Stanford



Max Lamparth

Research Fellow, HOOVER RESEARCH

Bio

BIO

Max Lamparth is a Research Fellow at the Stanford Intelligent Systems Laboratory, the Stanford Center for AI Safety, and the Hoover Institution's Technology Policy Accelerator at Stanford University.

He is building AI systems that act with us and for us, which requires that AI systems learn what we value correctly. Any objective learned from human data is an approximation of human intent, and errors compound silently from how judgments are elicited and aggregated to how they are optimized against and evaluated. He works on rewards that stay valid under optimization, proves when human data identifies an objective, and designs evaluations with domain experts that certify learned behavior and decide what we must elicit next. His work includes publications at NeurIPS, ICML, ICLR, CoLM, FAccT, and AIES, and has been recognized with spotlights. It has informed deployment practices at major AI developers and state, federal, and international policymakers across party lines through briefings, publications in Foreign Affairs, and invited feedback on AI legislation. He has designed and taught multipel courses at the intersection of computer science, machine learning, statistics, and societal impact of AI at Stanford University since 2023.

Prior to his current appointment, he was a postdoctoral fellow at Stanford University in the Department of Computer Science, the Stanford Center for AI Safety, with a fellowship from the Center for International Security and Cooperation and the Stanford Existential Risks Initiative advised by Prof. Clark Barrett, Prof. Steve Luby, and Prof. Paul Edwards. He received his Ph.D. in Physics from the Technical University of Munich in July 2023 and holds a B.Sc. and M.Sc. in Physics from the Ruprecht Karl University of Heidelberg.

ACADEMIC APPOINTMENTS

- Research Fellow, HOOVER RESEARCH

LINKS

- Homepage: http://www.maxlamparth.com
- Twitter/X: https://x.com/MLamparth
- LinkedIn: https://www.linkedin.com/in/maxlamparth/

Teaching

COURSES

2025-26

- Introduction to AI Safety: CS 120 (Aut)

2024-25

- Introduction to AI Governance: CS 134, STS 14 (Win)
- Introduction to AI Safety: CS 120 (Aut)

2023-24

- Introduction to AI Safety: CS 120, STS 10 (Spr)

Publications

PUBLICATIONS

- **A benchmark of expert-level academic questions to assess AI capabilities.** *Nature*
  Center for AI Safety, Scale AI, HLE Contributors Consortium, Phan, L., Gatti, A., Li, N., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., Hendrycks, D., Han, Z., et al
  2026; 649 (8099): 1139-1146

- **Escalation Risks from Language Models in Military and Diplomatic Decision-Making**
  Rivera, J., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., Schneider, J., Assoc Computing Machinery
  ASSOC COMPUTING MACHINERY.2024: 836-898

- **Analyzing And Editing Inner Mechanisms of Backdoored Language Models**
  Lamparth, M., Reuel, A., Assoc Computing Machinery
  ASSOC COMPUTING MACHINERY.2024: 2362-2373

- **Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations**
  Lamparth, M., Corso, A., Ganz, J., Mastro, O., Schneider, J., Trinkunas, H., Association for the Advancement of Artificial Intelligence
  ASSOC COMPUTING MACHINERY.2024: 807-817

- **Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation (Extended Abstract)**
  Grabb, D., Lamparth, M., Vasan, N., Association for the Advancement of Artificial Intelligence
  ASSOC COMPUTING MACHINERY.2024: 519

- **Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations**
  Lamparth, M., Corso, A., Ganz, J., Mastro, O., Schneider, J., Trinkunas, H.
  edited by Das, S., Green, B. P., Varshney, K., Ganapini, M., Renda, A.
  ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE.2024: 807-817

- **Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation (Extended Abstract)**
  Grabb, D., Lamparth, M., Vasan, N.
  edited by Das, S., Green, B. P., Varshney, K., Ganapini, M., Renda, A.
  ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE.2024: 519

- **BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices**
  Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., Kochenderfer, M. J.
  edited by Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C.
  NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).2024