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

As of November 2026: 22 peer-reviewed or in-press publications, 9 working papers and preprints, and 5 theses, in reverse chronological order. The 3 most important publications are marked with a star and listed first, with my contribution. Only peer-reviewed and in-press work is counted as published; preprints, submitted work and theses are listed separately and marked. **Open Access** marks publications freely available online; every entry links to the publisher's version or a preprint.

Three Most Important Publications

★ **Welfare, Improvability, and Variance: A Principal-Agent Approach to Optimal Benchmark Item Aggregation**
A. Haupt‡, J. Hartenstein‡, A. Reuel, M. Kochenderfer, S. Koyejo
Advances in Neural Information Processing Systems (Evaluations and Datasets Track), 2026. Open Access AI Evaluation Aggregation Market Design PDF HTML
First author; shared first authorship with J. Hartenstein (equal contribution, 50% each of the joint work); M. Kochenderfer and S. Koyejo are senior co-authors.

★ **Contextually Private Mechanisms**
A. Haupt‡, Z. Hitzig‡
American Economic Review 116 (9), p. 3223-3262 (Lead Article), 2026. Open Access Privacy Market Design PDF HTML
Equal contribution with Z. Hitzig (50% each); authors in alphabetical order, as is customary in economics. Lead article of the issue.

★ **Formal Contracts Mitigate Social Dilemmas in Multi-Agent Reinforcement Learning**
A. Haupt‡, P. Christoffersen‡, M. Damani, D. Hadfield-Menell
Autonomous Agents and Multi-Agent Systems 38, Article 51, 2024. Multi-Agent Market Design PDF CODE
First author; shared first authorship with P. Christoffersen (equal contribution, 50% each of the joint work); D. Hadfield-Menell is the senior co-author.

Peer-Reviewed and In-Press Publications

‡ indicates equal contribution or alphabetical author order, † equal supervision.

★ [1] **Welfare, Improvability, and Variance: A Principal-Agent Approach to Optimal Benchmark Item Aggregation**
A. Haupt‡, J. Hartenstein‡, A. Reuel, M. Kochenderfer, S. Koyejo
Advances in Neural Information Processing Systems (Evaluations and Datasets Track), 2026.
Open Access · AI Evaluation · Aggregation · Market Design · PDF · HTML

[2] **Reward Bias Substitution: Single-Axis Bias Mitigations Redirect Optimization Pressure**
M. Lamparth, D. Fein, **A. Haupt**, M. Hussing, M.J. Kochenderfer
Advances in Neural Information Processing Systems, 2026.
Open Access · Preferences · PDF

[3] **Position: AI Development Should Prioritize Cognitive Security**
B. El, S. Su, A. Pappu, P. Yin, J. Heng, E. Heng, R.Z. Wang, **A. Haupt**†, J. Zou†
Advances in Neural Information Processing Systems (Position Paper), 2026.
Open Access · Preferences · PDF

★ [4] **Contextually Private Mechanisms**
A. Haupt‡, Z. Hitzig‡
American Economic Review 116 (9), p. 3223-3262 (Lead Article), 2026.
Open Access · Privacy · Market Design · PDF · HTML

[5] **Machine Learning from Human Preferences**
S. Truong‡, **A. Haupt**‡, S. Koyejo‡
Princeton University Press (forthcoming)
In press · Preferences · HTML

[6] **Don't Walk the Line: Boundary Guidance for Filtered Generation**
S. Ball, **A. Haupt**
International Conference on Machine Learning, 2026.
Open Access · Multi-Agent · PDF

[7] **Bruno: An AI Product Manager for Scientists**
A. Haupt‡, P. Ranganathan‡
ICML Workshop on AI for Science (Best AI Scientist Award), 2026.
Work · HTML · VIDEO

[8] **The Collapse of Heterogeneity in Silicon Philosophers**
Y. Shi, **A. Haupt**
ACM Conference on Fairness, Accountability, and Transparency, 2026.
AI Evaluation · Preferences · PDF · CODE

