KUREHA YAMAGUCHI

London, UK | ky295@cantab.ac.uk | LinkedIn | Google Scholar | Github

EDUCATION

University of Cambridge Oct 2018 - Jun 2022

Information & Computer Engineering, ***B.A.*** *&* ***M.A.*** *&* ***MEng*** *with Honours*

Modules

- Bachelor's and Master's modules covering areas of Computational Statistics, Machine Learning, Deep Learning, Statistical Signal Analysis, Data Transmission, Information Theory and Coding

Scholarship

- ❀ Awarded the Diamond Jubilee Scholarship, a 4-year merit-based scholarship by The Institution of Engineering and Technology

North London Collegiate School Sep 2011 - Jun 2018

- Graduated with 4A*s at A-Levels (2018), 11A*s at GCSE/IGCSEs (2016), and A*s in my Extended Project Qualification (2018) and Free Standing Maths Qualification (2016)

PUBLIC RESEARCH OUTPUTS

Adversarial Manipulation of Reasoning Models using Internal Representations
ICML, Workshop on Reliable and Responsible Foundation Models (2025)

Kureha Yamaguchi, Benjamin Etheridge, and Andi Arditi

[Under Review] 2DSig-Detect: a semi-supervised framework for anomaly detection on image data using 2D-signatures
Pattern Recognition Journal, Elsevier (2025)

Xinheng Xie, **Kureha Yamaguchi**, Margaux Leblanc, Simon Malzard, Varun Chhabra, Victoria Nockles, and Yue Wu

An AI Red Team Playbook
SPIE Defense and Commercial Sensing, Assurance and Security for AI-enabled Systems (2024)

Anna Raney, Shiri Bendelac, Keith Manville, Mike Tan, and **Kureha Yamaguchi**

An AI Blue Ream Playbook
SPIE Defense and Commercial Sensing, Assurance and Security for AI-enabled Systems (2024)

Mike Tan, **Kureha Yamaguchi**, Anna Raney, Victoria Nockles, Margaux Leblanc, and Shiri Bendelac

RELEVANT EXPERIENCE

The Alan Turing Institute Sep 2023 - Present
Data Scientist *London, UK*

- Researching AI Security and Safety within the Turing's Defence and National Security programme, with UK Government stakeholders and academic collaborators
- Demonstrating and presenting my AI research at AIUK, DSTL's AI Showcase, Royal Astronomical Society, as well as at academic conferences and stakeholder meetings
- Serving as a technical advisor to shape Ministry of Defence (MoD) guidance on Dependable AI in Defence (JSP 936)
- Contributing to the open-source MITRE ATLAS matrix following collaborative research on 'Vulnerabilities in Critical Governmental Use of Large Language Model-based Analysis Tools'
- Engaging in cross-organizational collaboration to co-organise the Women in AI Security Workshop 2024 and 2025, bringing together experts across government, academia, and industry
- ❀ **Winning first place** at the UK MoD x Google Cloud Hackathon- the judging panel consisted of Senior HM Government leaders and Google's technology leadership, including GCHQ's Chief Data Scientist, Head of DAIC, and Google Cloud's Vice President of Engineering

· Building leading-edge Reinforcement Learning systems through close collaboration with stakeholders at the Foreign Office to improve detection of possible hostile activity in UK's Arctic Mission

Supervised Program for Alignment Research Feb 2025 - May 2025
Research Fellow *Virtual*

· Leading applied mechanistic interpretability research under the mentorship of Andy Arditi (former Anthropic AI Safety Fellow), and publishing the research at ICML 2025 R2-FM Workshop
· ❀ **Winning first place** for my research fellowship presentation on 'Adversarial Manipulation of Reasoning Models using Internal Representations'

Climate Change AI (CCAI) Oct 2021 - Dec 2023
Research Community Manager *Virtual*

· Reviewing academic papers submitted to CCAI's workshops on 'Tackling Climate Change with Machine Learning' at leading international AI/ML conferences, ICML and NeurIPS
· Managing a rapidly growing community platform of $+7,000$ members which serves to catalyse impactful work at the intersection of climate change and machine learning by driving interdisciplinary collaboration between technical and subject matter experts

Okinawa Institute of Science and Technology Graduate University Jul 2021 - Sep 2021
Software Engineer Intern *Okinawa, Japan*

· Providing practical embedded hardware and software support for multi-purpose, multi-sensor autonomous systems at the Neural Computation Unit under Professor Kenji Doya
· Developing software (Python, Java, Git) to efficiently and robustly implement the model-based policy search learning framework, 'Probabilistic Inference for Particle-Based Policy Search' (Parmas et al., 2018), for an ambitious real-world robot balancing task
· Exercising good coding practices when writing, testing and deploying production code, such as Git version control, technical documentation, issue tracking and unit testing

Cambridge University Robotics × Google DeepMind Dec 2019 - Aug 2020
AI Student Researcher *Cambridge, UK*

· Advancing reinforcement learning research in collaboration with Deepmind and the Bio-Inspired Robotics Laboratory, which required exercising my problem-solving skills and ability to learn quickly
· Developing edge AI methods in a simulated environment, adapted from OpenAI gym
· Collaborating with research colleagues and technical experts from Google DeepMind to apply novel AI algorithms, engineering and embedded hardware techniques

Impact Through Innovation Cambridge Jun 2020 - Jul 2021
Innovation Lead *Cambridge, UK*

· Managing an audacious portfolio of high impact projects in an interdisciplinary student team, through a culture of rapid prototyping and relentless problem-solving for innovating under ambiguity

City, University of London Jun 2017 - Jul 2017
Embedded Systems Intern *London, UK*

· Embodying the laboratory's cutting-edge research outputs through rapid hardware prototyping, programming, and constructing electrical circuits, supervised under Professor Panicos Kyriacou

ADDITIONAL INFORMATION

Spoken languages	English (native speaker), Japanese (fluent), German (basic)
Programming languages	Python (+5yrs professional experience) Bash, Java, MATLAB, C++, Julia (knowledgeable)
Machine Learning Frameworks	PyTorch (proficient) Scikit-learn, JAX, TensorFlow, NNSight, TransformersLens (knowledgeable)