JONATHAN NÖTHER | Curriculum Vitae

s8jonoet@stud.uni-saarland.de • Ottweilerstraße 38 / St. Wendel • 01773 5507105
GitHub

Interests

Secure Machine Learning, Attacks against ML Models, AI Safety, Reinforcement Learning

Education

SAARLAND UNIVERSITY — 12/2022 - Ongoing
M.Sc. in Data Science and Artificial Intelligence — Saarbrücken, Germany
ECTS: 1.7

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 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 2019/2020
Reactive Systems Group — Saarland University
Prepared Exercises, explained course topics to students, and graded tests and exams

Skills

PROGRAMMING LANGUAGE	**Experienced:** Python **Familiar:** C++
CONCEPTS	**Experienced:** Machine Learning | LLMs | Reinforcement Learning | Adversarial ML
	Familiar: Cybersecurity | Computer Vision | Diffusion Models
LIBRARIES	matplotlib | Pytorch | numpy
LANGUAGES	**Native:** German | **Fluent:** English (C1)

Publications

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

Under Review at COLM 2024 — Safety of LLMs

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

Under Review at TMLR — Robust Reinforcement Learning

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

PDF

AAMAS 2023 — Adversarial Reinforcement Learning

Projects

INTERVIEW PERFORMANCE PREDICTION AND LIE DETECTION

Project Seminar Data Science and Artificial Intelligence — Type: Computer Vision, NLP

Implementation of model that evaluated the performance and detected lies of a participant of mock-job interviews. Grade : 1.0

INPAITING DETECTION

High Level Computer Vision Course — Type: Computer Vision, Generative AI

Combine automatic segmentation with inpainting to automatically create edited images. Additionally experimented with detecting these faked images. Grade : 1.0

REINFORCEMENT LEARNING PROJECT

Reinforcement Learning Course — Type: Reinforcement Learning

Implementation of a RL-agent that solves the gridworld. Grade(course) : 1.0

SAFE STREETS

AI for the Social Good Seminar — Type: Data Science, Geospatial Data

Extend pedestrian route recommendation by taking into account the safety of the route (e.g. lights, open shops). Grade : 1.0