Alexander Robey

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

Education

Ph.D., Electrical and Systems Engineering — University of Pennsylvania
Advisors: Hamed Hassani and George J. Pappas — August 2018 – August 2024
Thesis: *Algorithms for Adversarially Robust Deep Learning*

B.A., Mathematics, B.S., Engineering — Swarthmore College
Advisor: Vidya Ganapati — August 2014 – May 2018
Thesis: *A Deep Learning Approach to Fourier Ptychographic Microscopy*

Work experience

Postdoctoral Researcher — Carnegie Mellon University
Advisor: J. Zico Kolter — September 2024–

Research consultant — Gray Swan
Contact: Matt Fredrikson — September 2024–

Visiting Instructor — Swarthmore College
Contact: Matt Zucker — January 2024 – May 2024

Student Researcher — Google Cloud AI
Advisors: Sayna Ebrahimi and Sercan Ö. Arik — June 2022 – February 2023

Research Assistant — Swarthmore College
Advisor: Vidya Ganapati — May 2018 – August 2018

Research Intern — Lawrence Livermore National Laboratory
Advisors: Abhinav Bhatele and Nikhil Jain — May 2017 – August 2017

Research Assistant — Swarthmore College
Advisor: Carr Everbach — May 2016 – August 2016

Awards and honors

Charles Hallac and Sarah Keil Wolf Dissertation Award — University of Pennsylvania
Department of Electrical and Systems Engineering — 2025

Rising Star in Cyber-Physical Systems — NSF
2025 NSF CPS Rising Stars Workshop — 2025

Best Poster Award — Princeton University
Symposium on Safe Deployment of Foundation Models in Robotics — 2024

Rising Star in Adversarial Machine Learning — NeurIPS
3rd Workshop on New Frontiers in Adversarial Machine Learning — 2024

Best Paper Award — ICML
2nd Workshop on New Frontiers in Adversarial Machine Learning — 2023

Research Fellowship — Amazon AWS
ASSET Center for AI-Enabled Systems — 2023

Teaching Assistant of the Year — University of Pennsylvania
Department of Electrical and Systems Engineering — 2020

Dean's Fellowship — University of Pennsylvania
Department of Electrical and Systems Engineering — 2018

Research Fellowship — Swarthmore College
Department of Engineering — 2016

Teaching experience

Visiting instructor

ENGR 012: Linear Physical Systems Analysis — Swarthmore College

Guest lecturer

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
ENGR 006: Engineering Mechanics — Swarthmore College

Professional activity

Reviewing

Conferences: NeurIPS, ICML, ICLR, SaTML, AAAI, COLM, L4DC, CDC, ICCV, ECCV, ISIT
Journals: JMLR, TMLR, PAMI, TAC, SIMODS, IJCV, Nature ML
Workshops and special tracks:
Red Teaming GenAI: What Can We Learn from Adversaries? (NeurIPS 2024)
Theoretical Foundations of Foundation Models (ICML 2024)
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)
Safe and Robust AI special track (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)

Organizing

Adversarial Robustness in the Real World (ECCV 2022)
Adversarial Robustness in the Real World (ICCV 2021)

Conference papers

[1] **Alexander Robey**, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas. Jailbreaking LLM-Controlled Robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

[2] Patrick Chao, **Alexander Robey**, Eric Wong, Hamed Hassani, George J. Pappas, and Edgar Dobriban. Jailbreaking Black Box Large Language Models in Twenty Questions. In *IEEE Conference on Secure and Trustworthy Machine Learning*. IEEE, 2025.

[3] 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. In *Advances in Neural Information Processing Systems*, 2024.

[4] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan 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. In *International Conference on Machine Learning*. PMLR, 2024.

[5] **Alexander Robey**⋆, Fabian Latorre⋆, George J. Pappas, Hamed Hassani, and Volkan Cevher. Adversarial Training Should Be Cast as a Non-Zero-Sum Game. In *International Conference on Learning Representations*, 2024.

[6] Haoze Wu⋆, Teruhiro Tagomori⋆, **Alexander Robey**⋆, Fengjun Yang⋆, Nikolai Matni, George J. Pappas, Hamed Hassani, Corina Pasareanu, and Clark Barrett. Toward Certified Robustness Against Real-World Distribution Shifts. In *IEEE Conference on Secure and Trustworthy Machine Learning*. IEEE, 2023.

