Nika Haghtalab

Associate Professor, EECS
University of California, Berkeley

Email : nika@berkeley.edu
Homepage: people.eecs.berkeley.edu/∼nika
Address: 2021 Berkeley Way West, BWW 8028
Berkeley, California

Research Interest and Expertise

I work on a broad range of **core methodological questions in machine learning and artificial intelligence**, with an emphasis on principled and often mathematical approaches. Much of this research contributes to an emerging foundations for machine learning and decision-making systems that explicitly accounts for the environments and contexts in which they operate. These include questions of **trustworthy AIML, sequential and interactive decisions-making**, principled approach to **generative modeling and LLMs**, **Human–AI interaction**, and learning with human or **agentic behavior**.

Employment

Associate Professor (with tenure), EECS, UC Berkeley	2025–
Assistant Professor, EECS, UC Berkeley	2021–2025
Assistant Professor, Computer Science, Cornell University	2019–2020
Postdoctoral Researcher, Microsoft Research – New England	2018–2019
Research Intern, Microsoft Research – New York City	06-09/2016
Research Intern, Microsoft Research – Redmond	05-08/2015

Education

Ph.D. in Computer Science, Carnegie Mellon University Advisors: Avrim Blum, Ariel Procaccia	2013-2018
M.Math in Computer Science, University of Waterloo Advisor: Shai Ben-David	2011-2013
B.Math in Comp. Sci. and Combinatorics & Optimization, University of Waterloo, Dean's Honors List with Distinction	2008-2011

Awards and Honors

Amazon Research Award	2025
Alfred P. Sloan Fellowship	2024
Schmidt Sciences AI2050 Fellowship	2024
Google Research Scholar Award	2023
Exemplary Paper Award, AI Track, EC 2023	2023
Outstanding Paper Award, NeurIPS 2022	2022
NSF CAREER Award	2022
SIGecom honorable mention Dissertation Award	2018
Best Paper Award, ICAPS 2018	2018
Carnegie Mellon University SCS Dissertation Award	2018
Carnegie Mellon University **ACM Dissertation Nomination**	2018
Rising Stars in EECS	2017
Siebel Scholarship	2017
Microsoft Research Ph.D. Fellowship	2016-2018
Facebook Ph.D. Fellowship	2016-2018
IBM Ph.D. Fellowship	2015-2016

Selected Leadership and Service

Conference Organization and Highest-Level Editorial Service

Program Co-Chair, Conference on Neural Information Processing Systems (NeurIPS) 2026
Program Co-Chair, Conference on Learning Theory (COLT) 2025
Program Co-Chair, Conference on Algorithmic Learning Theory (ALT) 2022
Editorial Member, Foundations and Trends in ML (FnT ML) 2024-

Other Major Leadership

Steering Committee member, Association for Computational Learning 2021–2024
Director, Association for Algorithmic Learning Theory 2021–2026
Director, Center for the Foundations of Learning, Inference, Information, Intelligence, Math and Microeconomics at Berkeley (CLIMB) 2021-
Co-founder, Learning Theory Alliance (LeT-All). The first mentorship program in learning theory 2020-

Regular Teaching

CS 170: Efficient Algorithms and Intractable Problems, UC Berkeley 2023-
Data 102: Data, Inference, and Decisions, UC Berkeley Sp2022
CS 272: Foundations of Learning, Decisions, and Games (New), UC Berkeley 2021-
CS4780/CS5780: Introduction to Machine Learning, Cornell University 2019-2020
CS6781: Foundations of Modern Machine Learning (New), Cornell University Sp2020

Research Advising

Ph.D. Students

Annie Ulichney, Stats co-advised with Jordan. 2025-present
Mark Bedaywi, EECS co-advised with Russell. 2024-present
Haozhe Jiang, EECS co-advised with Jiao. 2024-present
Brian Lee, EECS co-advised with Jordan. 2024-present
Jessica Dai, EECS co-advised with Recht. 2022-present
Nivasini Ananthakrishnan, EECS co-advised with Jordan. 2021-present
Kunhe Yang, EECS. 2021-present

