Clément DUMAS

Paris, France

✉ clement.dumas@ens-paris-saclay.fr

📄 linkedin/Clément Dumas | github.com/Butanium | Personal page

Education 📚

- 2024-2025: Computer Science research Master MVA (Mathematics, Vision, Learning) at **ENS Paris-Saclay**. The ENS is a selective institution that trains teachers and researchers
- 2023-2024: Computer Science research Master MPRI at **ENS Paris-Saclay**.
- 2022-2023: Double Bachelor's degree in Computer Science at **ENS Paris-Saclay**
- 2020-2022: Completed "classes préparatoires", an intensive two-year programme in the sciences with 12 hours of math per week, preparing for the competitive entrance exams to the ENS
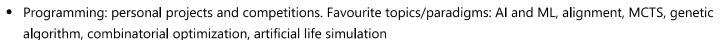
Skills 💻

- Proficient with functional (`OCaml`), imperative (`Python`) and object-oriented (`Java`) languages
- Some experience with: `Rust`, `CUDA`, `C++`, `C`, `C#`, `JavaScript`, `x86_64 Assembly`, `Scala`, `Haskell`, `Lisp`
- Proficient in **PyTorch**, **Stable Baselines 3**, **LaTeX**, **SLURM** and **Git**
- Strong mathematical and theoretical computer science background

Research and projects 🖱️

- January 2025: Incoming MATS scholar in Neel Nanda's stream
- July 2024: Attended the Human-aligned AI Summer School
- Summer 2024: 5-month research internship at EPFL with Robert West and Chris Wendler. Our work, Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers , was spotlight at **ICML 2024 mechanistic interpretability workshop**
- January 2024: Explored the emergence of XOR features in Large Language Models and the RAX hypothesis developed by Sam Marks. See our fork of the repository
- October 2023 - May 2024: Supervised Program for Alignment Research (SPAR) under the supervision of Walter Laurito. We tried to apply Contrast-Consistent Search to Reinforcement Learning models
- Summer 2023: Two months research internship with Jobst Heitzig on Aspiration-Based Q-Learning. See our LessWrong post and our Stable Baselines 3 fork
- 2022-2023: Participated in "Séminaire Turing", an AI alignment reading group at ENS Paris-Saclay
- December 2022: Participated in the AI testing hackathon organized by Esben Kran. Our submission about Trojans in transformers was ranked #4
- November 2022: Participated in the ML4G, a one-week French AI alignment camp organized by Effisciences
- November 2022: Participated in the Interpretability hackathon organized by Esben Kran featuring Neel Nanda
- October 2022: Participated in the AI alignment Hackathon organized by EffiSciences about the out of distribution and underspecification problems
- 2021-2022: Implemented a Monte-Carlo tree search for the travelling salesman problem which expand this paper to include local search in playouts
- 2021-2022: Created various heuristics for 6 CodinGame multiplayer games in Python and OCaml

Hobbies ♟️

- Programming: personal projects and competitions. Favourite topics/paradigms: AI and ML, alignment, MCTS, genetic algorithm, combinatorial optimization, artificial life simulation
- Behavioural biology (thanks to the online course Human Behavioural Biology, by Robert Sapolsky)
- Improv theater, escape games, board games, table-top and live action role-playing games
- Reading (particularly heroic fantasy and science fiction)
- Sport - badminton and tennis competitions, volleyball, football, team sports in general

Referees

Robert West

Associate Professor and Head of the Data Science Lab

Ecole Polytechnique Fédérale de Lausanne

dlab.epfl.ch/people/west

Chris Wendler

Postdoctoral Researcher at the Data Science Lab

Ecole Polytechnique Fédérale de Lausanne

wendlerc.github.io

Jobst Heitzig

Leader, FutureLab on Game Theory and Networks of Interacting Agents

Potsdam Institute for Climate Impact Research

www.pik-potsdam.de/members/heitzig

Walter Laurito

Research Engineer and Team Lead

Cadenza Lab

www.linkedin.com/in/walter-laurito-951565144

Matthias Fuegger

Head of the Distributed computing group

Formal Methods Laboratory

www.lsv.fr/~mfuegger

Charbel-Raphaël Segerie

Executive Director

CeSIA

crsegerie.github.io