Vikash Sehwag

vsehwag.github.io ☎ (609)-216-6036
✉ sehwag.vikash@gmail.com

Research Interests

My vision is to develop the next generation of generative artificial intelligence (AI) systems safely and responsibly. I am interested in identifying emerging safety and security challenges in large-scale AI systems, developing risk mitigation strategies for improved robustness and safety, and enhancing the overall trustworthiness of generative AI.
Research topics: Safer generative AI, Responsible data synthesis, Robust machine learning, Benchmarking progress in AI safety.
Sub-topics: Automated red teaming, Multimodal robust learning, Diffusion models, Adversarial robustness, Data watermarking and tracing, Data memorization and privacy leakage, Bias and Fairness, Trustworthy AI.
Phd thesis: Promises and Pitfalls of Generative AI: An AI-Safety Centric Approach (Princeton University)

Work Experience

- Research Scientist –*Sony AI*, USA — 2023 - present
  I lead research efforts on enhancing safety and utility of generative models at Sony AI.
- Research Internship –*Meta AI*, USA — Summer 2021
  Advisors – Caner Hazirbas, Cristian Canton Ferrer (*AI Red Team*)
  Project: Generating novel hard instances from low-density regions using generative models.
- Research Internship – *Microsoft Research*, Redmond (USA) — Summer 2019
  Advisors – Jay Stokes, Cha Zhang
  Project: Adversarial attacks and defenses beyond ℓ_p norms
- Research Internship – *Technische Universität Darmstadt*, Germany — Summer 2016
  Advisor – Heinz Koeppl
  Project: A study of stochastic SIS disease spreading on random graphs

Education

Program	Institution	Years
Ph.D., Electrical and Computer Engineering *Advisors – Prateek Mittal, Mung Chiang*	Princeton University NJ, USA	2017 - 2023
M.A., Electrical Engineering	Princeton University NJ, USA	2017 - 2019
B.Tech., Electronics and Electrical Communication Engg.	Indian Institute of Technology (IIT) Kharagpur, INDIA	2013 - 2017

Honors and Awards

- Received the 2023 Adversarial Machine Learning (AdvML) rising star award — 2023
- Graduate student award for excellence in service (ECE department, Princeton University) — 2022
- Charlotte Elizabeth Proctor Honorific Fellowship, one of the highest honors at Princeton University — 2022
- Best paper honorable mention award at ICLR workshop on Security and Safety in ML Systems — 2021
- Winner of Qualcomm Innovation Fellowship, North America Region — 2019
- Received best undergraduate thesis award (1 from 72 students) at IIT Kharagpur — 2017
- IEEE student award from IEEE student branch of IIT Kharagpur — 2016
- Awarded the WISE scholarship from German Academic Exchange Service (DAAD), Germany — 2016
- Received Merit-cum-Means Scholarship from MHRD, Government of India — 2013-17

Publications

Preprints and papers under review

- JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models
  Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, **Vikash Sehwag**, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, Eric Wong
  Arxiv, 2024
- Finding a needle in a haystack: A Black-Box Approach to Invisible Watermark Detection
  Minzhou Pan, Zhenting Wang, Xin Dong, **Vikash Sehwag**, Lingjuan Lyu, Xue Lin
  Under review, 2024
- How to Trace Latent Generative Model Generated Images without Artificial Watermark?
  Zhenting Wang, **Vikash Sehwag**, Chen Chen, Lingjuan Lyu, Dimitris N. Metaxas, Shiqing Ma
  Under review, 2024
- Position Paper: AI Risk Management Should Unambiguously Take into Account Both Safety and Security
  Qi at al, 2024
  Under review
- Scaling Compute Is Not All You Need for Adversarial Robustness
  Edoardo Debenedetti, Zishen Wan, Maksym Andriushchenko, **Vikash Sehwag**,
  Kshitij Bhardwaj, Bhavya Kailkhura
  Arxiv 2023, Under review
- DP-RAFT: A Differentially Private Recipe for Accelerated Fine-Tuning
  Ashwinee Panda, Xinyu Tang, **Vikash Sehwag**, Saeed Mahloujifar, Prateek Mittal
  Arxiv 2023, Under review