Eric Zhao, EECS co-advised with Jordan. *Now researcher at OpenAI.* 2021-2025
Abhishek Shetty, EECS. *Incoming faculty at GeorgiaTech CS* 2019-2024
Alex Wei, EECS co-advised with Jordan and Steinhardt. *Now researcher at OpenAI* 2021-2023

Postdoctoral Researchers

Keegan Harris, EECS. 2026-present
Yuval Dagan, FODSI. *Now faculty at Tel Aviv University* 2023-2024
Mingda Qiao, FODSI. *Incoming faculty UMass Amherst* 2023-2024
Paul Gölz, MSRI/FODSI. *Now faculty at Cornell ORIE* 2024
Chara Podimata, FODSI. *Now faculty at MIT Sloan* 2022-2023
Mahsa Derakhshan, FODSI. *Now faculty at Northeastern Univ.* 2021-2022

Undergraduate Research Advising

Naveen Durvasula, BEng/BBA 2023. *Now Ph.D student at Columbia U.* 2021-2023
Korinna Frangias, BEng 2023. *Incoming Ph.D student at CMU.* 2022-2023

Publications

The authors are ordered alphabetically in most of my publications. Exceptions are marked by (∗), in which typically the student authors are listed first and the faculty with the more significant contributions are listed last. In this case, (⋄) represent equal contribution.

Book Chapters

BC2 M.F Balcan, and **N. Haghtalab**. Noise in Classification. In T. Roughgarden, editors, Beyond the Worst-Case Analysis of Algorithms, chapter 16, Cambridge University Press, United Kingdom, 2020. [CUP version]

BC1 A. Blum, **N. Haghtalab**, and A.D. Procaccia. Learning to Play Stackelberg Security Games. In Ali E. Abbas, Milind Tambe, and Detlof von Winterfeldt, editors, Improving Homeland Security Decisions, chapter 25. Cambridge University Press, Cambridge, United Kingdom, 2017. [CUP version]

Journal Articles

J7 **N. Haghtalab**, T. Roughgarden, A. Shetty. Smoothed Analysis with Adaptive Adversaries. *Journal of the ACM, 71.3:1-34, 2024.* [JACM version]

J6 **N. Haghtalab**, M.O. Jackson, A.D. Procaccia. Belief Polarization in a Complex World: A Learning Theory Perspective. *Proc. of the National Academy of Science, 118 (19) e2010144118,* 2021. [PNAS version]

J5 A. Torrico, M. Singh, S. Pokutta, S. Naor, **N. Haghtalab**, N. Anari. Structured Robust Submodular Maximization: Offline and Online. *INFORMS Journal on Computing, 33(4):1590–1607,* 2021. [INFORMS version]

J4 M. Dudík, **N. Haghtalab**, H. Luo, R.E. Schapire, V. Syrgkanis, and J. Wortman Vaughan. Oracle-efficient learning and auction design. *Journal of the ACM 67(5):1–57,* 2020. [JACM version]

J3 M.F. Balcan, **N. Haghtalab**, and C. White. k-center Clustering under Perturbation Resilience. *ACM Transactions on Algorithms, 16(2):1–30,* 2020. [TALG version]

J2 A. Blum, J.P. Dickerson, **N. Haghtalab**, A.D. Procaccia, T. Sandholm, and A. Sharma. Ignorance is almost bliss: Near-optimal stochastic matching with few queries. *Operations Research, 68(1):16–34,* 2020. [INFORMS version]

J1 **N. Haghtalab**, A. Laszka, A.D. Procaccia, Y. Vorobeychik, and Xenofon Koutsoukos. Monitoring stealthy diffusion. *Knowledge and Information Systems, 52(3):1–29,* 2017. [KAIS version]

Peer-Reviewed Conference Proceedings

C61 M. Bedaywi, S. Emmons, **N. Haghtalab**, S. Russell. Short and Sweet: Computationally Efficient Collaborative Communication Through Indistinguishability and Coarsening. Forthcoming in *Proc. 67th Annual IEEE Symposium on Foundations of Computer Science, **(FOCS)***, November 2026.

