Stephen Casper

✉ scasper@mit.edu ☎ (208) 589-2225 🔗 stephencasper.com/

🔗 Twitter/X 🔗 BlueSky 🔗 Google Scholar

October 21, 2025

Introduction

Hi, my name is Stephen Casper, but you can call me Cas. I'm a final-year CS Ph.D. student at MIT and a former research resident at the UK AI Security Institute. I work on frontier AI risk managment problems including robustness, evals, and technically-rigorous governance. I love teaching, mentoring, and doing exciting research. So I'm thrilled to to be on the academic job market this fall!

Education

Harvard University *Sept 2017 – May 2021*
B.A in Statistics, Secondary field in Mathematical Sciences
Graduation cum laude (Latin) and with highest honors (English). GPA: 3.847/4.0

MIT *Sept 2021 – Feb 2024*
M.S. in Electrical Engineering and Computer Science
GPA: 4.8/5.0

MIT *Feb 2024 – May 2026 (Expected)*
Ph.D. in Electrical Engineering and Computer Science
Minor in Public Policy

Professional Experience

Research Assistant *Center for Brains, Minds, and Machines & Boston Children's Hospital* *June 2021 – Sept 2021*

- Interpreting and red-teaming image classifiers using feature-level adversarial attacks

Mentor *Machine Learning Alignment & Theory Scholars* *Jan 2024 - Present*

- Mentoring successful research projects on machine unlearning, latent adversarial training, evaluating LLMs under tampering attacks, and rigorous evaluation reporting.

Mentor *ERA Fellowship* *May 2024 - Present*

- Mentoring successful research projects on LLM jailbreaking, LLM unlearning, and LLM tamper resistance.

Teaching Assistant *Aug 2024 - Dec 2024*

- Creating assignments, holding recitations, guest lecturing, and grader management for MIT's 6.3950, AI Decision-Making and Society course

Research Resident *UK AI Security Institute* *Feb 2025 - Aug 2025*

- Leading UK AISI's technical work on open-weight risk management. I oversaw and executed a £750,000 research project on developing state-of-the-art tamper-resistant LLM safeguards for open-weight LLMs.

Mentor *GovAI Fellows Program* *Sept 2025 - Present*

- Mentoring an oingoing research project about regulatory gaps for internally deployed AI systems.

Awards

Hoopes Prize *Harvard University*
May 2021

Best Paper Award *NeurIPS ML Safety Workshop*
Dec 2022

Spotlight Paper *ICML GenLaw Workshop*
July 2023

Outstanding Paper Finalist *TMLR*
Dec 2024

Press & Newsletter Mentions

...covering me or papers that I was a leading contributor to:

- **Montreal AI Ethcis Institute:** Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback (Sept 2023)
- **The Globe and Mail:** Meet the gig workers making AI models smarter (Sept, 2023)
- **IEEE Spectrum:** OpenAI Builds AI to Critique AI (June 2024)
- **MIT FutureTech:** Presenting the AI Risk Repository (Aug 2024)
- **Center for AI Policy:** New Analysis of AI Agents Highlights a Serious Lack of Safety Oversight (Feb 2025)
- **Tech Policy Press:** Researchers Develop an AI Agent Index to Inform Governance of Agentic Systems (Mar 2025)
- **TechCrunch:** MIT study finds that AI doesn't, in fact, have values (April 2025)
- **AI Frontiers:** Smokescreen: How Bad Evidence Is Used to Prevent AI Safety (April 2025)
- **Business Insider:** AI isn't ready to do your job (April 2025)
- **TechCrunch:** Anthropic is launching a new program to study AI 'model welfare' (April 2025)
- **Wired:** Singapore's Vision for AI Safety Bridges the US-China Divide (May 2025)
- **IMDA:** Top scientific minds gathered for the first time in Singapore to advance AI that is trustworthy, reliable and secure (May 2025)
- **EuroNews:** There is a global consensus for AI safety despite Paris Summit backlash, new report finds (May 2025)
- **WebProNews:** Singapore's AI Diplomacy: Forging a Middle Path for Global AI Governance (May 2025)
- **OECD.AI:** Strengthening global AI Safety: A perspective on the Singapore Consensus (June 2025)
- **University of Oxford Press:** Study finds filtered data stops openly-available AI models from performing dangerous tasks (Aug 2025)
- **Washington Post:** AI systems 'ignorant' of sensitive data can be safer, but still smart (Aug 2025)
- **Tech Xplore:** Filtered data stops openly-available AI models from performing dangerous tasks, study finds (Aug 2025)
- **EdTech Innovation Hub:** Oxford University and partners build tamper-resistant safeguards into open-source AI models (Aug 2025)
- **Actu.ai:** Filtered data prevent publicly accessible AI models from performing dangerous tasks, according to a study (Aug 2025)
- **Perplexity:** AI models taught 'deep ignorance' resist bioweapon training (Aug 2025)
- **AISI Blog:** Managing Risks from Increasingly Capable Open-Weight AI Systems (Aug 2025)
- **ML Safety Newsletter:** Filtering Dangerous Training Data (Sept 2025)

