Stephen Casper

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

🔗 Twitter/X 🔗 BlueSky 🔗 Google Scholar

September 4, 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. I work on rigorous AI risk assessment, risk management, and technical 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

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, culminating in research introducing state-of-the-art tamper-resistant LLM safeguards.

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

Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback	*Montreal AI Ethcis Institute,* *Sept 2023*
Meet the gig workers making AI models smarter	*The Globe and Mail,* *Sept 2023*
OpenAI Builds AI to Critique AI	*IEEE Spectrum,* *Jun 2024*
Defending against Persistent Harmful Behaviors in LLMs with Latent Adversarial Training	*CHAI Workshop* *June 2024*
Presenting the AI Risk Repository	*MIT FutureTech,* *Aug 2024*
New Analysis of AI Agents Highlights a Serious Lack of Safety Oversight	*Center for AI Policy,* *Feb 2025*
Researchers Develop an AI Agent Index to Inform Governance of Agentic Systems	*Tech Policy Press,* *Mar 2025*
MIT study finds that AI doesn't, in fact, have values	*TechCrunch,* *April 2025*
Managing Risks from Increasingly Capable Open-Weight AI Systems	*AISI Blog,* *Aug 2025*
Smokescreen: How Bad Evidence Is Used to Prevent AI Safety	*AI Frontiers,* *April 2025*
AI isn't ready to do your job	*Business Insider,* *April 2025*
Anthropic is launching a new program to study AI 'model welfare'	*TechCrunch,* *April 2025*
Study finds filtered data stops openly-available AI models from performing dangerous tasks	*University of Oxford Press,* *Aug 2025*
AI systems 'ignorant' of sensitive data can be safer, but still smart	*Washington Post,* *Aug 2025*
Filtered data stops openly-available AI models from performing dangerous tasks, study finds	*Tech Xplore,* *Aug 2025*
Oxford University and partners build tamper-resistant safeguards into open-source AI models	*EdTech Innovation Hub,* *Aug 2025*
Filtered data prevent publicly accessible AI models from performing dangerous tasks, according to a study	*Actu.ai,* *Aug 2025*
AI models taught 'deep ignorance' resist bioweapon training	*Perplexity,* *Aug 2025*

Research Mentorship Experience

Past mentees of mine have gone on to research and leadership positions at OpenAI, Epoch, Eleuther, Anthropic, the UK AI Security Institute, and numerous PhD programs. One mentee founded Harmony Intelligence.

- **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
- **Carl Guo:** Measuring the success of diffusion models at imitating human artists
- **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

Invited Talks, Panels, & Podcasts

Interpretability for Engineers with Stephen Casper	*AXRP Podcast, May 2023*
Realistic Red-Teaming in Large Langauge Models	*CHAI Workshop, Jun 2023*
Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback	*University of Alberta AI Seminar Series, Sept 2023*
Lessons from RLHF on the Difficulties of Aligning Advanced AI	*EAG Boston, Oct 2023*
Why do LLM Outputs Disagree with Internal Representations of Truthfulness?	*FAR New Orleans Alignment Workshop, Dec 2023*
Generalized Adversarial Training and Testing	*FAR Vienna Alignment Workshop, July 2024*
Stephen Casper on Technical and Sociotechnical AI Safety Research	*CAIP Podcast, Aug 2024*
Powering Up AI Capability Evaluations	*FAR Bay Area Alignment Workshop,*

Oct 2024

Pitfalls of Evidence-Based AI Policy — *EAG Boston, Oct 2024*

AI Model Piracy, Virtual Panel — *Americans for Responsible Innovation, February 2025*

MIT on The AI Agent Index — *Human Feedback Paper Group, April 2025*

Open Problems and Fundamental Limitations of RLHF — *ICLR 2025 Workshop on Human-AI Coevolution, April 2025*

LLM Tamper Resistance as a Key Priority for AI Safety — *CHAI Workshop June 2025*

Research

See my Google Scholar page
H-Index: 22
I10-Index: 35
Citations: 2804

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.
- **Casper, S.**, Bailey, L., Hunter, R., Ezell, C., Cabalé, E., Gerovitch, M., … & Kolt, N. (2025). The AI Agent Index. arXiv preprint arXiv:2502.01635.
  - Website: https://aiagentindex.mit.edu/
- 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.
- **Casper, S.**, Bailey, L., & Schreier, T. (2025). Practical Principles for AI Cost and Compute Accounting. arXiv preprint arXiv:2502.15873.
- Staufer, L., Yang, M., Reuel, A., & **Casper, S.** (2025). Audit Cards: Contextualizing AI Evaluations. arXiv preprint arXiv:2504.13839.
- Caputo, N. A., Campos, S., **Casper, S.**, Gealy, J., Hung, B., Jacobs, J., Kossack, D., Lorente, T., Murray, M., Ó hÉigeartaigh, S., Oueslati, A., Papadatos, H., Schuett, J., Wisakanto, A. K., & Trager, R. (2025, June 16). Risk tiers: Towards a gold standard for advanced AI. Oxford Martin AI Governance Initiative.
- Short, C., & **Casper, S.** (2025). Public Perspectives on AI Governance: A Survey of Working Adults in California, Illinois, and New York. Zenodo. https://doi.org/10.5281/zenodo.16566059

