration on the Theory of Algorithmic Fairness
2025 UK AISI Alignment Conference
2025 Cornell Young Researcher's Workshop

Learning and Incentives in Human-AI Collaboration

2026 NYU CS Colloquium
2026 TTIC Colloquium
2025 MIT Algorithms and Complexity Seminar
2025 Northeastern Theory Seminar
2025 CMU Theory Lunch
2025 NYU Theory Seminar
2025 Rational Intelligence Seminar Series

Swap Regret and Correlated Equilibria Beyond Normal-Form Games

2025 ACM Conference on Economics and Computation (EC)
2025 Cornell Theory Seminar
2025 JHU Theory Seminar

2025 TCS+ Seminar

Repeated AI Interaction: How Agents Learn and Strategize over Time

2025 CHAI Workshop, Session on Cooperation and Coordination

Collaborative Prediction: Tractable Information Aggregation via Agreement

2025 Caltech RSRG/FALCON Seminar

2025 AWS Responsible AI Science Meeting

2025 EC Workshop on Human-AI Collaboration

Algorithmic Collusion Without Threats

2024 Innovations in Theoretical Computer Science (ITCS)

2025 Computer Science and Law Conference (CSLAW)

Tractable Agreement Protocols

2024 Northwestern/TTIC Junior Theorists Workshop

2024 JHU Theory Seminar

2024 AWS Responsible AI Science Meeting

2024 FOCS Workshop on Calibration

2024 TOC4Fairness Seminar

2025 NYC Student Theory Day

2025 Symposium on Theory of Computing (STOC)

Learning to Play Against Unknown Opponents

2024 INRIA Paris Optimization Seminar

2024 Princeton Mechanism Design Lunch

2025 ACM Conference on Economics and Computation (EC)

Pareto-Optimal Algorithms for Learning in Repeated Games

2024 Workshop on Algorithms, Learning and Economics (WALE)

2024 ACM Conference on Economics and Computation (EC)

2024 ESIF Economics and AI+ML Meeting

Repeated Contracting with Multiple Non-Myopic Agents: Policy Regret and Limited Liability

2024 EnCORE Seminar (Best Presentation Award)

2024 ACM Conference on Economics and Computation (EC)

2024 ESIF Economics and AI+ML Meeting

Efficient Prior-Free Mechanisms for No-Regret Agents

2024 ACM Conference on Economics and Computation (EC)

2024 INFORMS Meeting

Efficient Stackelberg Strategies for Finitely Repeated Games

2023 Conference on Autonomous Agents and Multi Agent Systems (AAMAS)

Dynamic Weighted Matching with Heterogeneous Arrival and Departure Rates

2020 Conference on Web, Internet and Network Economics (WINE)

On the (in)-approximability of Bayesian Mechanism Design for a Combinatorial Buyer

2020 ACM Conference on Economics and Computation (EC)

Leadership and Mentorship

Head Teaching Assistant, Advanced Topics in Machine Learning (CIS 6200) Fall 2025

Co-leader, Economics and Computation Gender Inclusion Workshop EC 2024, EC 2025

Co-founder and Leader, University of Pennsylvania Theory Seminar Fall 2022 - Fall 2025

Head Teaching Assistant, Advanced Topics in Machine Learning (CIS 6200) Fall 2024

Head Teaching Assistant, Algorithmic Game Theory (NETS 4120) Spring 2023

Teaching Assistant, New Horizons in Theoretical Computer Science Summer 2022

Head Teaching Assistant, Introduction to Algorithms (CIS 3200) Fall 2021

Service

PC Reviewer: ITCS 2026, EC 2026, WINE 2025, EC 2025, FAccT 2022

Subreviewer: SODA 2026, STOC 2025, FAccT 2024, SODA 2024, ITCS 2023, EC 2021.

Student Volunteer: EC 2025, CCC 2022.

Other Talks

A General Reduction from No-Regret to No-Swap Regret

Lecture in CIS 6200, University of Pennsylvania.