Stanford

Max Lamparth

Research Fellow

HOOVER RESEARCH

Bio

BIO

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

With his research, he is working towards making AI systems inherently more secure and safe, providing critical insights to inform and guide effective AI policies, and shape public discourse. He specializes in interpretability and robustness of AI systems, ethical decision-making of language models, and uncertainty quantification. His work aims to promote the safe and responsible use of AI in society, with a particular emphasis on language models for automated decision-making, and has been recognized through publications in leading technical and socio-technical conferences such as NeurIPS, CoLM, FAccT, and AIES, as well as policy-oriented outlets like Foreign Affairs. Additionally, his research has garnered attention from international media, with coverage in the MIT Technology Review, The Washington Times, The Japan Times, LaPress, Axios, Deutschlandfunk, and New Scientist.

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

ACADEMIC APPOINTMENTS

- Hoover 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