WILLIAM OVERMAN

655 Knight Way, Stanford, CA 94305, USA
wpo@stanford.edu | woverman.com | Google Scholar | LinkedIn

RESEARCH INTERESTS

AI Safety | AI Alignment | Reinforcement Learning | Causal Inference | Healthcare

EDUCATION

Stanford University — ***2022-2026 (exp.)***
Graduate School of Business
Ph.D. Operations, Information and Technology

University of California, Irvine — *2020-2022*
M.Sc., Computer Science

California Institute of Technology (Caltech) — *2016-2020*
B.S., Mathematics & Computer Science (Double Major)

INDUSTRY AND RESEARCH POSITIONS

Uber

PhD Scientist Intern, Uber Eats — 2025, New York, NY
PhD Scientist Intern, Rider Science — 2024, San Francisco, CA
PhD Scientist Intern, Airports — 2023, Sunnyvale, CA

Institute for Basic Science
Visiting Researcher — 2021, Daejeon, South Korea
Summer Undergraduate Research Fellowship — 2019, Daejeon, South Korea

Oxford University
Summer Undergraduate Research Fellowship — 2018, Oxford, United Kingdom
Mathematical Institute

PUBLICATIONS

[REFEREED CONFERENCE PAPERS]

- **Calibrating Conservatism for Scalable Oversight**
  W. Overman, M. Bayati.
  ICML 2026: International Conference on Machine Learning, 2026.

- **The Oversight Game: Learning to Cooperatively Balance an AI Agent's Safety and Autonomy**
  W. Overman, M. Bayati.
  ICML 2026: International Conference on Machine Learning, 2026.
  Early version in NeurIPS 2025 Workshop: ML×OR Workshop (Oral).

- **Annealed Softmax Greedy in Many-Armed Bayesian Bandits**
  W. Overman, M. Bayati.
  RLC 2026: Reinforcement Learning Conference, 2026.

- **Conformal Arbitrage: Risk-Controlled Balancing of Competing Objectives in Language Models**
  W. Overman, M. Bayati.
  NeurIPS 2025: Neural Information Processing Systems, 2025.

- **Aligning Model Properties via Conformal Risk Control**
  W. Overman, J. Vallon, M. Bayati.
  NeurIPS 2024: Neural Information Processing Systems, 2024.

- **Higher-Order Causal Message Passing for Experimentation with Complex Interference**
  M. Bayati, Y. Luo, **W. Overman**, S. Shirani, R. Xiong. (alphabetical).
  NeurIPS 2024: Neural Information Processing Systems, 2024.

- **Improved Regret Bound for Safe Reinforcement Learning via Tighter Cost Pessimism and Reward Optimism**
  K. Yu, D. Lee, **W. Overman**, D. Lee.
  RLC 2025: Reinforcement Learning Conference, 2025.

- **Beating Price of Anarchy and Gradient Descent without Regret in Potential Games**
  I. Sakos, S. Leonardos, S. Stavroulakis, **W. Overman**, I. Panageas, G. Piliouras.
  ICLR 2024: International Conference on Learning Representations, 2024.

- **Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games**
  S. Leonardos, **W. Overman**, I. Panageas, G. Piliouras (**alphabetical, sole student**).
  ICLR 2022: International Conference on Learning Representations, 2022.

- **Independent Natural Policy Gradient always converges in Markov Potential Games**
  R. Fox, S. Mcaleer, **W. Overman**, I. Panageas (**alphabetical, sole student**).
  AISTATS 2022: International Conference on Artificial Intelligence and Statistics, 2022.

[SUBMITTED AND WORKING PAPERS]

- **Planning Against Learning in Rank-1 Games**
  W. Overman
- **Causal Effects with Unobserved Unit Types in Interacting Human–AI Systems**
  W. Overman, S. Shirani, M. Bayati.
  Early version in **ICML 2026 Workshop on Technical AI Governance Research**

- **Occupancy Prediction with Patient Data: Evaluating Time-Series, Patient-Level Aggregation, and Deep Set Models**
  S.-H. Kim, **W. Overman**, J. Pauphilet, W. Cha. (**alphabetical, sole student**).
  Major Revision at MSOM: Manufacturing & Service Operations Management.

