

Neel Nanda's MATS 7.0 scholar with ICML 2024 mechanistic interpretability workshop spotlight publication looking for an AI safety researcher position.

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

Research Experience

💡: lead to a first-author publication

- Since January 2025 💡: MATS 7.0 scholar in Neel Nanda's stream [working on model diffing using methods like crosscoders](#)
- 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**
- July 2024: Attended the [Human-aligned AI Summer School](#)
- 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 Apart Research. [Our submission about Trojans in transformers](#) was ranked #4
- November 2022: Participated in the [Interpretability hackathon](#) organized by Apart Research
- November 2022: Participated in the [ML4G](#), a one-week French AI alignment camp organized by [EffSciences](#)
- 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

Programming Projects

- Developed mechanistic interpretability tooling: [nnterp](#), a wrapper around [NNsight](#) focused on LLMs and [tiny-dashboard](#), a tool to visualize activations of sparse dictionaries
- Early adopter of the mechanistic interpretability library [NNsight](#), actively engaging with the community to answer questions and improve it
- Developed a [CodinGame multiplayer game](#)
- Ranked in top percentiles in CodinGame multiplayer bot programming contests: top [0.5%](#), [3%](#) and [7%](#) in 2021-2022
- Proficient in OCaml, Python, Java, PyTorch, and NNsight, with working knowledge of Rust, CUDA, C++ and others

Referees

Neel Nanda

Mechanistic Interpretability Team Lead
DeepMind
[neelnanda27\[at-symbol\]gmail.com](mailto:neelnanda27[at-symbol]gmail.com)
www.neelnanda.io

Robert West

Associate Professor and Head of the Data Science Lab
Ecole Polytechnique Fédérale de Lausanne
[robert.west\[at-symbol\]epfl.ch](mailto:robert.west[at-symbol]epfl.ch)
dlab.epfl.ch/people/west

Chris Wendler

Postdoctoral Researcher at the Data Science Lab
Ecole Polytechnique Fédérale de Lausanne
[chris.wendler\[at-symbol\]epfl.ch](mailto:chris.wendler[at-symbol]epfl.ch)
wendlerc.github.io

Jobst Heitzig

Leader, FutureLab on Game Theory and Networks of Interacting Agents
Potsdam Institute for Climate Impact Research
[jobst.heitzig\[at-symbol\]pik-potsdam.de](mailto:jobst.heitzig[at-symbol]pik-potsdam.de)
www.pik-potsdam.de/members/heitzig

Walter Laurito

Research Engineer and Team Lead
Cadenza Lab
[lauritowal\[at-symbol\]yahoo.com](mailto:lauritowal[at-symbol]yahoo.com)
www.linkedin.com/in/walter-laurito-951565144

Matthias Fuegger

Head of the Distributed computing group
Formal Methods Laboratory
[mfuegger\[at-symbol\]lmf.cnrs.fr](mailto:mfuegger[at-symbol]lmf.cnrs.fr)
www.lsv.fr/~mfuegger

Charbel-Raphaël Segerie

Executive Director
CeSIA
[crsegerie\[at-symbol\]gmail.com](mailto:crsegerie[at-symbol]gmail.com)
crsegerie.github.io