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. arXiv preprint arXiv:2209.02167
- **Casper, S.**, Li, Y., Li, J., Bu, T., Zhang, K., Hariharan, K., Hadfield-Menell, D., (2023). Red Teaming Deep Neural Networks with Feature Synthesis Tools NeurIPS, 2023.
- **Casper, S.**, Lin, J., Kwon, J., Culp, G., & Hadfield-Menell, D. (2023). Explore, Establish, Exploit: Red Teaming Language Models from Scratch. arXiv preprint arXiv:2306.09442.
- Shah, R.*, Feuillade–Montixi, Q.*, Pour, S.*, Tagade, A.*, **Casper, S.**, Rando, J. (2023) Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation. arXiv preprint: arXiv:2311.03348

- **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.
- Lynch, A.*, Guo, P.*, Ewart, A.*, **Casper, S.**†, Hadfield-Menell, D.† (2024) Eight Methods to Evaluate Robust Unlearning in LLMs. arXiv preprint: ariXiv:2402.16835
- **Casper, S.**, Yun, J., Baek, J., Jung, Y., Kim, M., Kwon, K., ... & Hadfield-Menell, D. (2024). The SaTML'24 CNN Interpretability Competition: New Innovations for Concept-Level Interpretability. arXiv preprint arXiv:2404.02949.
- Kirch, N. M., Field, S., & **Casper, S.** (2024). What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks. arXiv preprint arXiv:2411.03343.
- Bailey, L., Serrano, A., Sheshadri, A., Seleznyov, M., Taylor, J., Jenner, E., Hilton, J., **Casper, S.**, Guestrin, C., & Emmons, S. (2024). Obfuscated Activations Bypass LLM Latent-Space Defenses. arXiv preprint arXiv:2412.09565.
- 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.
- McKenzie, I. R., Hollinsworth, O. J., Tseng, T., Davies, X., **Casper, S.**, Tucker, A. D., ... & Gleave, A. (2025). STACK: Adversarial Attacks on LLM Safeguard Pipelines. arXiv preprint arXiv:2506.24068.

AI Risk Management

- **Casper, S.***, Nadeau, M.*, Hadfield-Menell, D., Kreiman, G (2021). Robust Feature-Level Adversaries are Interpretability Tools. In NeurIPS, 2022.
- **Casper, S.***, Hariharan, K.*, Hadfield-Menell, D., (2022). Diagnostics for Deep Neural Networks with Automated Copy/Paste Attacks.
- Liu, S., Yao, Y., Jia, J., **Casper, S.**, Baracaldo, N., Hase, P., Xu, X., Yao, Y., Li, H., Varshney, K.R., Bansal, M., Koyejo, S., Liu, Y. (2024) Rethinking Machine Unlearning for Large Language Models. arXiv preprint: ariXiv:2402.08787
- **Casper, S.***, Schulze, L.*, Patel, O., Hadfield-Menell, D. (2024) Defending Against Unforeseen Failure Modes with Latent Adversarial Training arXiv preprint: ariXiv:2403.05030
- 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. arXiv preprint arXiv:2407.15549.
- Schwinn, L., Scholten, Y., Wollschläger, T., Xhonneux, S., **Casper, S.**, Günnemann, S., & Gidel, G. (2025). Adversarial Alignment for LLMs Requires Simpler, Reproducible, and More Measurable Objectives. arXiv preprint arXiv:2502.11910.
- 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. arXiv preprint arXiv:2508.06601.

