

Alexander Robey

arobey@andrew.cmu.edu
arobey1.github.io

Education

Ph.D., Electrical and Systems Engineering

Advisors: Hamed Hassani and George J. Pappas
Thesis: *Algorithms for Adversarially Robust Deep Learning*

University of Pennsylvania
2018–2024

B.A., Mathematics; B.S., Engineering (High Honors)

Advisor: Vidya Ganapati
Thesis: *A Deep Learning Approach to Fourier Ptychographic Microscopy*

Swarthmore College
2014–2018

Work experience

Member of the technical staff, Thinking Machines Lab

2026–

Postdoctoral fellow, Carnegie Mellon University

2024–2025

Research contractor, Gray Swan AI

2024–2025

Visiting instructor, Swarthmore College

2024

Research intern, Google Cloud AI

2022

Research intern, Lawrence Livermore National Laboratory

2017

Research assistant, Swarthmore College

2016

Awards and honors

Charles Hallac and Sarah Keil Wolf Dissertation Award, Penn Engineering

2025

Rising Star in Cyber-Physical Systems, National Science Foundation

2025

Best Paper Award, Princeton Symposium on Safe Deployment of Foundation Models in Robotics

2024

Rising Star in Adversarial Machine Learning, NeurIPS '24 AdvML Workshop

2024

Best Paper Award, ICML '23 AdvML Workshop

2023

Trustworthy AI Research Fellowship, Amazon

2023

Teaching Assistant of the Year, Penn Engineering

2020

Dean's Fellowship, Penn Engineering

2018

Research Fellowship, Swarthmore College Engineering Department

2016

Best Reviewer Award, ICML {'20, '22, '24}, NeurIPS {'21, '22}, ICLR {'21, '24}, AISTATS '25

Teaching experience

Visiting instructor

ENGR 012: Linear Physical Systems Analysis

Swarthmore College

Guest lecturer

COSC 227: Neural Safety Net

Amherst College

Artificial Intelligence & Criminal Justice

University of British Columbia

CS 7180: Verifiable Machine Learning

Northeastern University

CIS 7000: Trustworthy Machine Learning

University of Pennsylvania

ENGR 056: Modeling and Optimization for Engineering

Swarthmore College

Teaching assistant

ESE 605: Modern Convex Optimization

University of Pennsylvania

ESE 290: Introduction to Research Methodologies

University of Pennsylvania

ESE 530: Elements of Probability Theory

University of Pennsylvania

ENGR 019: Numerical Methods for Engineering

Swarthmore College

ENGR 011: Electrical Circuit Analysis

Swarthmore College

ENGR 012: Linear Physical Systems Analysis

Swarthmore College

Professional activity

Area chair

NeurIPS, ICML

Conference reviewer

NeurIPS, ICML, ICLR, SaTML, AAAI, COLM, L4DC, CDC, ICCV, ECCV, ISIT, ICCPS, IASEAI

Journal reviewer

JMLR, TMLR, PAMI, TAC, SIMODS, IJCV, Nature

Workshop reviewer

Red Teaming GenAI: What Can We Learn from Adversaries?