C60 B. Lee, **N. Haghtalab**$^{\diamond}$, M.I. Jordan$^{\diamond}$, R. Tibshirani$^{\diamond}$ (*). Revisiting Gradient Equilibrium. In *Proc. Conference on Learning Theory, **(COLT)***, 2026.

C59 I. Aden-Ali, N. Golowich, A. Liu, A. Shetty, A. Moitra, **N. Haghtalab**(*). Subliminal Effects in Your Data: A General Mechanism via Log-Linearity. In *Proc. International Conference on Machine Learning, **(ICML)***, 2026. [arXiv version]

C58 K. Oko, A. Ulichney, **N. Haghtalab**, H. Bao (*). Distortion of AI Alignment Revisited: RLHF is a Decent Utilitarian Aligner. In *Proc. International Conference on Machine Learning, **(ICML)***, 2026.

C57 J. Dai, S. Garcia, E. Pierson, B. Recht, **N. Haghtalab**(*). Societal Impact Evaluations from Social Media Data. In *Proc. International Conference on Machine Learning, **(ICML)***, 2026.

C56 **N. Haghtalab**, A. D. Procaccia, H. Shao, S. L. Wang, K. Yang. Pluralistic Leaderboards. In *Proc. International Conference on Machine Learning, **(ICML)***, 2026 .

C55 H. Jiang, **N. Haghtalab**, L. Chen (*). Diffusion Language Models are Provably Optimal Parallel Samplers. In *Proc. International Conference on Learning Representations, **(ICLR)***, 2026. [arXiv version]

C54 S. Balakrishnan, **N. Haghtalab**, D. Hsu, B. Lee, E. Zhao. Panprediction: Optimal Predictions for Any Downstream Task and Loss. In *Proc. International Conference on Artificial Intelligence and Statistics, **(AISTATS)***, 2026. [arXiv version]

C53 P. Golz, **N. Haghtalab**, K. Yang. Distortion of AI Alignment: Does Preference Optimization Optimize for Preferences?. In *Proc. 38th Annual Conference on Neural Information Processing Systems, **(NeurIPS)***, 2025. [arXiv version]

C52 E. Zhao, P. Awasthi, **N. Haghtalab**. From Style to Facts: Mapping the Boundaries of Knowledge Injection with Finetuning. In *Proc. 38th Annual Conference on Neural Information Processing Systems, **(NeurIPS)***, 2025. [arXiv version]

C51 **N. Haghtalab**, O. Montasser, M. Qiao. Sample-Adaptivity Tradeoff in On-Demand Sampling. In *Proc. 38th Annual Conference on Neural Information Processing Systems, **(NeurIPS)***, 2025. [arXiv version]

C50 **N. Haghtalab**, M. Qiao, K. Yang. Leakage-Robust Bayesian Persuasion. In *Proc. ACM Conference on Economics and Computation, **(EC)***, 2025 (forthcoming). [arXiv version]

C49 J. Dai, **N. Haghtalab**, E. Zhao. Learning With Multi-Group Guarantees For Clusterable Subpopulations. In *Proc. International Conference on Machine Learning, **(ICML)***, 2025. [arXiv version]

C48 **N. Haghtalab**, M. Qiao, K. Yang. Platforms for Efficient and Incentive-Aware Collaboration. In *Proc. of the Symposium on Discrete Algorithms **(SODA)***, 2025. [arXiv version]

C47 **N. Haghtalab**, M. Qiao, K. Yang, E. Zhao. Truthfulness of Calibration Measures. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2024. [arXiv version]

C46 N. Ananthakrishnan, **N. Haghtalab**, C. Podimata, K. Yang. Is Knowledge Power? On the (Im)possibility of Learning from Strategic Interactions. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2024. [arXiv version]