Systemization of Knowledge

- Rauker, T.*, Ho, A.*, **Casper, S.***, & Hadfield-Menell, D. (2022). Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. SATML 2023.
- **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
- Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., **Casper, S.**, Sourbut, O., Edelman, B. L., Zhang, Z., Günther, M., Korinek, A., Hernandez-Orallo, J., Hammond, L., Bigelow, E., Pan, A., Langosco, L., Korbak, T., Zhang, H., Zhong, R., Ó hÉigeartaigh, S., Recchia, G., Corsi, G., Chan, A., Anderljung, M., Edwards, L., Bengio, Y., Chen, D., Albanie, S., Maharaj, T., Foerster, J., Tramer, F., He, H., Kasirzadeh, A., Choi, Y., Krueger, D. (2024). Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv preprint arXiv:2404.09932.
- Bengio, Y., Minderman, S., Privitera, D., Besiroglu, T., **Casper, S.**, Choi, Y., Goldfarb, D., Heidari, H.,

Khalatbari, L., Longpre, S., Mavroudis, V., Mazeika, M., Yee Ng, K., Okolo, C., Raji, D., Skeadas, T., Tramer, F. (2024) International Scientific Report on the Safety of Advanced AI – Interim Report

- Reuel, A., Bucknall, B., **Casper, S.**, Fist, T., Soder, L., Aarne, O., … & Trager, R. (2024). Open Problems in Technical AI Governance. arXiv preprint arXiv:2407.14981.
- Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., **Casper, S.**, & Thompson, N. (2024). The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. preprint.
- Peppin, A., Reuel, A., **Casper, S.**, Jones, E., Strait, A., Anwar, U., … & Hooker, S. (2024). The Reality of AI and Biorisk. Proceedings of the 2025 ACM conference on fairness, accountability, and transparency. 2025.
- Barez, F., Fu, T., Prabhu, A., **Casper, S.**, Sanyal, A., Bibi, A., … & Gal, Y. (2025). Open Problems in Machine Unlearning for AI Safety. arXiv preprint arXiv:2501.04952.
- Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., … **Casper, S.** … & McGrath, T. (2025). Open Problems in Mechanistic Interpretability. arXiv preprint arXiv:2501.16496.
- Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., **Casper, S.**, … & Zeng, Y. (2025). International AI Safety Report. arXiv preprint arXiv:2501.17805.
- 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. arXiv preprint arXiv:2506.20702.

Miscellaneous

- Saleh, A., Deutsch, T., **Casper, S.**, Belinkov, Y., & Shieber, S. M. (2020). Probing Neural Dialog Models for Conversational Understanding. In Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI (pp. 132-143).
- **Casper, S.** (2020). Achilles Heels for AGI/ASI via Decision Theoretic Adversaries. arXiv
- Filan, D.*, **Casper, S.***, Hod, S.*, Wild, C., Critch, A., & Russell, S. (2021). Clusterability in Neural Networks. arXiv
- **Casper, S.***, Boix, X.*, D'Amario, V., Guo, L., Schrimpf, M., Vinken, K., & Kreiman, G. (2021). Frivolous Units: Wider Networks Are Not Really That Wide. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol 35,)
- Hod, S.*, **Casper, S.***, Filan, D.*, Wild, C., Critch, A., & Russell, S. (2021). Detecting Modularity in Deep Neural Networks. arXiv preprint
- **Casper, S.***, Hod, S.*, Filan, D.*, Wild, C., Critch, A., & Russell, S. (2022). Graphical clusterability and local specialization in deep neural networks. In ICLR 2022 Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data.
- Liu, K.*, **Casper, S.***, Hadfield-Menell, D., Andreas., J. (2023) Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness? EMNLP, 2023.

What am I up to this fall?

- I'm working on a paper titled "Open Technical Problems in Open-Weight Model Safety"
- I'm working on a paper titled "AI, The Concentration of Power, and the Diffusion of Responsibility"
- Like last year, I'm a writer again for the International AI Safety Report.
- I am considering if I should help to spearhead the creation of a yearly conference for technical AI governance research.
- I am trying to get the unicode consortium to create two new zero-width, invisible, non-breaking characters added which mean "This document was AI generated," and "The author of this document does not consent to AI training on it."
- I am considering working to get an open-letter written and signed by many Western and Eastern academics saying that AI securitization and arms races are self-fulfilling prophecies and not in anyone's public interest.

Miscellaneous

- Anecdotal places where my work has been influential:
  - Course syllabi 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
  - (Speculative) It is possible that my paper, Practical Principles for AI Cost and Compute Accounting, may have influenced Section 22757.15 of California's Proposed 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 the EU AI office was interested in the paper, and met with them in July, 2024.
  - (Speculative) Based on private conversations with New-York State Representative Alex Bores's office, some of my work on Public Perspectives on AI Governance seems to have been helpful in discussions around the proposed RAISE Act.
- Blog post: Reframing AI Safety as a Neverending Institutional Challenge
- Blog post: Managing risks from increasingly capable open-weight AI systems
- Blog post sequence: The Engineer's Interpretability Sequence (2023)
- I was a member of the EU AI Act Codes of Practice Plenary (2024-2025)
- As an April Fools joke, I compiled everything you need.
- I have a personal anonymous feedback form.
- My favorite AI-related quotes are:
  - *"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