Diego DORN

Research Engineer

✉ cv@ddorn.fr
🌐 ddorn.fr
github.com/ddorn

Diego works on the mitigation of **systemic risks** from **general-purpose artificial intelligence** systems.

He has extensive expertise in **software engineering** and experience in **teaching**, **leadership** and **communication** from his volunteering. He finishes his master in Communication Systems in August 2024.

Work Experience

Paris 🇫🇷
Feb. 2024 – present

★ **Research engineer at CeSIA (French Center for AI safety)**
*Lead the design of benchmarks to evaluate jailbreak and hallucination detectors for LLMs, red-teamed input-output safeguards. Published "*BELLS: A Framework Towards Future Proof Benchmarks for the Evaluation of LLM Safeguards*" in the NextGen AI Safety workshop at ICML 2024.*

Across Europe 🇪🇺
Aug. 2023 – present

★ **Head Teacher for four ML4Good, a summer school on systemic AI risk** (ml4good.org)
Delivery and improvement of 10 days of technical and conceptual workshopsw for ~20 participants, covering threat modeling, technical safety and AI policy. Management of the teaching team of 2~3.

Cambridge 🇬🇧
July – Sep. 2023

Research assistant, Machine Learning Group, Cambridge University
*Research on goal misgeneralisation in Reinforcement Learning (RL) with N. Alex and D. Krueger. Published "*Goal Misgeneralization as Implicit Goal Conditioning*" in the GCRL workshop at Neurips 2023*

Lausanne 🇨🇭
Jan. 22 – May 23

Lead developer for the startup SPRIG (sprigproofs.org)
Developing a distributed platform to increase confidence in mathematical proofs.

Education

Lausanne 🇨🇭
Sep. 21 – Aug. 2024

Master's in Communication Systems, Ecole Polytechnique Fédérale de Lausanne (EPFL)
Focus on artificial intelligence, formal verification and advanced algorithms. Minor in Mathematics.

Interlaken 🇨🇭
July 2023

★ **Summer school "Science and Policy – How to bridge the gap?"**
5 days on science for policy, science communication, open science and the Swiss policy landscape.

London 🇬🇧
May – June 2023

ARENA, Alignment Research Engineer Accelerator (arena.education)
6 weeks intensive training on interpretability, RL and training at scale.

Lausanne 🇨🇭
Sep. 18 – July 2021

Bachelor's in Mathematics at EPFL
Passed with a 5.42/6 average and top 5/100 of my year.

Volunteering

Lausanne 🇨🇭
Sep. 22 – March 24

★ **Founder and President of the Safe AI Lausanne student association**
Led a team of 8 through the design of a strategy, resulting in a 10-day winter school on systemic AI risks, 3 talks and 2 panel discussions with a total of 10 experts, and a talk for TEDxEcublens.

Lausanne 🇨🇭
Sep. 20 – Sep. 21

President of CQFD, the mathematics students' association of EPFL
Management of a team of 14 people, dialogue with the direction of the faculty.

Many places 🇫🇷
Sep. 20 – May 21

National organisation committee of the french tournament of young mathematicians, TFJM²
Coordination of 9 events for 600 participants, communication, and creation of a new online infrastructure.

Awards & Extra

Bruxelles 🇧🇪
February 2024

★ **1st place in the hackathon the "Digital Services Act RAG Race"**
Creation of a Q&A system for questions on the DSA based on open-source models, in a team of 3, during a 7 hours hackathon organised by the PEReN and the European Commission.

Earth 🌍
2014 – present

Game development, tool design, websites (ddorn.fr/showcase)
Creation of 10+ small games under strong time constraints and pressure for game jams, a 2D EsoLang (Asciidots), multiple software tools and websites. Teamwork and sprints.

Hard Skills

★ **Python** (pytorch, huggingface, streamlit, click, mypy, pytest...)	*6000h*
JavaScript / CSS / HTML (VueJS, TailwindCSS)	*500h*
Rust, C++, Scala, LaTeX	*300h each*
System Administration (Git, Docker, Bash, remote machines...)	*200h*

Soft Skills

- Training in Non-Violent Communication
- Public speaking
- Native in French (C2)
- Fluent in English (C1)