C45 D. Halawi, A. Wei, E. Wallace, T. Wang, **N. Haghtalab**$^{\diamond}$, Jacob Steinhardt$^{\diamond}$. Covert Malicious Finetuning: Challenges in Safeguarding LLM Adaptation. In *Proc. International Conference on Machine Learning, **(ICML)***, 2024 (∗). [arXiv version]

C44 N. Ananthakrishnan, S. Bates, M.I. Jordan, and **N. Haghtalab**. Delegating Data Collection in Decentralized Machine Learning. In *Proc. International Conference on Artificial Intelligence and Statistics, **(AISTATS)***, 2024 (∗). [arXiv version]

C43 J. Dai, B. Flanigan, **N. Haghtalab**, M. Jagadeesan, C. Podimata. Can Probabilistic Feedback Drive User Impacts in Online Platforms? In *Proc. International Conference on Artificial Intelligence and Statistics, **(AISTATS)***, 2024. [arXiv version]

C42 C. Daskalakis, N. Golowich, **N. Haghtalab**, and A. Shetty. Smooth Nash Equilibria: Algorithms and Complexity. In *Proc. Innovations in Theoretical Computer Science, **(ITCS)***, 2024. [arXiv version]

C41 N. Haghtalab, N. Immorlica, B. Lucier, M. Mobius, and D. Mohan. Communicating with Anecdotes. In *Proc. Innovations in Theoretical Computer Science, **(ITCS)***, 2024. [arXiv version]

C40 A. Wei, **N. Haghtalab**$^{\diamond}$, J. Steinhardt$^{\diamond}$. Jailbroken: How Does LLM Safety Training Fail? In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2023. (Oral) (∗). [arXiv version]

C39 **N. Haghtalab**, C. Podimata, K. Yang. Calibrated Stackelberg Games: Learning Optimal Commitments Against Calibrated Agents. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2023. [arXiv version]

C38 A. Bhatt, **N. Haghtalab**, A. Shetty. Smoothed Analysis of Sequential Probability Assignment. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2023. [arXiv version]

C37 **N. Haghtalab**, M.I. Jordan, E. Zhao. A Unifying Perspective on Multi-Calibration: Game Dynamics for Multi-Objective Learning. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2023. [arXiv version]

C36 M. Jagadeesan, M.I. Jordan, J. Steinhardt$^{\diamond}$, **N. Haghtalab**$^{\diamond}$. Improved Bayes Risk Can Yield Reduced Social Welfare Under Competition. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2023 (∗). [arXiv version]

C35 P. Awashti, **N. Haghtalab**, and E. Zhao. Open Problem: The Sample Complexity of Multi-Distribution Learning for VC Classes. In *Proc. 36th Annual Conference on Learning Theory **(COLT)***, 2023. [arXiv version]

C34 N. Durvasula, **N. Haghtalab**, and M. Zampetakis. Smoothed Analysis of Online Non-parametric Auctions. In *Proc. 24th ACM Conference on Economics and Computation **(EC)***, 2023. [EC version]

C33 W. Guo, **N. Haghtalab**, K. Kandasamy, and E. Vitercik. Leveraging Reviews: Learning to Price with Buyer and Seller Uncertainty. In *Proc. 24th ACM Conference on Economics and Computation **(EC)***, 2023. **EXEMPLARY PAPER AWARD IN THE AI TRACK**. [arXiv version]

C32 M. Derakhshan, N. Durvasula, and **N. Haghtalab**. Stochastic Minimum Vertex Cover in General Graphs: a 3/2-Approximation. In *Proc. of the ACM Symposium on Theory of Computing **(STOC)***, 2023. [arXiv version]

C31 M. Jagadeesan, M.I. Jordan, **N. Haghtalab**. Competition, Alignment, and Equilibria in Digital Marketplaces. In *Proc. of the AAAI Conference on Artificial Intelligence **(AAAI)***, 2023 (*). [arXiv version]