Conference and Journal Publications

- Differentially Private Image Classification by Learning Priors from Random Processes
  Xinyu Tang, Ashwinee Panda, **Vikash Sehwag**, Prateek Mittal
  *Neural Information Processing Systems (**NeurIPS**), 2023 - Spotlight presentation*
- Extracting Training Data from Diffusion Models
  Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, **Vikash Sehwag**,
  Florian Tramèr, Borja Balle, Daphne Ippolito, Eric Wallace
  ***USENIX Security Symposium**, 2023*
- Uncovering Adversarial Risks of Test-Time Adaptation
  Tong Wu, Feiran Jia, Xiangyu Qi, Jiachen T. Wang, **Vikash Sehwag**, Saeed Mahloujifar, Prateek Mittal
  *International Conference on Machine Learning (**ICML**), 2023*
- MultiRobustBench: Benchmarking Robustness Against Multiple Attacks
  Sihui Dai, Saeed Mahloujifar, Chong Xiang, **Vikash Sehwag**, Pin-Yu Chen, Prateek Mittal
  *International Conference on Machine Learning (**ICML**), 2023*
- A Light Recipe to Train Robust Vision Transformers
  Edoardo Debenedetti, **Vikash Sehwag**, Prateek Mittal
  *IEEE Conference on Secure and Trustworthy Machine Learning (**SaTML**), 2023*
- Generating High Fidelity Data from Low-density Regions using Diffusion Models
  Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, Cristian Canton Ferrer
  *Conference on Computer Vision and Pattern Recognition (**CVPR**), 2022*
- Understanding Robust Learning through the Lens of Representation Similarities
  Christian Cianfarani*, Arjun Nitin Bhagoji*, **Vikash Sehwag***, Ben Zhao, Prateek Mittal, Haitao Zheng
  *Neural Information Processing Systems (**NeurIPS**), 2022*
- Robust Learning Meets Generative Models: Can Proxy Distributions Improve Adversarial Robustness?
  Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, Prateek Mittal
  *International Conference on Learning Representations (**ICLR**), 2022*

* refers to equal contribution.

- RobustBench: a standardized adversarial robustness benchmark
Francesco Croce*, Maksym Andriushchenko*, **Vikash Sehwag***, Edoardo Debenedetti*, Nicolas Flammarion, Mung Chiang, Prateek Mittal, Matthias Hein
*Neural Information Processing Systems (**NeurIPS**), 2021 - Datasets and Benchmarks Track*
*Won **best paper honorable mention prize** at ICLR 2021 workshop on Security and Safety in Machine Learning Systems.*
- Lower Bounds on Cross-Entropy Loss in the Presence of Test-time Adversaries
Arjun Nitin Bhagoji, Daniel Cullina, **Vikash Sehwag**, Prateek Mittal
*International Conference on Machine Learning (**ICML**), 2021*
- SSD: A Unified Framework for Self-Supervised Outlier Detection
Vikash Sehwag, Mung Chiang, Prateek Mittal
*International Conference on Learning Representations (**ICLR**), 2021*
Short version accepted at NeurIPS 2020 Workshop on Self-Supervised Learning - Theory and Practice
- Beyond ℓ_p Norms: Delving Deeper into Robustness to Physical Image Transformations
Vikash Sehwag, Jay Stokes, Cha Zhang
IEEE Military Communications Conference (MILCOM), 2021
- PatchGuard: Provable Defense against Adversarial Patches Using Masks on Small Receptive Fields
Chong Xiang, Arjun Nitin Bhagoji, **Vikash Sehwag**, Prateek Mittal
***USENIX Security Symposium**, 2021*
- HYDRA: Pruning Adversarially Robust Neural Networks
Vikash Sehwag, Shiqi Wang, Prateek Mittal, Suman Jana
*Neural Information Processing Systems (**NeurIPS**), 2020*
- Fast-Convergent Federated Learning
Hung T. Nguyen, **Vikash Sehwag**, Seyyedali Hosseinalipour, Christopher G. Brinton, Mung Chiang, H. Vincent Poor
*IEEE Journal on Selected Areas in Communications (**J-SAC**) - Series on Machine Learning for Communications and Networks, 2020*

Peer-reviewed Workshop Publications

- Differentially Private Generation of High Fidelity Samples From Diffusion Models
Vikash Sehwag*, Ashwinee Panda*, Ashwini Pokle, Xinyu Tang, Saeed Mahloujifar, Mung Chiang, Zico Kolter, Prateek Mittal
ICML workshop on Deployable Generative AI, 2023
- Just Rotate it: Deploying Backdoor Attacks via Rotation Transformation
Tong Wu, Tianhao Wang, **Vikash Sehwag**, Saeed Mahloujifar, Prateek Mittal
*In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (**AISec**), 2022*
- Robustness from Perception
Saeed Mahloujifar, Chong Xiang, **Vikash Sehwag**, Sihui Dai, Prateek Mittal
ICLR workshop on Security and Safety in Machine Learning Systems, 2021
- Time for a Background Check! Uncovering the impact of Background Features on Deep Neural Networks
Vikash Sehwag, Rajvardhan Oak, Mung Chiang, Prateek Mittal
ICML workshop on Object-Oriented Learning, 2020
- On separability of self-supervised representations
Vikash Sehwag, Mung Chiang, Prateek Mittal
ICML workshop on Uncertainty & Robustness in Deep Learning, 2020
- On Pruning Adversarially Robust Neural Networks
Vikash Sehwag, Shiqi Wang, Prateek Mittal, Suman Jana
ICLR workshop on Towards Trustworthy ML, 2020
- Analyzing the robustness of open-world machine learning
Vikash Sehwag*, Arjun Nitin Bhagoji*, Liwei Song*, Chawin Sitawarin, Daniel Cullina, Mung Chiang, Prateek Mittal
*In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (**AISec**), 2019*