Invited Talks, Panels, & Podcasts

- **AXRP Podcast:** Interpretability for Engineers with Stephen Casper (May 2023)
- **CHAI Workshop:** Realistic Red-Teaming in Large Langauge Models (June 2023)
- **University of Alberta AI Seminar Series:** Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback (Sept 2023)
- **EAG Boston 2023:** Lessons from RLHF on the Difficulties of Aligning Advanced AI (Oct 2023)
- **FAR New Orleans Alignment Workshop:** Why do LLM Outputs Disagree with Internal Representations of Truthfulness? (Dec 2023)
- **CHAI Workshop:** Defending against Persistent Harmful Behaviors in LLMs with Latent Adversarial

Training (June 2024)

- **FAR Vienna Alignment Workshop:** Generalized Adversarial Training and Testing (Jul 2024)
- **CAIP Podcast:** Stephen Casper on Technical and Sociotechnical AI Safety Research (Aug 2024)
- **FAR Bay Area Alignment Workshop:** Powering Up AI Capability Evaluations (Oct 2024)
- **EAG Boston 2024:** Pitfalls of Evidence-Based AI Policy (Oct 2024)
- **Americans for Responsible Innovation:** AI Model Piracy, Virtual Panel (Feb 2025)
- **Human Feedback Paper Group:** MIT on The AI Agent Index (April 2025)
- **ICLR 2025 Workshop on Human-AI Coevolution:** Open Problems and Fundamental Limitations of RLHF (April 2025)
- **CHAI Workshop:** LLM Tamper Resistance as a Key Priority for AI Safety (June 2025)
- **ICML 2025 2nd Workshop on Models of Human Feedback for AI Alignment (MoFA):** Panel (July 2025)
- **ICML 2025 Technical AI Governance Workshop:** Panel: Technical AI Governance as a Field (Moderator) (July 2025)

Research Mentorship Experience

Mentees of mine have gone on to research and leadership positions at Anthropic, DeepMind, Epoch, Eleuther, Meta, OpenAI, the UK AI Security Institute, and numerous PhD programs. Two mentees went on to found Harmony Intelligence and Fulcrum.

Here, I list papers that I worked on with mentees where I was the joint-first or last author and the mentee(s) were the first or joint-first:

- **Max Nadeau:** Robust feature-level adversaries are interpretability tools
- **Kaivalya Hariharan:** Diagnostics for deep neural networks with automated copy/paste attacks
- **Gatlen Culp, Jason Lin, Joe Kwon:** Explore, establish, exploit: Red teaming language models from scratch
- **Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang:** Red teaming deep neural networks with feature synthesis tools
- **Anson Ho, Tilman Räuker:** Toward transparent AT: A survey on interpreting the inner structures of deep neural networks
- **Xander Davies:** Open problems and fundamental limitations of reinforcement learning from human feedback
- **Lennart Schulze:** Defending against unforeseen failure modes with latent adversarial training
- **Carson Ezell:** Black-box access is insufficient for rigorous ai audits
- **Soroush Pour, Rusheb Shah, Quentin Feuilade-Montixi, Arush Tagade:** Scalable and transferable black-box jailbreaks for language models via persona modulation
- **Phillip Guo, Aengus Lynch, Aidan Ewart, Cindy Wu, Abhay Sheshadri:** Latent adversarial training improves robustness to persistent harmful behaviors in llms
- **Nathalie Kirch, Severin Field:** What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks
- **Zora Che:** Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities
- **Ariba Khan:** Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs
- **Leon Staufer, Mick Yang:** Audit Cards: Contextualizing AI Evaluations
- **Kyle O'Brien:** Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs

Research Highlights

Google Scholar: See my Google Scholar page for a complete list.
H-Index: 23
I10-Index: 34
Citations: 3110

Technical AI Safeguards

- **Casper, S.***, Hariharan, K.*, Hadfield-Menell, D., (2022). Diagnostics for Deep Neural Networks with Automated Copy/Paste Attacks.
  - Best Paper Award, NeurIPS ML Safety Workshop 2022 (link)
- Sheshadri, A., Ewart, A., Guo, P., Lynch, A., Wu, C., Hebbar, V., Sleight, H., Cooper Stickland A., Perez, E., Hadfield-Menell, D., & **Casper, S.** (2024). Targeted Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. TMLR.
- O'Brien, K.*, **Casper, S.***, Anthony, Q., Korbak, T., Kirk, R., Davies, X., ... & Biderman, S. (2025). Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs. Oral presentation, NeurIPS 2025 workshop on Biosecurity for Generative AI

AI Evals, Red-Teaming, & Risk Assessment

- **Casper, S.**, Hadfield-Menell, D., Kreiman, G. (2022). Red-Teaming with Mind-Reading: White-Box Adversarial Policies in Deep Reinforcement Learning.
  - Hoopes Prize (link)
- **Casper, S.***, Ezell, C.*, Siegmann, C., Kolt, N., Curtis, T., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., Hadfield-Menell, D. (2024) Black-Box Access is Insufficient for Rigorous AI Audits. Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2024.
- Che, Z.,* **Casper, S.,*** Kirk, R., Satheesh, A., Slocum, S., McKinney, L. E., ... & Hadfield-Menell, D. (2025). Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities. TMLR.