NeurIPS 2024

ICML 2024

Theoretical Foundations of Foundation Models

NeurIPS 2023

Robustness of Few- & Zero-shot Learning in Large Foundation Models

NeurIPS 2023

Distribution Shifts: New Frontiers with Foundation Models

NeurIPS 2023

Adversarial Robustness in the Real World

ICCV 2023

Out-of-Distribution Generalization in Computer Vision

ICCV 2023

Adversarial Machine Learning Frontiers

ICML 2023

Domain Generalization

ICLR 2023

Special Track on Safe and Robust AI

AAAI 2023

Distribution Shifts

NeurIPS 2022

Robustness in Sequence Modeling

NeurIPS 2022

Out-Of-Distribution Generalization in Computer Vision

ECCV 2022

Adversarial Robustness in the Real World

ECCV 2022

Adversarial Machine Learning Frontiers

ICML 2022

Distribution Shifts: Connecting Methods and Applications

NeurIPS 2021

Adversarial Robustness in the Real World

ICCV 2021

Adversarial Robustness in the Real World

ECCV 2020

Workshop organizer

Adversarial Robustness in the Real World

ECCV 2022

Adversarial Robustness in the Real World

ICCV 2021

Conference papers

- [1] Francesco Marchiori, Rohan Sinha*, Christopher Agia*, **Alexander Robey**, George J. Pappas, Mauro Conti, and Marco Pavone. Preventing Robotic Jailbreaking via Multimodal Domain Adaptation. *ICRA*, 2026.
- [2] Barry Wang, Avi Schwarzschild, **Alexander Robey**, Ali Payani, Charles Fleming, Mingjie Sun, and Daphne Ippolito. Command-V: Pasting LLM Behaviors via Activation Profiles. *ICLR*, 2025.
- [3] Jared Perlo, **Alexander Robey**, Fazl Barez, Luciano Floridi, and Jakob Mökander. Embodied AI: Emerging Risks and Opportunities for Policy Action. *NeurIPS*, 2025.
- [4] Pratyush Maini*, Sachin Goyal*, Dylan Sam*, **Alexander Robey**, Yash Savani, Yiding Jiang, Andy Zou, Zachary C. Lipton, and Zico J. Kolter. Safety Pretraining: Toward the Next Generation of Safe AI. *NeurIPS*, 2025.
- [5] Yash Savani*, Asher Trockman*, Zhili Feng, Avi Schwarzschild, **Alexander Robey**, Marc Finzi, and J. Zico Kolter. Antidistillation Sampling. *NeurIPS*, 2025.
- [6] Jiabao Ji*, Bairu Hou*, **Alexander Robey***, George J. Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. Defending Large Language Models against Jailbreaking Attacks via Semantic Smoothing. *IJCNLP-AACL*, 2025.

- [7] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas. Jail-breaking LLM-Controlled Robots. *ICRA*, 2025.
- [8] Patrick Chao, Alexander Robey, Eric Wong, Hamed Hassani, George J. Pappas, and Edgar Dobriban. Jailbreaking Black Box Large Language Models in Twenty Questions. *IEEE Conference on Secure and Trustworthy Machine Learning*, 2025.
- [9] Patrick Chao*, Edoardo Debenedetti*, Alexander Robey*, Maksym Andriushchenko*, Vikash Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *NeurIPS*, 2024.
- [10] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyang Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson. A Safe Harbor for AI Evaluation and Red Teaming. *ICML*, 2024.
- [11] Alexander Robey*, Fabian Latorre*, George J. Pappas, Hamed Hassani, and Volkan Cevher. Adversarial training should be cast as a non-zero-sum game. *ICLR*, 2023.
- [12] Haoze Wu*, Tagomori Teruhiro*, Alexander Robey*, Fengjun Yang*, Nikolai Matni, George J. Pappas, Corina Pasareanu, and Clark Barrett. Toward Certified Robustness Against Real-World Distribution Shifts. *IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.
- [13] Cian Eastwood*, Alexander Robey*, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. Probable Domain Generalization via Quantile Risk Minimization. *NeurIPS*, 2022.
- [14] Anton Xue, Lars Lindemann, Alexander Robey, Hamed Hassani, George J. Pappas, and Rajeev Alur. Chordal Sparsity for Lipschitz Constant Estimation of Deep Neural Networks. *IEEE Conference on Decision and Control*, 2022.
- [15] Alexander Robey, Luiz Chamon, George J. Pappas, and Hamed Hassani. Probabilistically Robust Learning: Balancing Average and Worst-case Performance. *ICML*, 2022.
- [16] Allan Zhou*, Fahim Tajwar*, Alexander Robey, Tom Knowles, George J. Pappas, Hamed Hassani, and Chelsea Finn. Do Deep Networks Transfer Invariances Across Classes? *ICLR*, 2022.
- [17] Alexander Robey*, Luiz F. O. Chamon*, George J. Pappas, Hamed Hassani, and Alejandro Ribeiro. Adversarial Robustness with Semi-Infinite Constrained Learning. In *NeurIPS*, 2021.
- [18] Alexander Robey, George J. Pappas, and Hamed Hassani. Model-Based Domain Generalization. *NeurIPS*, 2021.
- [19] Stephen Tu, Alexander Robey, Tingnan Zhang, and Nikolai Matni. On the Sample Complexity of Stability Constrained Imitation Learning. *Learning for Dynamics and Control*, 2022.
- [20] Alexander Robey, Lars Lindemann, Stephen Tu, and Nikolai Matni. Learning Robust Hybrid Control Barrier Functions for Uncertain Systems. *IFAC Conference on Analysis and Design of Hybrid Systems*, 2021.
- [21] Alexander Robey, Arman Adibi, Brent Schlotfeldt, George J. Pappas, and Hamed Hassani. Optimal Algorithms for Submodular Maximization with Distributed Constraints. *Learning for Dynamics and Control*, 2021.
- [22] Lars Lindemann, Haimin Hu, Alexander Robey, Hanwen Zhang, Dimos V Dimarogonas, Stephen Tu, and Nikolai Matni. Learning Hybrid Control Barrier Functions from Data. *Conference on Robot Learning*, 2021.
- [23] Alexander Robey*, Haimin Hu*, Lars Lindemann, Hanwen Zhang, Dimos V. Dimarogonas, Stephen Tu, and Nikolai Matni. Learning Control Barrier Functions from Expert Demonstrations. *IEEE Conference on Decision and Control*, 2020.

