

Stephen Casper

✉ scasper@mit.edu stephencasper.com/

[Twitter/X](#) [BlueSky](#) [Google Scholar](#)

July 7, 2026

Hi, my name is Stephen Casper, but you can call me Cas. I am a computer scientist, an incoming assistant professor of Public Policy at the Harvard Kennedy School, and an incoming affiliate in Computer Science at the Harvard School of Engineering and Applied Sciences. Previously, I was a PhD student at MIT and a research resident at the UK AI Security Institute. I work on frontier AI risk management problems, including safeguards, incidents, and governance. A central goal of my research is to inform the policy community about best technical practices for managing AI risks.

My Work

Harvard Kennedy School <i>Assistant Professor of Public Policy</i>	<i>July 2026 – Present</i>
Harvard School of Engineering and Applied Sciences <i>Affiliate</i>	<i>July 2026 – Present</i>

Education

MIT <i>Ph.D. in Electrical Engineering and Computer Science</i> <i>Minor in Public Policy</i>	<i>Feb 2024 – May 2026</i>
MIT <i>M.S. in Electrical Engineering and Computer Science</i> GPA: 4.8/5.0	<i>Sept 2021 – Feb 2024</i>
Harvard University <i>B.A. in Statistics, Secondary field in Mathematical Sciences</i> Cum laude (Latin) and with highest honors (English). GPA: 3.847/4.0	<i>Sept 2017 – May 2021</i>

Awards

Best paper runner-up	<i>NeurIPS BioSafeGenAI Workshop</i> <i>Dec 2025</i>
Outstanding Paper Finalist	<i>TMLR</i> <i>Dec 2024</i>
Spotlight Paper	<i>ICML GenLaw Workshop</i> <i>Jul 2023</i>
Best Paper Award	<i>NeurIPS ML Safety Workshop</i> <i>Dec 2022</i>
Vitalik Buterin Fellowship	<i>Future of Life Institute</i> <i>Apr 2022</i>
Hoopes Prize Thesis Award	<i>Harvard University</i> <i>May 2021</i>

Press & Newsletter Mentions

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

- **Futurism:** [Top AI Researchers Terrified of a “Chernobyl Moment”: a Mass Casualty Event, or Worse, That Turns the World Against AI Forever](#)
- **MIT Technology Review:** [The Shock of Seeing your Body Used in Deepfake Porn](#)
- **CNET:** [AI Agents Are Getting Better. Their Safety Disclosures Aren’t](#)