[9] **Discussion of "The ICML 2023 Ranking Experiment" by Su et al.**
A. Haupt, S. Koyejo
Journal of the American Statistical Association 121 (554), p. 853-854, 2026.
AI Evaluation · Aggregation

[10] **Position: ML Conferences Should Establish a Refutations and Critiques Track**
R. Schaeffer, J. Kazdan, Y. Denisov-Blanch, B. Miranda, M. Gerstgrasser, S. Zhang, **A. Haupt**, I. Gupta, E. Obbad, J. Dodge, et al.
Advances in Neural Information Processing Systems (Position Paper), 2025.
Open Access · AI Evaluation · PDF

[11] **Scaling Human Judgment in Community Notes with LLMs**
H. Li, S. De, M. Revel, **A. Haupt**, B. Miller, K. Coleman, J. Baxter, M. Saveski, M.A. Bakker
Journal of Online Trust and Safety 3 (1), 2025.
Open Access · AI Evaluation · Aggregation · Platforms · PDF

[12] **AI should not be an imitation game: Centaur evaluations**
A. Haupt, E. Brynjolfsson
International Conference on Machine Learning (Position Paper), 2025.
Open Access · AI Evaluation · Work · PDF

[13] **Convex Markov Games: A New Frontier for Multi-Agent Reinforcement Learning**
I. Gemp, **A. Haupt**, L. Marris, S. Liu, G. Piliouras
International Conference on Machine Learning, 2025.
Open Access · Multi-Agent · PDF

[14] **Risk Preferences of Learning Algorithms**
A. Haupt‡, A. Narayanan‡
Games and Economic Behavior 148, p. 415-426, 2024.
Multi-Agent Preferences PDF

★ [15] **Formal Contracts Mitigate Social Dilemmas in Multi-Agent Reinforcement Learning**
A. Haupt‡, P. Christoffersen‡, M. Damani, D. Hadfield-Menell
Autonomous Agents and Multi-Agent Systems 38, Article 51, 2024.
Multi-Agent Market Design PDF CODE

[16] **Steering No-Regret Learners to a Desired Equilibrium**
B.H. Zhang, G. Farina, I. Anagnostides, F. Cacciamani, S.M. McAleer, **A. Haupt**, A. Celli, N. Gatti, V. Conitzer, T. Sandholm
ACM Conference on Economics and Computation, 2024.
Open Access Multi-Agent Market Design PDF

[17] **Certification Design for a Competitive Market**
A. Haupt‡, N. Immorlica‡, B. Lucier‡
ACM Conference on Economics and Computation, 2024.
Market Design PDF

[18] **Black-Box Access is Insufficient for Rigorous AI Audits**
S. Casper, C. Ezell, C. Siegmann, N. Kolt, T.L. Curtis, B. Bucknall, **A. Haupt**, K. Wei, J. Scheurer, M. Hobbhahn, et al.
ACM Conference on Fairness, Accountability, and Transparency, 2024.
AI Evaluation PDF

[19] **Computing Optimal Equilibria and Mechanisms via Learning in Zero-Sum Extensive-Form Games**
B. Zhang, G. Farina, I. Anagnostides, F. Cacciamani, S. McAleer, **A. Haupt**, A. Celli, N. Gatti, V. Conitzer, T. Sandholm
Advances in Neural Information Processing Systems, 2023.
Open Access Multi-Agent Market Design PDF

[20] **Understanding Multi-Homing and Switching by Platform Drivers**
X. Guo‡, **A. Haupt**‡, H. Wang, R. Qadri, J. Zhao
Transportation Research Part C: Emerging Technologies 154, 2023.
Platforms Work PDF

[21] **Towards Psychologically-Grounded Dynamic Preference Models**
M. Curmei‡, **A. Haupt**‡, B. Recht, D. Hadfield-Menell
ACM Conference on Recommender Systems, 2022.
Preferences PDF