[7] Cian Eastwood⋆, **Alexander Robey**⋆, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. Probable Domain Generalization via Quantile Risk Minimization. In *Advances in Neural Information Processing Systems*, 2022.

[8] Anton Xue, Lars Lindemann, **Alexander Robey**, Hamed Hassani, George J. Pappas, and Rajeev Alur. Chordal Sparsity for Lipschitz Constant Estimation of Deep Neural Networks. In *2022 61st IEEE Conference on Decision and Control (CDC)*. IEEE, 2022.

[9] **Alexander Robey**, Luiz F. O. Chamon, George J. Pappas, and Hamed Hassani. Probabilistically Robust Learning: Balancing Average-and Worst-case Performance. In *International Conference on Machine Learning*. PMLR, 2022.

[10] Allan Zhou⋆, Fahim Tajwar⋆, **Alexander Robey**, Tom Knowles, George J. Pappas, Hamed Hassani, and Chelsea Finn. Do Deep Networks Transfer Invariances across Classes? In *International Conference on Learning Representations*, 2022.

[11] **Alexander Robey**⋆, Luiz F. O. Chamon⋆, George J. Pappas, Hamed Hassani, and Alejandro Ribeiro. Adversarial Robustness with Semi-Infinite Constrained Learning. In *Advances in Neural Information Processing Systems*, 2021.

[12] **Alexander Robey**, George J. Pappas, and Hamed Hassani. Model-Based Domain Generalization. In *Advances in Neural Information Processing Systems*, 2021.

[13] Stephen Tu, **Alexander Robey**, Tingnan Zhang, and Nikolai Matni. On the Sample Complexity of Stability Constrained Imitation Learning. In *Learning for Dynamics and Control*. PMLR, 2022.

[14] **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.

[15] **Alexander Robey**, Arman Adibi, Brent Schlotfeldt, George J. Pappas, and Hamed Hassani. Optimal Algorithms for Submodular Maximization with Distributed Constraints. In *Learning for Dynamics and Control*. PMLR, 2021.

[16] 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.

[17] **Alexander Robey**⋆, Haimin Hu⋆, Lars Lindemann, Hanwen Zhang, Dimos V Dimarogonas, Stephen Tu, and Nikolai Matni. Learning Control Barrier Functions from Expert Demonstrations. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3717–3724. IEEE, 2020.

[18] Mahyar Fazlyab, **Alexander Robey**, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, pages 11427–11438, 2019.

Journal articles

[1] 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.

[2] Edgar Dobriban, Hamed Hassani, David Hong, and **Alexander Robey**. Provable Tradeoffs in Adversarially Robust Classification. *IEEE Transactions on Information Theory*, 2022.

[3] **Alexander Robey** and Vidya Ganapati. Optimal Physical Preprocessing for Example-based Super Resolution. *Optics Express*, 26(24):31333–31350, 2018.

Preprints

[1] Zachary Ravichandran, **Alexander Robey**, Vijay Kumar, George J. Pappas, and Hamed Hassani. Safety Guardrails for LLM-Enabled Robots. *arXiv preprint arXiv:2503.07885*, 2025.

[2] Hanjiang Hu, **Alexander Robey**, and Changliu Liu. Steering Dialogue Dynamics for Robustness against Multi-turn Jailbreaking Attacks. *arXiv preprint arXiv:2503.00187*, 2025.

[3] 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. *arXiv preprint arXiv:2403.19103*, 2024.

[4] **Alexander Robey**, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *arXiv preprint arXiv:2310.03684*, 2023.

[5] Thomas Waite, **Alexander Robey**, Hassani Hamed, George J. Pappas, and Radoslav Ivanov. Data-Driven Modeling and Verification of Perception-Based Autonomous Systems. *arXiv preprint arXiv:2312.06848*, 2023.

[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. *arXiv preprint arXiv:2402.16192*, 2024.

[7] **Alexander Robey**, Hamed Hassani, and George J. Pappas. Model-Based Robust Deep Learning. *arXiv preprint arXiv:2005.10247*, 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.