- **Heise Online:** AI Agents: popularity is skyrocketing – despite lack of security
- **FindArticles:** MIT Study Warns AI Agents Are Out Of Control and MIT Index Ranks Top 30 AI Agents By Autonomy And Use
- **IBT:** Study Finds Most AI Agents Skip or Lack Safety Disclosure Raising Transparency Concerns
- **GeneOnline:** Study Finds Most AI Systems Lack Comprehensive Safety Documentation
- **Ground News:** AI Agents are Getting Better. Their Safety Disclosures Aren't
- **Mirage:** Research: Most AI Bots Omit Basic Safety Disclosures
- **Newage:** Crucial safety info missing on AI 'agents': study
- **Gizmodo:** New Research Shows AI Agents Are Running Wild Online, With Few Guardrails in Place
- **The Register:** AI agents abound, unbound by rules or safety disclosures
- **Bioengineer:** Study Reveals Most AI Bots Lack Fundamental Safety Disclosures
- **ABS-CBN:** Crucial safety info missing on AI 'agents': study
- **Tech Xplore:** Most AI bots lack basic safety disclosures, study finds
- **Cambridge Press:** Most AI bots lack basic safety disclosures, study finds
- **The Standard:** Crucial safety info missing on AI 'agents': study
- **Zdnet:** AI agents are fast, loose and out of control, MIT study finds and These top 30 AI agents deliver a mix of functions and autonomy
- **Wired:** Deepfake 'Nudify' Technology Is Getting Darker—and More Dangerous (Jan 2026)
- **ML Safety Newsletter:** Filtering Dangerous Training Data (Sept 2025)
- **AISI Blog:** Managing Risks from Increasingly Capable Open-Weight AI Systems (Aug 2025)
- **Perplexity:** AI models taught 'deep ignorance' resist bioweapon training (Aug 2025)
- **Actu.ai:** Filtered data prevent publicly accessible AI models from performing dangerous tasks, according to a study (Aug 2025)
- **EdTech Innovation Hub:** Oxford University and partners build tamper-resistant safeguards into open-source AI models (Aug 2025)
- **Tech Xplore:** Filtered data stops openly-available AI models from performing dangerous tasks, study finds (Aug 2025)
- **Washington Post:** AI systems 'ignorant' of sensitive data can be safer, but still smart (Aug 2025)
- **University of Oxford Press:** Study finds filtered data stops openly-available AI models from performing dangerous tasks (Aug 2025)
- **OECD.AI:** Strengthening global AI Safety: A perspective on the Singapore Consensus (Jun 2025)
- **WebProNews:** Singapore's AI Diplomacy: Forging a Middle Path for Global AI Governance (May 2025)
- **EuroNews:** There is a global consensus for AI safety despite Paris Summit backlash, new report finds (May 2025)
- **IMDA:** Top scientific minds gathered for the first time in Singapore to advance AI that is trustworthy, reliable and secure (May 2025)
- **Wired:** Singapore's Vision for AI Safety Bridges the US-China Divide (May 2025)
- **TechCrunch:** Anthropic is launching a new program to study AI 'model welfare' (Apr 2025)
- **Business Insider:** AI isn't ready to do your job (Apr 2025)
- **AI Frontiers:** Smokescreen: How Bad Evidence Is Used to Prevent AI Safety (Apr 2025)
- **TechCrunch:** MIT study finds that AI doesn't, in fact, have values (Apr 2025)
- **Lumenova:** What You Should Know: The AI Agent Index (Apr 2025)
- **Tech Policy Press:** Researchers Develop an AI Agent Index to Inform Governance of Agentic Systems (Mar 2025)
- **Center for AI Policy:** New Analysis of AI Agents Highlights a Serious Lack of Safety Oversight (Feb 2025)
- **MIT FutureTech:** Presenting the AI Risk Repository (Aug 2024)
- **IEEE Spectrum:** OpenAI Builds AI to Critique AI (Jun 2024)

- **The Globe and Mail:** [Meet the gig workers making AI models smarter](#) (Sept, 2023)
- **Montreal AI Ethics Institute:** [Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback](#) (Sept 2023)

Invited Talks, Panels, & Podcasts

- **Beijing Academy of AI:** [Checking in on the Agentic AI Ecosystem: Safety, Geopolitics, and Consolidation of Power](#) (Jun 2026)
- **AI Doom Debates** [This Harvard Professor Says AI Alignment Will BACKFIRE](#) (Jun 2026)
- **FAR.AI CHAI Workshop:** [Non-Consensual AI Deepfakes: AI Safety's Trial by Fire](#) (Jun 2026)
- **FAR.AI Technical Innovations for AI Policy Workshop:** [Non-Consensual AI Deepfakes: AI Safety's Trial by Fire](#) (Mar 2026)
- **FAR.AI London Alignment Workshop:** [ML Researchers as Policymakers](#) (Mar 2026)
- **The AI Policy Podcast:** [Inside The Second Int'l AI Safety Report with Stephen Clare & Stephen Casper](#) (Feb 2026)
- **FAR San Diego Alignment Workshop:** [Powerful Open-Weight Models: Wonderful, Terrible, Inevitable. How can we make them Safer?](#) (Dec 2025)
- **ICML 2025 Technical AI Governance Workshop:** [Panel: Technical AI Governance as a Field \(Moderator\)](#) (Jul 2025)
- **ICML 2025 2nd Workshop on Models of Human Feedback for AI Alignment (MoFA):** [Panel](#) (Jul 2025)
- **CHAI Workshop:** [LLM Tamper Resistance as a Key Priority for AI Safety](#) (Jun 2025)
- **ICLR 2025 Workshop on Human-AI Coevolution:** [Open Problems and Fundamental Limitations of RLHF](#) (Apr 2025)
- **Human Feedback Paper Group:** [MIT on The AI Agent Index](#) (Apr 2025)
- **Americans for Responsible Innovation:** [AI Model Piracy, Virtual Panel](#) (Feb 2025)
- **EAG Boston 2024:** [Pitfalls of Evidence-Based AI Policy](#) (Oct 2024)
- **FAR Bay Area Alignment Workshop:** [Powering Up AI Capability Evaluations](#) (Oct 2024)
- **CAIP Podcast:** [Stephen Casper on Technical and Sociotechnical AI Safety Research](#) (Aug 2024)
- **FAR Vienna Alignment Workshop:** [Generalized Adversarial Training and Testing](#) (Jul 2024)
- **CHAI Workshop:** [Defending against Persistent Harmful Behaviors in LLMs with Latent Adversarial Training](#) (Jun 2024)
- **FAR New Orleans Alignment Workshop:** [Why do LLM Outputs Disagree with Internal Representations of Truthfulness?](#) (Dec 2023)
- **EAG Boston 2023:** [Lessons from RLHF on the Difficulties of Aligning Advanced AI](#) (Oct 2023)
- **University of Alberta AI Seminar Series:** [Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback](#) (Sept 2023)
- **CHAI Workshop:** [Realistic Red-Teaming in Large Language Models](#) (Jun 2023)
- **AXRP Podcast:** [Interpretability for Engineers with Stephen Casper](#) (May 2023)