AI, Governance, and Society

- **Casper, S.***, Guo, Z.*, Mogulothu, S., Marinov, Z., Deshpande, C., Yew, R. J., Dai, Z., & Hadfield-Menell, D. (2023). Measuring the Success of Diffusion Models at Imitating Human Artists.
  - Spotlight Paper: Genlaw Workshop 2024 (link)
- **Casper, S.**, Bailey, L., Hunter, R., Ezell, C., Cabalé, E., Gerovitch, M., ... & Kolt, N. (2025). The AI Agent Index. Website.
- Khan, A., **Casper, S.**, & Hadfield-Menell, D. (2025). Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. Proceedings of the 2025 ACM conference on fairness, accountability, and transparency, 2025.
- **Casper, S.**, Krueger, D., & Hadfield-Menell, D. (2025). Pitfalls of Evidence-Based AI Policy. ICLR 2025 Blog post.

Systemization of Knowledge

- **Casper, S.***, Davies, X.*, Shi, C., Gilbert, T., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., Hadfield-Menell, D. (2023) Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. TMLR Survey Certification, TMLR Featured Certification
  - Outstanding Paper Finalist, TMLR 2024 (link)
- Reuel, A., Bucknall, B., **Casper, S.**, Fist, T., Soder, L., Aarne, O., ... & Trager, R. (2024). Open Problems in Technical AI Governance. TMLR Survey Certification
- Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., **Casper, S.**, ... & Zeng, Y. (2025). International AI Safety Report.
- Bengio, Y., Maharaj, T., Ong, L., Russell, S., Song, D., Tegmark, M., Xue, L., Zhang, Y., Casper, S.... & Žikelić, D. (2025). The Singapore Consensus on Global AI Safety Research Priorities.

Miscellaneous

- Anecdotal places where my work has been influential:
  - Course curricula at Stanford, Princeton, Toronto, BlueDot, and the AISES Alignment Textbook.
  - The International AI Safety Report (including sections that I both was an was not a writer for).
  - The California Report on Frontier AI Policy
  - Litigation in the US District Court, Northern District of California San Francisco Division
  - New-York State Representative Alex Bores's office communicated to me that Public Perspectives on AI Governance was helpful in discussions around the proposed RAISE Act.
  - (Speculative) It is possible that my paper, Practical Principles for AI Cost and Compute Accounting, may have influenced Section 22757.15 of California's SB 53 bill. To my knowledge, SB 53 is the only frontier AI bill that makes special provisions for regulators to update compute thresholds over time in response to technical developments. It was introduced soon after our paper was published, and after we had communicated it to stakeholders in California involved in the legislative process. Separately, I also found the EU AI office was interested in the paper, and met with them in July, 2024.
- Blog post: Reframing AI Safety as a Neverending Institutional Challenge (2025)
- Blog post: Managing risks from increasingly capable open-weight AI systems (2025)
- Blog post sequence: The Engineer's Interpretability Sequence (2023)
- I was a member of the EU AI Act Codes of Practice Plenary (2024-2025)
- I am working to get two new invisible Unicode characters established: one for watermarking AI-generated text, and one for indicating author non-consent to AI model training.
- I have a personal anonymous feedback form.
- My favorite AI-related quotes:
  - *"They are wrong who think that politics is like an ocean voyage or a military campaign, something to be done with some particular end in view, something which leaves off as soon as that end is reached. It is not a public chore, to be got over with. It is a way of life."*
– Plutarch
  - *"Eternal vigilance is the price of liberty."*
– Wendell Phillips
  - *"The unleashed power of the atom has changed everything except our modes of thinking, and we thus drift toward unparalleled catastrophe." – Albert Einstein*
  - *"Technology is neither good nor bad; nor is it neutral."*
– Melvin Kranzberg
  - *"Don't ask if artificial intelligence is good or fair, ask how it shifts power."*
– Pratyusha Kalluri
  - *"Deliberation should be the goal of AI Safety, not just the procedure by which it is ensured."*
– Roel Dobbe, Thomas Gilbert, and Yonatan Minz
  - *Nothing about this form of AI coming to the fore or even existing at all was inevitable; it was the culmination of thousands of subjective choices, made by people who had the power to be in the decision-making room. In the same way, future generations of AI technologies are not predetermined... Who will get to shape them?*
- Karen Hao