JONATHAN NÖTHER | Curriculum Vitae

✉ jnoether@mpi-sws.org • GitHub • LinkedIn

Interests

Secure Machine Learning, Attacks against ML Models, Reinforcement Learning, Agentic Systems

Education

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS
10/2024 - Ongoing

PhD in Computer Science

Saarbrücken, Germany

SAARLAND UNIVERSITY
12/2022 - 08/2024

M.Sc. in Data Science and Artificial Intelligence

Saarbrücken, Germany

ECTS: 1.3

SAARLAND UNIVERSITY
10/2019 - 11/2022

B.Sc. in Data Science and Artificial Intelligence

Saarbrücken, Germany

ECTS: 1.7

Experience

RESEARCH ASSISTANT
08/2022-07/2024

Machine Teaching Group

MPI-SWS

Conducted research projects and presented my work and related papers

Teaching experience

TEACHING ASSISTANT FOR THE COURSE "GENERATIVE AI"
Winter 2024/2025

Machine Teaching Group

MPI-SWS

Prepare exercise sheets, answered student's questions and graded exercises and the exam

TEACHING ASSISTANT FOR THE SEMINAR "TRUSTWORTHINESS OF FOUNDATION MODELS"
Summer 2024

Multi-Agent Systems Group

MPI-SWS

Prepare the seminar project on red-teaming and watermarking of foundation models

TEACHING ASSISTANT FOR THE LECTURE "STATISTICS LAB"
Summer 2022

Modeling and Simulation Group

Saarland University

Explained course topics to students and graded tests and exams

TEACHING ASSISTANT FOR THE LECTURE "ARTIFICIAL INTELLIGENCE"
Summer 2022

Foundations of Artificial Intelligence Group

Saarland University

Prepared Exercises, explained Course topics to students, and graded tests and exams

TEACHING ASSISTANT FOR "PROGRAMMING 1"
Winter 2020/2021

Reactive Systems Group

Saarland University

Prepared Exercises, explained course topics to students, and graded tests and exams

Skills

PROGRAMMING LANGUAGES	**Experienced:** Python **Familiar:** C++
CONCEPTS	**Experienced:** Machine Learning | Reinforcement Learning | Adversarial ML | Agentic Systems | Large Language Models **Familiar:** Cybersecurity | Computer Vision | Diffusion Models
LIBRARIES	matplotlib | Pytorch | numpy | AutoGen | TRL | transformers
LANGUAGES	**Native:** German | **Fluent:** English (C1)

Publications

MAMA: A GAME-THEORETIC APPROACH FOR DESIGNING SAFE AGENTIC SYSTEMS

PDF

Preprint, under Review — Safety of Agentic Systems

TL;DR: Automatic Design of Safe Agentic Systems using a two-player game between a system designer and an attacker

AGENTICRED: OPTIMIZING AGENTIC SYSTEMS FOR AUTOMATED RED-TEAMING

PDF

Preprint, under Review

Jailbreaking of LLM's

TL;DR: Automatically design red-teaming workflows without human intervention

BENCHMARKING THE ROBUSTNESS OF AGENTIC SYSTEMS TO ADVERSARIALLY-INDUCED HARMFUL ACTIONS

PDF

Preprint, under Review — Safety of Agentic Systems

TL;DR: Benchmark for testing the robustness of LLM-based agents against adversaries that aim to manipulate them into performing dangerous actions

TEXT-DIFFUSION RED-TEAMING OF LARGE LANGUAGE MODELS: UNVEILING HARMFUL BEHAVIORS WITH PROXIMITY CONSTRAINTS

PDF

AAAI 2025 (Oral) — Safety of LLMs

TL;DR: Applying text-diffusion models to red-teaming to satisfy proximity constraints with regards to a reference prompt

POLICY TEACHING VIA DATA POISONING IN LEARNING FROM HUMAN PREFERENCES

PDF

AISTATS 2025 — Safety of LLMs

TL;DR: Forcing an LLM to adapt a target policy by synthesizing preference data

DEFENDING AGAINST UNKNOWN CORRUPTED AGENTS: REINFORCEMENT LEARNING OF ADVERSARIALLY ROBUST NASH EQUILIBRIA

PDF

TMLR 08/2024 — Robust Reinforcement Learning

TL;DR: Training robust agents in an MARL setting where an attacker can arbitrarily corrupt a subset of peer agents of a given cardinality

IMPLICIT POISONING ATTACKS IN TWO-AGENT REINFORCEMENT LEARNING: ADVERSARIAL POLICIES FOR TRAINING-TIME ATTACKS

PDF

AAMAS 2023 — Adversarial Reinforcement Learning

TL;DR: Attacking an agent by poisoning the policy of a peer agent during training

Awards

GÜNTER-HOTZ MEDAILLE

Award for the top computer science graduates of Saarland University

TOP-REVIEWER

Neurips 2025

Reviewing

AAAI 2024 | AAAI 2025 | NeurIPS 2025 | ICLR 2026