Andreas Haupt

AI Institute Fellow-in-Residence, Schmidt Sciences
Digital Fellow, Stanford Digital Economy Lab
Technology and Human Rights Fellow, Harvard Kennedy School

h4upt@stanford.edu andyhaupt.com Scholar

Appointments

2026–2027 **Schmidt Sciences** **AI Institute Fellow-in-Residence**
AI Institutes

2026–2027 **Stanford Digital Economy Lab** **Digital Fellow**
Stanford Institute for Human-Centered AI

2026–2027 **Harvard Kennedy School** **Technology and Human Rights Fellow**
Harvard Kennedy School

2024–2026 **Stanford University** **Human-Centered AI Fellow; Stanford Impact Labs Fellow**
Departments of Computer Science and Economics
HAI, DEL, StatsML, CRFM, STAIR
- Worked with Erik Brynjolfsson (Economics) and Sanmi Koyejo (Computer Science)

Education

2019–2024 **MIT** **Ph.D., Engineering-Economic Systems**
Computer Science and AI Lab; degree conferred February 2025.
Committee: Alessandro Bonatti, Dylan Hadfield-Menell, Eric Maskin, David Parkes.
Fields: microeconomic theory, market design, industrial organization, reinforcement learning.
Dissertation: *The Economic Engineering of Personalized Experiences*.

2017–2019 **University of Frankfurt** **B.S., Computer Science**

2015–2018 **University of Bonn** **M.S., Economics (with distinction)**

2014–2017 **University of Bonn** **M.S., Mathematics**
Year 1 at EPF Lausanne

2012–2014 **University of Bonn** **B.S., Mathematics**

References

Erik Brynjolfsson
Professor of Economics
Stanford University
erikb@stanford.edu

Eric Maskin
Professor of Economics and Mathematics
Harvard University
emaskin@fas.harvard.edu

Sanmi Koyejo
Professor of Computer Science
Stanford University
sanmi@cs.stanford.edu

David Parkes
Professor of Computer Science
Harvard University
parkes@seas.harvard.edu

Helen Nissenbaum
Professor of Information Science
Cornell Tech
hn288@cornell.edu

Dylan Hadfield-Menell
Professor of AI and Decision-Making
MIT
dhm@csail.mit.edu

Publications

2026 **Contextually Private Mechanisms**, *American Economic Review 116 (9), p. 3223–3262 (Lead Article)* (with Z. Hitzig)

Machine Learning from Human Preferences, *Princeton University Press (forthcoming)* (with S. Truong and S. Koyejo)

Discussion of "The ICML 2023 Ranking Experiment" by Su et al., *Journal of the American Statistical Association 121 (554), p. 853–854* (with S. Koyejo)

2025 **Scaling Human Judgment in Community Notes with LLMs**, *Journal of Online Trust and Safety 3 (1)* (with H. Li et al.)

2024 **Risk Preferences of Learning Algorithms**, *Games and Economic Behavior 148, p. 415–426* (with A. Narayanan)

Formal Contracts Mitigate Social Dilemmas in Multi-Agent Reinforcement Learning, *Autonomous Agents and Multi-Agent Systems 38, Article 51* (with P. Christoffersen, M. Damani, and D. Hadfield-Menell)

2023 **Understanding Multi-Homing and Switching by Platform Drivers**, *Transportation Research Part C: Emerging Technologies 154* (with X. Guo, H. Wang, R. Qadri, and J. Zhao)

Working Papers

2026 **TaskNet: Building Occupational Task Inventories from Worker Interviews at Scale**, *SSRN Preprint 7670419* (with J. Wang et al.)

AISPA: User-Centric System Prompt Auditing for Large Language Model Applications, *Working paper* (with X. Lin et al.)

General Preference Reinforcement Learning, *Working paper* (with M. Umer et al.)

Latent Adversarial Regularization for Offline Preference Optimization, *Working paper* (with E. Jiang et al.)

