Dr. Vasilios Mavroudis

Principal Research Scientist
AI for Cyberdefence Research Centre
Alan Turing Institute, UK

Email: vas@mavroud.is
Web: https://mavroud.is
Group: https://turing.ac.uk/aicd

Research Interests

- ML for Security: LLMs for Cybersecurity tasks, RL for Cyber Decision-Making.
- AI Safety: Evaluations, Red Teaming, Guardrails, Dangerous Capabilities

Languages & Packages

Python, PyTorch, Tensorflow, Numpy, Stable Baselines 3, Ray RLlib, CleanRL, Transformers HuggingFace, JavaCard

Selected Conferences

NeurIPS, ACM CCS, NDSS, PETs Symposium, BlackHat US, Defcon, RSA Conference, Chaos Communication Congress

Employment

Principal Research Scientist Alan Turing Institute, UK 2022–Now
- Founded the "AI for Cyberdefence" research centre.
- Managing a research team of 12 researchers.
- Awarded and managing £3,550,000 research budget from three funders.

Research Associate Alan Turing Institute, UK 2020-22
- Novel work on reinforcement learning for active cybersecurity attacks.
- LLMs for security and privacy estimation in anonymity systems.
- Developed and prototyped novel schemes for digital identity services in millions of legacy mobile devices (B&M Gates Foundation).

Researcher Visa Research, US 2020
- Privacy-preserving Tabular GANs for synthetic data.
- Patent application pending.

Visiting Researcher ETH Zurich, Switzerland 2019
- Decentralized provably secure system for low-latency onchain cryptocurrency payments.
- NDSS 2020, 17.4% acceptance rate; patent pending.

Visiting Researcher University of California Santa Barbara, US 2015–16
- Privacy leakage in cross-device advertising solutions.

Internship University of California, Santa Barbara, US 2014
- Malicious javascript and evasion techniques.

Research Assistant Center for Research & Technology Hellas, Greece 2013–14
- ML models for the early detection of large-scale attacks against telecommunication networks.

Research Assistant Deutsche Bank, Technology Security dept, Germany 2012
- Designed and built a proof-of-concept system to protect web-banking customers from targeted malware.

Education

- PhD in Computer Science, UCL — 2020
  Information Leakage Attacks and Countermeasures
- MSc Information Security, UCL — 2015
  Distinction; Thesis: Privacy-preserving Statistics for Anonymity Networks
- BSc Computer Science, UoM — 2012
  Distinction

Honours, Awards & Activities

- Gave evidence on AI Risk in the House of Commons, UK Parliament — 2024
- "Scientific Report on the Safety of Advanced AI" with Yoshua Bengio et. al — 2024
- Semi-finalist in the AIxCC competition (Autonomous Vulnerability Discovery) — 2024
- BlackHat US talk on Backdoors in Reinforcement Learning — 2024
- Spotlight presentation at AISec 2023 for our MARL paper — 2023
- 1st place in the CAGE autonomous cyberdefence competition (again) — 2023
- 1st place in the CAGE autonomous cyberdefence competition — 2022
- Spark Award nomination for the most promising invention of the year — 2021
- 1st place prize (Future of Blockchain) for "Snappy Payments" — 2019
- Grant for the development of "Cryptogame", UCL PEU — 2018
- Award Finalist CSAW Europe 2018 Applied Research Award — 2018
- Honor Heidelberg Laureate Forum's 10-out-of-200 young researchers list — 2018
- Werner Romberg Grant by the Heidelberg Laureate Forum — 2018
- Research Grant from the Allan & Nesta Ferguson Charitable Trust — 2018
- Grant Data Transparency Lab engagement funding — 2016
- Award Dean's List for outstanding academic performance, UCL — 2016
- Honor Distinction in Information Security M.Sc; 1st in cohort, UCL — 2015
- Award First place at UCL code breaking competition — 2015

Funding

- AI Safety for LLMs, co-PI (AI Safety Institute, £250,000) — 2024–2025
- Generalizable RL for Security, co-PI (NCSC, £250,000) — 2023–2025
- AI for Cyberdefence, co-PI (Dstl, £3,000,000) — 2022–2025
- Human-Machine Teaming, co-PI (US Army Research Labs, £300,000) — 2022–2024
- Reinforcement Learning for Systems Security, co-PI (GCHQ, £250,000) — 2021

Selected Publications

- Online Convex Optimisation: The Optimal Switching Regret for all Segmentations Simultaneously; Pasteris S., Hicks C., Mavroudis V. Herbster M., Annual Conference on Neural Information Processing Systems (NeurIPS), 2024 [Spotlight 3%]
- Benchmarking OpenAI o1 in Cyber Security; Ristea D., Mavroudis V., Hicks C., arXiv:2410.21939 ,2024
- Nearest Neighbour with Bandit Feedback; Pasteris S., Hicks C., Mavroudis V., Annual Conference on Neural Information Processing Systems (NeurIPS), 2023
- Adaptive Webpage Fingerprinting from TLS Traces; Mavroudis V., Hayes J., 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2023
- Autonomous network defence using reinforcement learning; Foley M., Hicks C., Highnam K., Mavroudis V., Asia Conference on Computer and Communications Security (AsiaCCS), 2022
- On the Privacy and Security of the Ultrasound Tracking Ecosystem; Mavroudis V., Hao S., Fratantonio Y., Maggi F., Kruegel C., Vigna G., Proceedings of the Privacy Enhancing Technologies Symposium (PETs), 2017

Please find the full list of publications on my website.