Professional Experience

Fellow *Harvard Berkman Klein Center* *March 2026 - August 2026*

- Research on AI governance and risk management.

Mentor *GovAI Fellows Program* *Sept 2025 - January 2026*

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

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

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

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

Mentor *ERA Fellowship*

May 2024 - Present

- Mentoring two successful research projects on LLM jailbreaking, LLM unlearning, and LLM tamper resistance, both resulting in papers.

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.

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

Jun 2021 - Sept 2021

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

Service & Field Building

- I am beginning the organization process for an annual, archival academic technical AI governance research (TAIGR) conference, which will begin in 2027.
- I am a workshop organizer for the submitted [ICML 2026 Technical AI Governance Workshop](#), the second iteration of this workshop which has become the world's only venue specializing in technical AI governance challenges (pending, Apr 2026)
- I am an area chair for the NeurIPS 2026 Position Paper Track.
- I am an area chair for the ICLR 2026 Blog Post Track.
- I was a workshop organizer for the [Post-AGI Futures workshop](#), which focuses on the economics, culture, and governance possibilities for a future with transformative AI technology (Dec 2025).
- I was a member of the [EU AI Act Codes of Practice Plenary](#) (Nov 2024-Apr 2025)
- I organized the [SaTML CNN Interpretability Competition](#) which contributed to the development of state-of-the-art feature synthesis techniques for practical debugging in vision models (Apr 2024).

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 multiple PhD programs. Several went on to found [Harmony Intelligence](#), [Fulcrum](#), [Watertight AI](#), [Athena](#), and [Geodesic](#).

Here I list mentees for 17 projects where I was the joint-first or last author and the mentee(s) were the first or joint-first. The H-index of these 17 papers is 13.

- **Joe Kwon:** [Internal Deployment Gaps in AI Regulation](#)
- **Leon Staufer:** [The 2025 AI Agent Index](#)
- **Max Kamachee:** [Video Deepfake Abuse: How Company Choices Predictably Shape Misuse Patterns](#)
- **Kyle O'Brien:** [Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs](#)
- **Leon Staufer, Mick Yang:** [Audit Cards: Contextualizing AI Evaluations](#)
- **Ariba Khan:** [Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs](#)
- **Zora Che:** [Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities](#)
- **Nathalie Kirch, Severin Field:** [What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks](#)
- **Phillip Guo, Aengus Lynch, Aidan Ewart, Cindy Wu, Abhay Sheshadri:** [Latent adversarial training improves robustness to persistent harmful behaviors in llms](#)
- **Soroush Pour, Rusheb Shah, Quentin Feuilade-Montixi, Arush Tagade:** [Scalable and transferable black-box jailbreaks for language models via persona modulation](#)
- **Carson Ezell:** [Black-box access is insufficient for rigorous ai audits](#)
- **Lennart Schulze:** [Defending against unforeseen failure modes with latent adversarial training](#)
- **Xander Davies:** [Open problems and fundamental limitations of reinforcement learning from human feed-](#)