C30 **N. Haghtalab**, Y. Han, A. Shetty, and K. Yang. Oracle-Efficient Online Learning for Beyond Worst-Case Adversaries. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2022. [arXiv version]

C29 **N. Haghtalab**, M.I. Jordan, and E. Zhao. On-Demand Sampling: Learning Optimally from Multiple Distributions. In *Proc. of the Conference on Neural Information Processing Systems, **(NeurIPS)***, 2022. **WINNER OF OUTSTANDING PAPER AWARD**. [arXiv version]

C28 **N. Haghtalab**, T. Lykouris, S. Nietert, A. Wei. Learning in Stackelberg Games with Non-myopic Agents. In *Proc. of the ACM Conference on Economics and Computation **(EC)***, 2022. [arXiv version]

C27 **N. Haghtalab**, T. Roughgarden, A. Shetty. Smoothed Analysis with Adaptive Adversaries. In *Proc. of the Symposium on Foundations of Computer Science **(FOCS)***, 2021. [arXiv version]

C26 A. Blum, **N. Haghtalab**, R. Phillips, H. Shao. One for One, or All for All: Equilibria and Optimality of Collaboration in Federated Learning. In *Proc. of the International Conference on Machine Learning,* ***(ICML)***, 2021. [arXiv version]

C25 **N. Haghtalab**, T. Roughgarden, A. Shetty. Smoothed Analysis of Online and Differentially Private Learning. In *Proc. of the Conference on Neural Information Processing Systems,* ***(NeurIPS)***, 2020. [arXiv version]

C24 **N. Haghtalab**, N. Immorlica, B. Lucier, J. Wang. Maximizing Welfare with Incentive-Aware Evaluation Mechanisms. In *Proc. of the International Joint Conference on Artificial Intelligence* ***(IJCAI)***, 2020. [arXiv version]

C23 L.T. Liu, A. Wilson, **N. Haghtalab**, A. Kalai, C. Borgs, and J. Chayes. The Disparate Equilibria of Algorithmic Decision Making when Individuals Invest Rationally. In *Proc. of the ACM Conference on Fairness, Accountability, and Transparency,* ***(FAT*)***, 2020(*). [arXiv version]

C22 **N. Haghtalab**, C. Musco, and B. Waggoner. Toward a Characterization of Loss Functions for Distribution Learning. In *Proc. of the Conference on Neural Information Processing Systems,* ***(NeurIPS)***, 2019. [arXiv version]

C21 A. Blum, **N. Haghtalab**, M. Hajiaghayi, and S. Seddighin. Computing Stackelberg Equilibria of Large General-Sum Games. In *Proc. of the International Symposium on Algorithmic Game Theory* ***(SAGT)***, 2019. [arXiv version]

C20 C. Borgs, J. Chayes, **N. Haghtalab**, A. Kalai, E. Vitercik. Algorithmic Greenlining: An Approach to Increase Diversity. In *Proc. of the ACM Conference on AI, Ethics, and Society* ***(AIES)***, 2019. [AIES version]

C19 N. Anari, **N. Haghtalab**, S. Naor, S. Pokutta, M. Singh, and A. Torrico. Robust Submodular Maximization: Offline and Online Algorithms. In *Proc. of the International Conference on Artificial Intelligence and Statistics* ***(AISTATS)***, 2019. [PMLR version]

C18 **N. Haghtalab**, S. Mackenzie, A.D. Procaccia, O. Salzman, S. Srinivasa. The Provable Virtue of Laziness in Motion Planning. In *Proc. of the International Conference on Automated Planning and Scheduling* ***(ICAPS)***, 2018.
WINNER OF THE BEST PAPER AWARD . [arXiv version]

C17 **N. Haghtalab**, R. Noothigattu, and A.D. Procaccia. Weighted Voting Via No-Regret Learning. In *Proc. of the AAAI Conference on Artificial Intelligence* ***(AAAI)***, 2018. [arXiv version]

