Ilias Tsingenopoulos

Postdoctoral Researcher | AI Safety & Security
Department of Computer Science | KU Leuven, Belgium

+32486319227 | ilias.tsingenopoulos@cs.kuleuven.be | Webpage
LinkedIn | itsiggen | Google Scholar

About me

I am a postdoctoral researcher specializing in the intersection of AI Security, Reinforcement Learning, and Foundation Models. My expertise spans both theoretical and practical aspects of robust learning across diverse AI systems and modalities: from web-bot detection like Google reCaptcha, to hardening antivirus models against adversarial malware, and more generally through adapting and optimizing RL-based attacks and defenses against each other under the competitive game they form.

Currently I am developing RL-based attacks and defenses on LLMs, across a range of tasks like forbidden question answering, private information retrieval, and text classification. Beyond identifying and patching vulnerabilities however, my goal is to develop systematic approaches for robust-by-design AI that address fundamental challenges like reward hacking and the inner/outer misalignment.

Professional & Research Experience

DistriNet, KU Leuven — Leuven, Belgium
Postdoctoral Researcher in AI Safety & Security — *October 2024 – Present*
- LLM security through RL-based optimization of jailbreaks.
- Investigating problem-space activation engineering with RL.

DistriNet, KU Leuven — Leuven, Belgium
Doctoral Researcher in Robust & Secure AI — *May 2019 – September 2024*
- Adversarial attack adaptation and optimization.
- Model hardening and active defense development.
- Domains: image classification, bot & malware detection.

S2Lab, University College London — London, UK
Visiting Scholar — *January, 2023 – April 2023*
- Rendering a commercial antivirus robust to evasive malware.

DistriNet, KU Leuven — Leuven, Belgium
Research Assistant — *May, 2018 – April 2019*
- Optimization of black-box adversarial attacks.
- Preventing abusive DNS registrations in the .eu domain.

Information Technologies Institute, CERTH — Thessaloniki, Greece
Research Assistant — *January, 2017 – April 2018*
- Research and implementation in several Horizon 2020 projects.
- Formulating and writing research proposals.

Education

Aristotle University of Thessaloniki — Thessaloniki, Greece
Diploma (5-year degree, M.Eng equivalent) in Electrical & Computer Engineering — *Graduated Jul. 2016*
Dissertation: Fuzzy Clustering Algorithms in Feature Subspaces

Technical University of Berlin — Berlin, Germany
Exchange Semester, Department of Electrical Engineering & Computer Science — *Spring 2012*

Selected Publications

- **Tsingenopoulos I.**, Rimmer V., Preuveneers D., Pierazzi F., Cavallaro L., Joosen W. *The Adaptive Arms Race: Redefining Robustness in AI Security*. In: 28th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2025).
- **Tsingenopoulos I.**, Cortellazzi J., Bosansky B., Aonzo S., Preuveneers D., Joosen W., Pierazzi F., Cavallaro L. *How to Train your Antivirus: RL-based Hardening through the Problem-Space*. In: 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2024).
- **Tsingenopoulos I.**, Rimmer V., Preuveneers D., Pierazzi F., Cavallaro L., Joosen W. *On Adaptive Decision-Based Attacks and Defenses*. In: DLSP Workshop, IEEE S&P 2024.
- **Tsingenopoulos I.**, Preuveneers D., Desmet L., Joosen W. *Captcha me if you can: Imitation Games with Reinforcement Learning*. In: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P 2022).
- **Tsingenopoulos I.**, Shafiei A. M., Desmet L., Preuveneers D., Joosen W. *Adaptive Malware Control: Decision-Based Attacks in the Problem Space of Dynamic Analysis*. In: 1st Workshop on Robust Malware Analysis (WoRMA 2022).
- Hernández-Castro C. J., Liu Z., Serban A., **Tsingenopoulos I.**, Joosen W. *Adversarial Machine Learning*. In: Security and Artificial Intelligence: A Crossdisciplinary Approach. Springer, 2022.

Knowledge & Technical Skills

Domain Expertise:

- **Machine Learning**: Clustering, SVMs, Decision Trees/Random Forests, XGBoost
- **Deep Learning**: Convolutional/Graph/Recurrent Neural Networks, GANs, Transformers
- **Reinforcement Learning**: Q-Learning, Policy Optimization, POMDPs, Multi-agent
- **Adversarial ML**: White/Black-box Attacks, Adversarial Training
- **Foundation Models**: RLHF/GRPO Alignment, RAG, Activation Engineering, Agentic Workflows
- **Performance Optimization**: Benchmarking/Profiling, Quantization, Pruning.

Development and Tools:

- **Programming Languages**: Python, C, C++, Java, Matlab
- **Deep Learning**: PyTorch, CUDA, Tensorflow, Keras
- **Reinforcement Learning**: Gym/Gymnasium, Stable Baselines
- **Distributed Training** : HPC, Ray, PyTorch Lightning
- **Others**: Git, LaTex

Online Courses & Summer Schools

- **Advanced LLM Agents**: MOOC Spring 2025
- **LLM Agents**: MOOC Fall 2024
- **CS 285:** Deep Reinforcement Learning UC Berkeley, Fall 2021
- **M2L 2023**: Mediterranean Machine Learning Summer School
- **Security & Privacy in the Age of AI**: Organizer and participant in editions 2022-2024

Languages

Greek: Native | **English**: Excellent - C2 | **German**: Very Good - B2/C1 | **Dutch**: Good - B1

Other Interests

An inquisitive and driven spirit, in my free time I enjoy science fiction, creative writing, and tabletop/computer games; I am also a capoeirista and water polo player.