- **Can We Validate Counterfactual Estimations in the Presence of General Network Interference?**
  S. Shirani, Y. Luo, **W. Overman**, R. Xiong, M. Bayati.
  Major Revision at Management Science
  ***Oral Presentation CODE@MIT 2025**: Conference on Digital Experimentation @ MIT, 2025*

- **On Aligning Prediction Models with Clinical Experiential Learning: A Prostate Cancer Case Study**
  J. J. Vallon, **W. Overman**, W. Xu, N. Panjwani, X. Ling, S. Vij, H. P. Bagshaw, J. T. Leppert, S. Shah, G. Sonn, S. Srinivas, E. Pollom, M. K. Buyyounouski, M. Bayati.

[JOURNAL PAPERS]

- **Some Ordered Ramsey Numbers of Graphs on Four Vertices**
  W. Overman, J. Alm, K. Coffey, C. Langhoff.
  Australasian Journal of Combinatorics, 2024.

☆ AWARDS & SERVICE

Thomas A. Tisch Prize for Undergraduate Teaching in Computing and Mathematical Sciences, Caltech, *2020*

Dean's Award, UC Irvine, for outstanding research potential. *2020*
Morgan Ward Prize, Caltech, best original work in mathematics by an underclassman. *2018*
Conference Reviewer: NeurIPS (2024, 2025), ICLR (2026), AISTATS (2026)
Silver Reviewer Award ICML 2026

TEACHING

TA, Business Intelligence from Big Data Stanford, Graduate School of Business, 2026 Winter
TA, Operations Stanford, Graduate School of Business, 2024 Fall
TA, Optimization and Simulation Modeling Stanford, Graduate School of Business, 2023 Fall
TA, Introduction to Algorithms University of California, Irvine 2022 Spring
TA, Introduction to Cryptography University of California, Irvine 2021 Winter
TA, Introduction to Algorithms California Institute of Technology 2020 Spring
TA, Decidability and Tractability California Institute of Technology 2020 Winter
TA, Discrete Mathematics California Institute of Technology 2019 Fall
TA, Introduction to Algorithms California Institute of Technology 2019 Spring
TA, Discrete Mathematics California Institute of Technology 2018 Fall
TA, Introduction to Algorithms California Institute of Technology 2018 Spring

INVITED PRESENTATIONS

Seoul National University Data Science Colloquium, Seoul, South Korea *Sep. 2026*
Control of Increasingly Capable AI Systems with Safety Guarantees

NeurIPS ML x OR Workshop Spotlight Presentation, San Diego, CA *Dec. 2025*
The Oversight Game: Learning AI Control & Corrigibility in Markov Games

KAIST ISE Winter Symposium, Daejeon, South Korea *Dec. 2025*
Control and Oversight of Black-Box AI Models with Formal Guarantees

Seoul National University Data Science Seminar, Seoul, South Korea *Oct. 2025*
Black-Box Model Alignment: Why It Matters and How We Can Approach It

KAIST AI Invited Talk, Seoul, South Korea *Sep. 2025*
Alignment of Black-Box Models via Conformal Risk Control

Seoul National University GSDS Invited Talk, Seoul, South Korea *Sep. 2025*
Alignment of Black-Box Models via Conformal Risk Control

GSB Bridging Humans and Machines: Advancing Alignment in AI, Stanford, CA *Apr. 2025*
Aligning Model Properties via Conformal Risk Control

INFORMS Annual Meeting, Seattle, WA *Oct. 2024*
Aligning Model Properties via Conformal Risk Control

Science Jam Uber Technologies, San Franscisco, CA *Aug. 2024*
Testing for Network Interference in A/B Tests

INFORMS Annual Meeting, Phoenix, AZ *Oct. 2023*
Calibrate to Operate: Aligning Patient-Level Predictions with System-Wide Forecasts

Discrete Math Seminar, IBS Discrete Mathematics Group, Daejeon, South Korea *Apr. 2021*
Some Ordered Ramsey Numbers of Graphs on Four Vertices

Caltech CS Theory Seminar, Pasadena, CA *May 2020*
Boolean Functions on the Infinite Hamming Cube