LUKAS FLURI

https://lukas-fluri.com

EDUCATION

ETH ZURICH — *ZURICH, CH*

05/2025 – today

PhD in AI Safety: Working on making AI safe(r). Supervised by Prof. Florian Tramèr at SPY Lab

09/2020 – 07/2023

M.Sc. in Data Science; Final GPA: 5.66 (above class average), 1.0 – 6.0, with 6.0 being best

- *Master thesis: "Detecting Failures of Superhuman AI Models" – Honored with ETH Medal*

B.Sc. in Computer Science; Final GPA: 5.29 (above class average), 1.0 – 6.0, with 6.0 being best — 09/2016 – 07/2019

- *Bachelor thesis: "Towards Quantum Graph Processing and Analytics"*

PROFESSIONAL & SCIENTIFIC EXPERIENCE

CENTER FOR HUMAN COMPATIBLE AI — *BERKELEY, US*

Research Intern, working on AI safety — 10/2024 – today

Assessed the risks of harmful reward hacking with LLM reward models

- *Supervised by Micah Carroll*

UNIVERSITY OF CAMBRIDGE — *CAMBRIDGE, UK*

Research Paper: "The Perils of Optimizing Learned Reward Functions: Low Training Error Does Not Guarantee Low Regret" — 10/2023 – 05/2024

- *Analyzed theoretical safety guarantees for reward learning, and RLHF in particular*
- *Paper on Arxiv and currently under submission as a conference paper to 'ICLR 2025'*
- *At the chair of Prof. David Krueger, supervised by Joar Skalse, in collaboration with Leon Lang*

SPY LAB ETH ZURICH — *ZURICH, CH*

Research Paper "Evaluating Superhuman Models with Consistency Checks" — 10/2022 – 07/2023

- *Developed a black-box evaluation method that allows to find failures in ML models*
- *Accepted at the 2024 'IEEE Conference on Secure and Trustworthy Machine Learning'*
- *Earlier version accepted as Spotlight paper at 'Neurips SoLaR' 2023 workshop*
- *In collaboration with Daniel Paleka and Florian Tramèr*

LEARNING & ADAPTIVE SYSTEMS GROUP ETH ZURICH — *ZURICH, CH*

Semester Project in the field of data science — 03/2022 – 10/2022

- *Used meta-learning to reduce the sample-frequency of active reward learning for reinforcement learning problems. Supervised by David Lindner and Jonas Rothfuss.*

AI SAFETY CAMP — *REMOTE*

Semester Project — 01/2022 – 07/2022

- *Surveyed the field of negative side-effect minimization*

DATA SCIENCE LAB ETH ZURICH — *ZURICH, CH*

Semester Project — 09/2021 – 02/2022

- *Used Graph Neural Networks to predict and understand cyclone tracks*

ORACLE LABS — *ZURICH, CH*

Research Intern — 05/2020 – 07/2020

- *Implemented ML pipeline for study startup time prediction*

ETH ZURICH — *ZURICH, CH*

Research Assistant in the field of quantum graph processing — 09/2019 – 12/2019

- *Conducted further research based on my Bachelor thesis*

EXTRACURRICULAR ACTIVITIES

ZURICH AI ALIGNMENT GROUP — *ZURICH, CH*

Main Organizer — 08/2022 – today

- *Organizing the 'ZAIA' reading group and the local instance of the 'AI Safety Fundamentals' course*

SKILLS & LANGUAGES

Languages: German (*native language*), English (*fluent*), French (*conversational*), Italian (*basic*)

IT: Python (*advanced*), Classical ML (*advanced*), Deep Learning (*advanced*), LLMs (*advanced*), Pytorch (*advanced*), Git (*advanced*), Latex (*advanced*), C (*intermediate*), Java (*intermediate*), Project Q (*intermediate*), Reinforcement Learning (*intermediate*), NLP (*intermediate*), Matlab (*intermediate*), C++ (*basic*), Haskell (*basic*), TensorFlow (*basic*), PostgreSQL (*basic*), Spark (*basic*), Neo4j (*basic*), Rumble (*basic*)

Links: Website: https://lukas-fluri.com | Github: https://github.com/luk-s