C16 A. Blum and **N. Haghtalab**. Generalized topic modeling. In *Proc. of the AAAI Conference on Artificial Intelligence* ***(AAAI)***, 2018. [arXiv version]

C15 O. Dekel, A. Flajolet, **N. Haghtalab**, P. Jaillet. Online Learning with a Hint. In *Proc. of the Conference on Neural Information Processing Systems,* ***(NeurIPS)***, 2017. [NeurIPS version]

C14 A. Blum, **N. Haghtalab**, A.D. Procaccia, and M. Qiao. Collaborative PAC Learning. In *Proc. of the Conference on Neural Information Processing Systems,* ***(NeurIPS)***, 2017. [NeurIPS version]

C13 M. Dudík, **N. Haghtalab**, H. Luo, R.E. Schapire, V. Syrgkanis, and J. Wortman Vaughan. Oracle-efficient learning and auction design. In *Proc. of the Symposium on Foundations of Computer Science* ***(FOCS)***, 2017. [arXiv version]

C12 P. Awasthi, A. Blum, **N. Haghtalab**, Y. Mansour. Efficient PAC learning from the Crowd. In *Proc. of the Conference on Learning Theory* ***(COLT)***, 2017. [arXiv version]

C11 A. Blum, I. Caragiannis, **N. Haghtalab**, A.D. Procaccia, E.B. Procaccia, and R. Vaish. Opting into optimal matchings. In *Proc. of the Symposium on Discrete Algorithms* ***(SODA)***, 2017. [arXiv version]

C10 P. Awasthi, M.F. Balcan, **N. Haghtalab**, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proc. of the Conference on Learning Theory **(COLT)***, 2016. [PMLR version]

C9 M.F. Balcan, **N. Haghtalab**, and C. White. k-center clustering under perturbation resilience. In *Proc. of the International Colloquium on Automata, Languages, and Programming, **(ICALP)***, 2016. [arXiv version]

C8 **N. Haghtalab**, F. Fang, T. Nguyen, A. Sinha, A.D. Procaccia, and M. Tambe. Three strategies to success: Learning adversary models in security games. In *Proc. of the International Joint Conference on Artificial Intelligence **(IJCAI)***, 2016 (*). [IJACI version]

C7 P. Awasthi, M.F. Balcan, **N. Haghtalab**, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proc. of the Conference on Learning Theory **(COLT)***, 2015. [arXiv version]

C6 **N. Haghtalab**, A. Laszka, A.D. Procaccia, Y. Vorobeychik, and X. Koutsoukos. Monitoring stealthy diffusion. In *IEEE International Conference on Data Mining **(ICDM)***, 2015. [ICDM version]

C5 M.F. Balcan, A. Blum, **N. Haghtalab**, and A.D. Procaccia. Commitment without regrets: Online learning in Stackelberg security games. In *Proc. of the Conference on Economics and Computation **(EC)***, 2015. [EC version]

C4 A. Blum, J.P. Dickerson, **N. Haghtalab**, A.D. Procaccia, T. Sandholm, and A. Sharma. Ignorance is almost bliss: Near-optimal stochastic matching with few queries. In *Proc. of the Conference on Economics and Computation **(EC)***, 2015. [arXiv version]

C3 A. Blum, **N. Haghtalab**, and A.D. Procaccia. Learning optimal commitment to overcome insecurity. In *Advances in Neural Information Processing Systems **(NeurIPS)***, 2014. [NeurIPS version]

C2 S. Ben-David and **N. Haghtalab**. Clustering in the presence of background noise. In *Proc. of the International Conference on Machine Learning **(ICML)***, 2014. [PMLR version]

C1 A. Blum, **N. Haghtalab**, and A.D. Procaccia. Lazy defenders are almost optimal against diligent attackers. In *Proc. of the AAAI Conference on Artificial Intelligence **(AAAI)***, 2014. [AAAI version]