Diego DORN

Research Engineer

cv@ddorn.fr

cozyfractal.com

github.com/ddorn

Diego finishes his master in August 2024, with ~1 year of professional expertise in **software engineering** and **teaching**, especially in the mitigation of **systemic risks** from **artificial intelligence** systems.

Work Experience

Paris — *Feb. 2024 – present*
Research engineer at EffiSciences
Automated supervision of LLM-agents, design of a benchmark to evaluate detection of out-of-distribution failure modes by monitoring systems.

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.
"Goal Misgeneralization as Implicit Goal Conditioning" *in the GCRL workshop at Neurips 2023*

Berlin — *August 2023*
Teacher at ML4Good, a summer school on AI risk
Delivery and improvement of 10 days of technical and conceptual content. 21 participants.

Lausanne — *2022 – 2023*
Lead developer for SPRIG (`sprigproofs.org`)
Developing a distributed platform to increase confidence in mathematical proofs.

Lausanne — *2019 - 2021*
Teaching assistant at EPFL
TA for 8 courses for 1st, 2nd and 3rd year bachelors: Analysis (real, vectorial, complex), C++, mathematical logic, computer science basics.

Game development & small projects (`cozyfractal.com/showcase`)
Creation of 10+ small games under strong time constraints for jams, a 2D EsoLang (Asciidots)...

Volunteering

Lausanne — *2022 – 2024*
Founder of the Safe AI Lausanne student association
Events on reducing systemic and catastrophic risks from AI. Organisation of a 10-day bootcamp, talks and a reading group. Moderation of two round table discussions.

Lausanne — *2021 – 2023*
Co-founder of Chocopoly, the hot chocolate association of EPFL
Followed by 400+ students, collaborated with 19 associations and served 1288L of hot chocolate.

Lausanne — *2020 – 2021*
President of CQFD
The association of mathematics students of EPFL. Management of a team of 14 people.

Many places — *2020 – 2021*
Member of the national organisation committee of the TFJM²
The french tournament of young mathematicians. Coordination of 9 events across France, development of a new online infrastructure and communication.

Education

Lausanne — *2021 – present*
Master's at EPFL in Communication Systems, minor in Mathematics
Focus on artificial intelligence, formal verification and complexity theory.

Interlaken — *July 2023*
Summer school "Science and Policy – How to bridge the gap?"
Topics: science for policy, science communication, open science, Swiss policy landscape.

London — *May – June 2023*
ARENA, Alignment Research Engineer Accelerator
6 weeks intensive training on interpretability, RL and training at scale.

Lausanne — *2021 – 2022*
Semester research projects in Mathematical Logic and Game Theory
Guided research under Jacques Duparc's supervision
"Infinite games in the Baire space", *Bachelor thesis, Spring 2021*
"Between decidable logics: ω-automata and infinite games", *Master's semester project, Spring 2022*

Lausanne — *2018 – 2021*
Bachelor's in Mathematics at EPFL
Passed with a 5.42/6 average and top 5/100 of my year.

Skills

- **Programming** Main hobby for the 10 last years. Many projects can be seen at `cozyfractal.com/showcase`
  - **Python (6000h)** Some of the modules I enjoyed using in more than 2 projects each include: `asyncio`, `click`, `einops`, `fastAPI`, `flask`, `huggingface`, `jaxtyping`, `joblib`, `kivy`, `matplotlib`, `moderngl`, `mypy`, `numba`, `numpy`, `pillow`, `plotly`, `poetry`, `pre-commit`, `pygame`, `pytest`, `pytorch`, `stable_baselines3`, `streamlit`, `transformer_lens`, `typeguard`,
  - **Rust (300h)**, **Scala (200h)** and **C/C++ (300h)**
  - **JavaScript / CSS / HTML (500h)** Also using, VueJS, TailwindCSS, typescript
  - **Other languages** LaTeX (200h), Typst, 6502/NES assembly, Haskell, Matlab, Lean
  - **Tools** Vim, Jetbrains IDEs, VS Code, git, Docker, slurm, runAI, inkscape, OBS, Google Suite, ArchLinux (i3wm/sway)...
- **Collaboration:** Non-violent communication
- **Languages:** French (native), English (fluent)