[back](#)

- **Anson Ho, Tilman R  uker:** [Toward transparent AT: A survey on interpreting the inner structures of deep neural networks](#)
- **Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang:** [Red teaming deep neural networks with feature synthesis tools](#)
- **Gatlen Culp, Jason Lin, Joe Kwon:** [Explore, establish, exploit: Red teaming language models from scratch](#)
- **Kaivalya Hariharan:** [Diagnostics for deep neural networks with automated copy/paste attacks](#)
- **Max Nadeau:** [Robust feature-level adversaries are interpretability tools](#)

My Favorite Papers I have Worked on

Google Scholar: See my [Google Scholar page](#) for a complete list of papers.

H-Index: 26

I10-Index: 38

Citations: 4216

Technical AI Safeguards

- Kamachee, M.*, **Casper, S.***, Ding, M. L., Yew, R. J., Reuel, A., Biderman, S., & Hadfield-Menell, D. (2025). [Video Deepfake Abuse: How Company Choices Predictably Shape Misuse Patterns](#). arXiv preprint arXiv:2512.11815.
- **Casper, S.**, O'Brien, K., Longpre, S., Seger, E., Klyman, K., Bommasani, R., Nrusimha, A., Shumailov, I., Mindermann, S., Basart, S., Rudzicz, F., Pelrine, K., Ghosh, A., Strait, A., Kirk, R., Hendrycks, D., Henderson, P., Kolter, Z., Irving, G., Gal, Y., Bengio, Y., & Hadfield-Menell, D. (2025). [Open technical problems in open-weight AI model risk management](#).
- 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](#).
 - [Best paper runner-up, NeurIPS 2025 Workshop on Biosecurity Safeguards for Generative AI](#)
- 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.
- **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](#))

AI Evals, Red-Teaming, & Risk Assessment

- 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.
- **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.
- **Casper, S.**, Hadfield-Menell, D., Kreiman, G. (2022). [Red-Teaming with Mind-Reading: White-Box Adversarial Policies in Deep Reinforcement Learning](#).
 - [Hoopes Prize](#) ([link](#))

AI, Governance, and Society

- **Casper, S.**, Krueger, D., & Hadfield-Menell, D. (2025). [Pitfalls of Evidence-Based AI Policy](#). ICLR 2025 Blog post.
- 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.**, Bailey, L., Hunter, R., Ezell, C., Cabal  , E., Gerovitch, M., ... & Kolt, N. (2025). [The AI Agent Index](#). [Website](#).

- **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](#))

Systemization of Knowledge

- Bengio, Y., Maharaaj, 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](#).
- Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., **Casper, S.**, ... & Zeng, Y. (2025). [International AI Safety Report](#).
- Reuel, A., Bucknall, B., **Casper, S.**, Fist, T., Soder, L., Aarne, O., ... & Trager, R. (2024). [Open Problems in Technical AI Governance](#). TMLR Survey Certification
- **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., Krashennnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, 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](#))

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](#) which cites 13 of my papers.
 - [The California Report on Frontier AI Policy](#) which cites [some of my work](#) as its basis for discussing “evidence-generating policy.”
 - [Litigation](#) in the US District Court, Northern District of California San Francisco Division which cites [some of my work](#) in discussing claims of trade dress infringement.
 - New-York State Representative Alex Bores’s office communicated to me that [a paper of mine](#) was helpful in discussions around the proposed RAISE Act.
- Blog posts:
 - [Reframing AI Safety as a Neverending Institutional Challenge](#) (2025)
 - [Managing risks from increasingly capable open-weight AI systems](#) (2025)
 - [The Engineer’s Interpretability Sequence](#) (2023)
- 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
 - *“Technological change is neither additive nor subtractive. It is ecological...One significant change generates total change.”*
– Neil Postman
 - *“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