

Gabriele La Malfa

Talbot Yard, London SE1 1XT | gabriele.la_malfa@kcl.ac.uk | [GitHub](#) | [Personal Website](#)

While studying for my master's degree in accounting, I got hugely passionate about game theory. As I continued my business studies, I focused on moving toward science. I learned programming, mathematics, and statistics on my own, which helped me gain admission to the PhD program in computer science at King's College London. There, I was able to fully pursue my passion for science. Now, in my final year of doctoral studies, I am researching multi-agent reinforcement learning, as well as the safety and fairness of AI. Lately, my work has focused on self-play, large language models, the Theory of Space, and vision-language models.

I am a human-oriented person. Projects are exciting, but people around me and collaborators always come first.

ACADEMIC POSITIONS

PhD candidate in [Safe & Trusted AI](#) at King's College London Oct. 2022 – Current
Safety and Fairness in Multi-Agent Systems.

– Supervisors: Prof. Elizabeth Black, Dr. Jie M. Zhang and Prof. Michael Luck.

Visiting Researcher at [Nokia Bell Labs](#) Jul. 2025 – Dec 2025
Classification of AI agent risks in the workplace.

– Researching AI impacts on society under the supervision of Prof. Daniele Quercia.

Research Assistant at the University of Cambridge, [Centre for Risk Studies](#) Oct. 2020 – Aug. 2022
Research and modelling business, financial and cyber risks.

– Co-published: [Mitigating ransomware risk: determining optimal strategies for businesses.](#)

WORK EXPERIENCE

Quantitative Risk Officer Jun. 2020 – Aug. 2020
Deutsche Börse AG

EDUCATION

EMLyon Business School 2020
MSc Quantitative Finance

University of Eastern Piedmont 2017
Master's Degree in Accounting and Business Administration

PUBLICATIONS

[Explore, Map, Remember, Decide: Are Embodied VLMs Ready for Safety-Critical Scenarios?](#)

Under submission.

Gabriele La Malfa, Nitay Alon, Emanuele La Malfa, Reuth Mirsky, Stefan Sarkadi.

[Unaccountable Delegation, Fading Skills: Mapping the Risks of Workplace AI Agents.](#)

Proceedings of the Ninth AAAI/ACM Conference on AI, Ethics, and Society (AIES-26) - Main Track.

Gabriele La Malfa, Lakmal Meegahapola, Edyta Bogucka, Jie M. Zhang, Michael Luck, Elizabeth Black, Daniele Quercia.

[The Attacker in the Mirror: Breaking Self-Consistency in Safety via Anchored Bipolicy Self-Play.](#)

Under submission.

Gabriele La Malfa, Emanuele La Malfa, Saar Cohen, Jie M. Zhang, Michael Luck, Michael Wooldridge, Elizabeth Black.

[Fairness Aware Reinforcement Learning via Proximal Policy Optimization.](#)

Vol. 8 No. 1 (2025): Proceedings of the AAAI Conference on Artificial Intelligence, 40(27), 22725–22733.

Gabriele La Malfa, Jie M. Zhang, Michael Luck, Elizabeth Black.

[Large Language Models Miss the Multi-Agent Mark.](#)

The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track.

Gabriele La Malfa, Emanuele La Malfa, Samuele Marro, Jie M. Zhang, Elizabeth Black, Michael Luck, Philip Torr, Michael Wooldridge

[Using Protected Attributes to Consider Fairness in Multi-Agent Systems.](#)

Proceedings of the AEQUITAS 2024 ECAI Workshop on Fairness and Bias in AI (2024).

Gabriele La Malfa, Jie M. Zhang, Michael Luck, Elizabeth Black.

[Deep Neural Networks via Complex Network Theory: a Perspective.](#)

Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Main Track. Pages 4361–4369.

Gabriele La Malfa, Emanuele La Malfa, Giuseppe Nicosia, Vito Latora.

CURRENT/FUTURE PROJECTS

Self-play red teaming for autonomous agents Operative Systems (OS) security. I extend the concept of single agent oversight to AI agent "anti-malware" system, in which a population of AI agents (blue team) monitors health of an OS across both user and agent spaces to detect, intercept, and shield against faulty or malicious behaviours of other AI agents (red team). Through a lightweight, distributed Parameter-Efficient Fine-Tuning (PEFT) this framework trains attackers and defenders in a self-reinforcing loop, aiming at secure and safety deployment.

Self-play red teaming and mosaic attacks. Building on my paper "The Attacker in the Mirror: Breaking Self-Consistency in Safety via Anchored Bipolicy Self-Play", I am working on an advanced version of self-play red team to make LLMs safer against mosaic attacks, i.e., multi-prompt inputs that appear benign individually but produce harmful outcomes when combined.

Multi-agent Theory of Space in safety critical scenarios. Building on my paper "Explore, Map, Remember, Decide: Are Embodied VLMs Ready for Safety-Critical Scenarios?", I am exploring extending the work to further safety-critical scenarios in the presence of deception of focal points and multi-agent coordination.

TEACHING

Artificial Intelligence <i>Practical</i>	2023–2025
Mathematics and Statistics for AI <i>Group Tutorial Lead</i>	2024–2025
Machine Learning <i>Group Tutorial Lead</i>	2023–2024
Logic and Knowledge Representation <i>Group Tutorial Lead</i>	2023–2024
Elementary Logic and Applications <i>Group Tutorial Lead</i>	2023–2024
Big Data Technology <i>Group Tutorial Lead</i>	2023–2024

SERVICE & ORGANISATIONS

LOD Conference Series (2020–2025)
Organizing Committee

AAAI, NeurIPS
Reviewer

VOLUNTEERING

I am committed to social justice, ranging from dialogue leadership at KCL to humanitarian support for students and people who are not as lucky as I am.

Lunchtime Dialogues: Identity & Social Justice 2023 – 2025
Lead Facilitator: Organised a dialogue series across all King's College London faculties to raise awareness on social justice and human rights issues.

Ukrainian Global University 2022 – 2023
Interviewer: Supported the international placement program by conducting admissions interviews for Ukrainian students seeking to study at partner universities abroad.

The Kharkiv and Przemyśl Project August 2022
Volunteer: Provided on-the-ground support in Przemyśl, Poland, assisting refugees arriving from and returning to Ukraine.

Cambridge University Amnesty International (CUAI) 2021
Executive Committee Member: Led local campaigns including Write for Rights (W4R) and various global human rights initiatives.

Eutanasia Legale July 2021
Volunteer: Coordinated signature collection for a national referendum to decriminalise euthanasia in Italy, contributing to a total of 1.2 million signatures.

Centro Gazza Ladra 2018
Project Organiser: Managed crowdfunding and event logistics for an association dedicated to supporting children with special needs.