2025 **Preference Measurement Error, Concentration in Recommendation Systems, and Persuasion**, *Allied Social Science Associations Meeting*

The Self-Preferencing Prohibition as a Fairness Provision, *ACM Symposium on Computer Science and Law (Poster)*

Platform Preferencing and Price Competition I: Evidence from Amazon, *SSRN Preprint 5126918* (with O. Hartzell)

2023 **Recommending to Strategic Users**, *Symposium on the Foundations of Responsible Computing* (with D. Hadfield-Menell and C. Podimata)

Opaque Contracts, *Working paper* (with Z. Hitzig)

Refereed Conference Proceedings

2026 **Welfare, Improvability, and Variance: A Principal-Agent Approach to Optimal Benchmark Item Aggregation**, *Advances in Neural Information Processing Systems (Evaluations and Datasets Track)* (with J. Hartenstein, A. Reuel, M. Kochenderfer, and S. Koyejo)

Reward Bias Substitution: Single-Axis Bias Mitigations Redirect Optimization Pressure, *Advances in Neural Information Processing Systems* (with M. Lamparth, D. Fein, M. Hussing, and M.J. Kochenderfer)

Position: AI Development Should Prioritize Cognitive Security, *Advances in Neural Information Processing Systems (Position Paper)* (with B. El et al.)

Don't Walk the Line: Boundary Guidance for Filtered Generation, *International Conference on Machine Learning* (with S. Ball)

Bruno: An AI Product Manager for Scientists, *ICML Workshop on AI for Science (Best AI Scientist Award)* (with P. Ranganathan)

The Collapse of Heterogeneity in Silicon Philosophers, *ACM Conference on Fairness, Accountability, and Transparency* (with Y. Shi)

2025 **Position: ML Conferences Should Establish a Refutations and Critiques Track**, *Advances in Neural Information Processing Systems (Position Paper)* (with R. Schaeffer et al.)

AI should not be an imitation game: Centaur evaluations, *International Conference on Machine Learning (Position Paper)* (with E. Brynjolfsson)

Convex Markov Games: A New Frontier for Multi-Agent Reinforcement Learning, *International Conference on Machine Learning* (with I. Gemp, L. Marris, S. Liu, and G. Piliouras)

2024 **Steering No-Regret Learners to a Desired Equilibrium**, *ACM Conference on Economics and Computation* (with B.H. Zhang et al.)

Certification Design for a Competitive Market, *ACM Conference on Economics and Computation* (with N. Immorlica and B. Lucier)

Black-Box Access is Insufficient for Rigorous AI Audits, *ACM Conference on Fairness, Accountability, and Transparency* (with S. Casper et al.)

2023 **Computing Optimal Equilibria and Mechanisms via Learning in Zero-Sum Extensive-Form Games**, *Advances in Neural Information Processing Systems* (with B. Zhang et al.)

2022 **Towards Psychologically-Grounded Dynamic Preference Models**, *ACM Conference on Recommender Systems* (with M. Curmei, B. Recht, and D. Hadfield-Menell)

2021 **The Optimality of Upgrade Pricing**, *Web and Internet Economics* (with D. Bergemann, A. Bonatti, and A. Smolin)

Honors and Fellowships

2026 Best AI Scientist Award, ICML Workshop on AI for Science (for *Bruno*)

2025 Rising Star in Data Science

2025 Young Scientist, 8th Economics Lindau Nobel Laureate Meeting

2025 Young Researcher, 12th Heidelberg Laureate Forum

2024 Price Theory Summer Camp, Becker Friedman Institute

2019 Presidential Fellowship, MIT

2012–2018 Full Scholarship, Villigst Foundation

Policy and Teaching Experience

2024–2026 **Stanford University**, Co-Primary Instructor: CS329H (Machine Learning from Human Preferences)

2023 **Federal Trade Commission**, Summer Fellow: memos on AI, competition, and consumer protection

2021 **European Commission**, DG COMP Trainee: voice assistants inquiry; Amazon data case

2017–2019 **Teach First Deutschland**, High School Teacher: mathematics and computer science

Professional Activities

Talks NYU, CMU, Harvard Kennedy School, Lindau, Microsoft Research, Google DeepMind, HBS

Organized SimulacraBench (NeurIPS 2026), The Future of Authorship (Stanford DEL 2026), LLM Incentives (ACM EC 2026), H+AI@EC (ACM EC 2025)

Service Area Chair, ICLR/ICML/AAAI (2026–27); Co-Chair, Stanford HAI Affinity Group on Global AI Governance (2024–26); MIT AI Ethics & Policy (2021–23); MIT Science Policy Initiative (2022–23)