Kamile Lukosiute

lukosiutekamile@gmail.com

kamilelukosiute.com

AI cyber misuse risk researcher. My work directly informs frontier safety policy decisions at frontier AI labs. I advise safety teams at major AI labs on cyber threat modeling and engage with the EU AI Office on AI risk governance. Former Anthropic alignment Resident, former astrophysicist.

EXPERIENCE

Centre for the Governance of AI — *Research Scholar* Aug 2025 – Present

San Francisco, CA

- Conduct threat modeling research that directly informs frontier safety policy decision-making at frontier AI laboratories; ongoing engagement with safety teams at all major labs
- Write technical reports and policy memos on AI cyber threat models for lab and government audiences
- Advise EU AI Office on state of the art AI risk modeling practices
- Build and maintain relationships with lab safety teams to identify research priorities that address deployment decisions
- Supervise GovAI Winter Fellow on AI/cyber policy research

Cisco Systems (via Robust Intelligence acquisition) — *AI Security Researcher* Jun 2024 – Jun 2025

San Francisco, CA

- Conducted product-focused security research, training and fine-tuning language models for threat classification
- Developed BERT-based models for defensive security applications (log analysis, anomaly detection) in collaboration with Cisco Secure Malware Analytics (Threat Grid) and Splunk teams
- Supervised research intern on novel jailbreaking techniques (paper forthcoming); contributed initial codebase and served as research advisor
- First author, "LLM Cyber Evaluations Don't Capture Real-World Risk" (arXiv:2502.00072)

Independent Research Jan 2024 – Jun 2024

San Francisco, CA

- Developed LLM evaluation methodologies in collaboration with Center for AI Safety
- Collaboration with CAIS on "Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?" (arXiv:2407.21792)
- Published practitioner-focused writing on evaluation design and limitations

Anthropic — *Resident Researcher, AI Alignment* Oct 2022 – Jul 2023

San Francisco, CA

- Contributed experimental work to "model-written evaluations" methodology for discovering problematic model behaviors, including experiments on measuring bias (published at ACL 2023)
- Contributed to foundational safety research: scalable oversight and debate
- Collaborated with alignment and reinforcement learning teams

University of Amsterdam — *Instructor & PhD Candidate* Jan 2021 – Apr 2022

Amsterdam, NL

- Designed and taught machine learning curriculum for MSc Physics students
- Departed PhD to focus full-time on AI safety research

EDUCATION

MS Physics & Astronomy — University of Amsterdam, NL 2021

Thesis: Machine learning methods for astrophysical event classification

BA Physics, cum laude — Wellesley College, MA 2019

SELECTED PUBLICATIONS

- K. Lukosiute, J. Halstead, L. Righetti, "Global cybercrime damages: A baseline for frontier AI risk assessment," GovAI Technical Report, forthcoming
- K. Lukosiute & A. Swanda, "LLM Cyber Evaluations Don't Capture Real-World Risk," arXiv:2502.00072
- E. Perez, K. Lukosiute, et al., "Discovering Language Model Behaviors with Model-Written Evaluations," Findings of ACL, 2023
- K. Lukosiute, G. Raaijmakers, Z. Doctor, M. Soares-Santos, B. Nord, "KilonovaNet: Surrogate Models of Kilonova Spectra with Conditional Variational Autoencoders," Monthly Notices of the Royal Astronomical Society, 2022
- Additional writing at kamilelukosiute.com