- [24] Mahyar Fazlyab, **Alexander Robey**, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks. *NeurIPS*, 2019.

Journal articles

- [1] Zachary Ravichandran, **Alexander Robey**, Vijay Kumar, George J. Pappas, and Hamed Hassani. Safety Guardrails for LLM-Enabled Robots. *IEEE RAL*, 2026.
- [2] Yutong He, **Alexander Robey**, Naoki Murata, Yiding Jiang, Joshua Williams, George J. Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J. Zico Kolter. Automated Black-box Prompt Engineering for Personalized Text-to-Image Generation. *TMLR*, 2025.
- [3] **Alexander Robey**, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *TMLR*, 2025.
- [4] Lars Lindemann, **Alexander Robey**, Lejun Jiang, Stephen Tu, and Nikolai Matni. Learning Robust Output Control Barrier Functions from Safe Expert Demonstrations. *IEEE Open Journal of Control Systems*, 2024.
- [5] Edgar Dobriban, Hamed Hassani, David Hong, and **Alexander Robey**. Provable Tradeoffs in Adversarially Robust Classification. *IEEE Transactions on Information Theory*, 2022.
- [6] **Alexander Robey** and Vidya Ganapati. Optimal Physical Preprocessing for Example-based Super Resolution. *Optics Express*, 2018.

Workshop papers

- [1] Davis Brown*, Mahdi Sabbaghi*, Luze Sun, **Alexander Robey**, George J. Pappas, Eric Wong, and Hamed Hassani. Benchmarking misuse mitigation against covert adversaries. *NeurIPS Workshop on Biosecurity Safeguards for Generative AI*, 2025.
- [2] Eliot Krzysztof Jones, **Alexander Robey**, Andy Zou, Zachary Ravichandran, George J. Pappas, Hamed Hassani, Matt Fredrikson, and J. Zico Kolter. Adversarial Attacks on Robotic Vision Language Action Models. *RSS 2025 Workshop on Reliable Robotics: Safety and Security in the Face of Generative AI*, 2025.
- [3] Zhili Feng, Yixuan Even Xu, Pratyush Maini, **Alexander Robey**, Avi Schwarzschild, and J. Zico Kolter. Evaluating LLM Memorization Using Soft Token Sparsity. *ICLR Workshop on Sparsity in LLMs*, 2025.
- [4] Rishi Rajesh Shah, Chen Henry Wu, Ziqian Zhong, **Alexander Robey**, and Aditi Raghunathan. Jailbreaking in the Haystack. *ICML Workshop on Long-Context Foundation Models*, 2025.

Preprints

- [1] Dylan Sam, Sachin Goyal, Pratyush Maini, **Alexander Robey**, and J. Zico Kolter. When Should We Introduce Safety Interventions During Pretraining? *arXiv*, 2026.
- [2] Sarah Ball*, Niki Hasrati*, **Alexander Robey**, Avi Schwarzschild, Frauke Kreuter, J. Zico Kolter, and Andrej Risteski. Toward Understanding the Transferability of Adversarial Suffixes in Large Language Models. *arXiv*, 2025.
- [3] Dylan Sam, **Alexander Robey**, Andy Zou, Matt Fredrikson, and J. Zico Kolter. Evaluating Language Model Reasoning about Confidential Information. *arXiv*, 2025.
- [4] Zhili Feng*, Yixuan Even Xu*, **Alexander Robey**, Robert Kirk, Xander Davies, Yarin Gal, Avi Schwarzschild, and J. Zico Kolter. Existing Large Language Model Unlearning Evaluations Are Inconclusive. *arXiv*, 2025.
- [5] Kai Hu, Weichen Yu, Li Zhang, **Alexander Robey**, Andy Zou, Chengming Xu, Haoqi Hu, and Matt Fredrikson. Transferable Adversarial Attacks on Black-Box Vision-Language Models. *arXiv*, 2025.

- [6] Hanjiang Hu, **Alexander Robey**, and Changliu Liu. Steering Dialogue Dynamics for Robustness against Multi-turn Jailbreaking Attacks. *arXiv*, 2025.
- [7] Thomas Waite, **Alexander Robey**, Hassani Hamed, George J. Pappas, and Radoslav Ivanov. Data-Driven Modeling and Verification of Perception-Based Autonomous Systems. *arXiv*, 2023.
- [8] **Alexander Robey**, Hamed Hassani, and George J. Pappas. Model-Based Robust Deep Learning. *arXiv*, 2020.

Patents

- [1] **Alexander Robey**, Hamed Hassani, and George J Pappas. Model-Based Robust Deep Learning, April 2024. U.S. Patent No. 11,961,283.