- Not All Pixels are Born Equal: An Analysis of Evasion Attacks under Locality Constraints
  Vikash Sehwag, Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, Prateek Mittal
  Poster at ACM SIGSAC Conference on Computer and Communications Security (CCS), 2018.

Academic services

Teaching

- Lecture on basics of adversarial machine learning at Princeton-Intel REU Seminar 2021
- Teaching assistant for ECE 574: Security & Privacy Fall 2021
- Taught a mini-course on adversarial attacks & defenses in Wintersession at Princeton University 2020
- Teaching assistant for ECE 535: Machine Learning and Pattern Recognition Fall 2019

Mentoring

I continue to mentor the next generation of researchers.

- *Christian Cianfarani* - Graduate student at University of Chicago. 2021-now
- *Edoardo Debenedetti* - Master's student at École polytechnique fédérale de Lausanne (EPFL) 2021-2022
- *Rajvardhan Oak* - Master's student at University of California, Berkeley Summer 2020
- *Tinashe Handina* (B.S.E., Electrical Engineering 2021) - now a graduate student at Caltech.
- *Matteo Russo* (B.S.E., Computer Science 2020) - now a masters student at ETH Zurich.

Peer reviewing Conference on Neural Information Processing Systems (NeurIPS) - *2021, 2022, 2023*; IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) - *2023, 2024*; International Conference on Learning Representations (ICLR) - *2022, 2024*; ACM Computing Surveys - *2023*; Transactions on Machine Learning Research (TMLR) - *2022*; International Conference on Machine Learning (ICML) - *2022*; Conference on Computer Vision and Pattern Recognition (CVPR) - *2022*; International Conference on Computer Vision (ICCV) - *2021, 2023*; Privacy Enhancing Technologies Symposium (PETS) - *2021*, *2022*; Conference on Information Sciences and Systems (CISS) - *2020*, *2022*; PLOS Computational Biology - *2020*; ACM Transactions on Privacy and Security (TOPS) - *2019*; USENIX Security Symposium - *2018*, *2019*

Other Services

- Workshop organizer - ICCV 2023 ARROW workshop, CVPR 2023 Workshop of Adversarial Machine Learning on Computer Vision: Art of Robustness 2023
- Program committe member for IEEE Conference on Secure and Trustworthy Machine Learning 2023
- Organized more than 20 talks on security & privacy in machine learning (SPML seminar series) 2022
- Part of core maintaining team of Adversarial Robustness Benchmark (robustbench.github.io) 2020-now
- Volunteered as junior mentor at Princeton-OLCF-NVIDIA GPU Hackathon 2020

Invited Talks

- On Safety Risks of Generative AI - From ChatGPT to DallE.3 Nov 2023
  One of the three invited speakers at Responsible AI Webinar, Columbia University
- Prospects and Pitfalls of modern generative models - An AI safety perspective Feb 2023
  Workshop on Practical Deep Learning in the Wild (AAAI 2023)
- Enhancing machine learning using synthetic data distilled from generative models Jan 2023
  Microsoft Research, Cambridge
- Role of synthetic data in trustworthy machine learning May 2022
  University of Chicago; University of California, Berkeley
- A generative approach to robust machine learning Mar 2022
  Annual Conference on Information Sciences and Systems (CISS)
- A generative approach to robust machine learning (link) Jan 2022
  RIKEN-AIP TrustML Young Scientist Seminar, Japan

- Generating novel hard-instances form low-density regions using generative models Aug 2021
  Facebook AI, USA
- A primer on adversarial machine learning July 2021
  Princeton-Intel REU Seminar
- Embedding data distribution to make machine learning more reliable March 2021
  Adversarial robustness seminar, École polytechnique fédérale de Lausanne (EPFL)
- Private Deep Learning Made Practical Oct 2019
  Qualcomm, San Diego