[22] **The Optimality of Upgrade Pricing**
D. Bergemann‡, A. Bonatti‡, **A. Haupt**‡, A. Smolin‡
Web and Internet Economics, 2021.
Market Design PDF

Working Papers and Preprints

[23] **TaskNet: Building Occupational Task Inventories from Worker Interviews at Scale**
J. Wang[‡], **A. Haupt**[‡], I. Robinson, P. Lewis, E. Brynjolfsson[†], D. Yang[†]
SSRN Preprint 7670419, 2026.
Preprint, not peer reviewed | Open Access | Work | AI Evaluation | PDF

[24] **AISPA: User-Centric System Prompt Auditing for Large Language Model Applications**
X. Lin, S. Zhu, S. Yang, Z. Zhang, H. Zhang, Y. Zhao, C. Qian, T. Wang, Z. Zhang, Z. Yuan, D. Wu, J. Wu, Y. Si, J. Liu, B. Bi, R. Mahari, T. South, D. Greenwood, Z. He, R. Bommasani, S. Kazinnik, **A. Haupt**, S. Marro, E. Brynjolfsson, A. Pentland, J. Pei
Preprint, 2026.
Preprint, not peer reviewed | Open Access | AI Evaluation | Platforms | PDF

[25] **General Preference Reinforcement Learning**
M. Umer, M.A. Mohsin, A. Bilal, A. Chaudhry, **A. Haupt**, S. Koyejo, E. Fox, J.M. Cioffi
Preprint, 2026.
Preprint, not peer reviewed | Open Access | Preferences | PDF

[26] **Latent Adversarial Regularization for Offline Preference Optimization**
E. Jiang, Y.J. Zhang, Y. Xu, **A. Haupt**, N. Amato, S. Koyejo
Preprint, 2026.
Preprint, not peer reviewed | Open Access | Preferences | PDF

[27] **Preference Measurement Error, Concentration in Recommendation Systems, and Persuasion**
A. Haupt
Allied Social Science Associations Meeting, 2025.
Preprint, not peer reviewed | Open Access | Preferences | Platforms | PDF

[28] **The Self-Preferencing Prohibition as a Fairness Provision**
A. Haupt
ACM Symposium on Computer Science and Law (Poster), 2025.
Poster, non-archival | Open Access | Platforms | Market Design | PDF

[29] **Platform Preferencing and Price Competition I: Evidence from Amazon**
O. Hartzell[‡], **A. Haupt**[‡]
SSRN Preprint 5126918, 2025.
Preprint, not peer reviewed | Open Access | Platforms | PDF

[30] **Recommending to Strategic Users**
A. Haupt, D. Hadfield-Menell, C. Podimata
Symposium on the Foundations of Responsible Computing, 2023.
Peer-reviewed, non-archival | Open Access | Preferences | Platforms | PDF

[31] **Opaque Contracts**
A. Haupt, Z. Hitzig
Preprint, 2023.
Preprint, not peer reviewed | Open Access | Privacy | Market Design | PDF

Theses

[32] **The Economic Engineering of Personalized Experiences**
A. Haupt
Ph.D. Dissertation, Massachusetts Institute of Technology, 2025.
Open Access Preferences PDF

[33] **Multi-Agent Influence Diagrams and Commitment**
A. Haupt
B.S. Thesis, Goethe Universität Frankfurt, 2019.
Open Access Multi-Agent PDF

[34] **Voting with Restricted Communication**
A. Haupt
M.S. Thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2018.
Open Access Market Design Aggregation PDF

[35] **A Data Application of Graphon Theory**
A. Haupt
M.S. Thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2017.
Open Access Platforms PDF

[36] **Die Integrality Ratio der Subtour-Relaxierung**
A. Haupt
B.S. Thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2014.
Open Access PDF