Thomas **Winninger**

Student in gap year

☎ (+33) 695507673 ✉ thomas.winninger@telecom-sudparis.eu 🏠 https://le-magicien-quantique.github.io/en/about Sckathach Thomas Winninger

Whoami?

Aka *the quantum warlock, the masked camel, the fanOfThermodynamics, the pipe and clouds engineer, the whale orchestra conductor*, or just **Sckathach**. I'm a french student at Télécom SudParis, and soon, a ML(4SEC) researcher!

Fond of mathematics and physics with a strong interest in computer sciences, I choose an engineering path, to be able to create things, and most importantly, break things. I started cybersecurity at Télécom SudParis, where I enjoyed crafting and toying with technology. However, I found that I needed more mathematics in my life. As I discovered AI security a bit earlier, I knew the best place for me was in the AI field: with loads of mathematics everywhere. And I can still create and destroy things! And there is also physics! And there are also graphs! That is why I am currently in a gap year, so I can tryhard mathematics and AI before my PhD.

My interests range from geopolitics to the depth of spinning blackholes, but I'm currently focused on the thing I find the craziest at the moment: **Mechanistic Interpretability**.

Experience

Research internship in AI security

Thales 2024 - Present

To start my gap year, I wanted to have an experience at the european leader of defence: Thales. The team is brillant, and I can fully focus on LLM attack and defence, which is exactly what I was looking for. I can even talk quantum telecommunications and graph theory, the dream...

Head of training

HackademINT 2023 - 2024

HackademINT is the cybersecurity club at Télécom SudParis. It's where I spend most of my time, because I have to train myself, organise competitions, but also train others! For example, I give training courses on Tuesday evenings in AI attack, web attack and geopolitics; I create services on our Kubernetes infrastructure; and I create challenges for our various competitions. This club is a great opportunity for me to invest myself full-time in ambitious and rewarding projects.

Organiser of the 404 CTF

HackademINT 2024

The 404 CTF is a three-week online cybersecurity competition organised by HackademINT in partnership with Télécom SudParis, the DGSE, OVHcloud and Vivatech. This year, I'm in charge of part of the infrastructure, and I'm responsible for the artificial intelligence and quantum algorithm categories. In particular, I'm very lucky to be able to create AI challenges with INRIA, and quantum challenges with Quandela.

Projects

Exploring the use of Mechanistic Interpretability to Craft Adversarial Attacks

Sckathach/mi-gcg

Mechanistic Interpretability 2024

- Implemented an advanced version of the GCG attack on LLM with the Transformer Lens library.
- Adapted the gradient method to take into account the refusal direction as presented in the *Refusal in LLMs is mediated by a single direction* paper.

Exploring the use of Mechanistic Interpretability to Detect Poisoning Attacks

Sckathach/venusaur

Mechanistic Interpretability 2024

- Implemented a simple poisoning scheme using a federated learning framework.
- Analysed the effect of backdoors in CNN with the tools of mechanistic interpretability, (e.g., SAE, probes, activation patching, etc.)

Survey on GNN based IDS and their robustness

Sckathach/adversarial-gnn-based-ids

Adversarial attacks on GNN

- Survey on the state-of-the-art concerning GNN based IDS with a report and a presentation.

- Examples of adversarial attacks on classification GNN with the PyTorch Geometric library.

Skills

Languages	**Python**, **Ocaml**, Typst, TypeScript, Rust, Lua, C, Bash, C++
Spoken Languages	**English**, **French**, Korean
Frameworks/ Other	**PyTorch**, PyG, Docker, Podman, Kubernetes, Qiskit, Sage

Education

Master of Science in Cybersecurity

Institut polytechnique de Paris (IPP)

2024 - 2026

To finish my engineering curriculum with a flourish. I'm planning to continue with a PhD in AI Security (INRIA?? 🐫🐫🐫)

Engineering Degree

Télécom SudParis

2022 - 2026

The French equivalent for universities to become engineers or researchers. Even though Télécom SudParis offers general scientific courses in mathematics and physics, there is a focus on telecommunication sciences (information theory, signal theory with our best friend Fourier, graph theory, etc.), which is ideal to learn cybersecurity. I participated to the research curriculum which allowed me to take the time to work on adversarial attacks on graph neural networks and model poisoning (CNN). I love graphs. I was also part of the cybersecurity club and the code club.

Classe préparatoire MP*

Lycée Kléber - Strasbourg

2019 - 2022

These are preparatory classes where we focus on learning mathematics and physics during two years. I was already fond of mathematics, but now I know: Mathematics is the way. (Especially graphs and algebraic topology <3).

Online Courses

AI Safety Fundamentals - AI Alignment Course

Blue Dot

2024

The world of AI safety is extremely vast, from the ethical issues raised by generative AI to the risks of losing control of a super-intelligence, it's easy to get lost in this rapidly expanding field. Fortunately, there's Blue Dot! This association is offering 12 weeks of courses to take you on a tour of this new world. From my point of view, the most interesting part of the course is the weekly discussions in small groups. It was a very enriching experience to be able to talk to people from different countries and with completely different backgrounds. Incidentally, I'm on the list of best projects: https://aisafetyfundamentals.com/projects/.

Alignment Research Engineer Accelerator

Mechanistic Interpretability

2024

The logical follow-up to the fundamental Blue Dot courses: 4 technical modules on AI alignment. More specifically, I focus on the interpretability part (Transfomer Lens - SAE Lens), and on the reinforcement learning part (RL).

LLM Agents MOOC

UC Berkeley

2024

Hacking LLM is cool, but hacking agents is cooler :)

Deep Learning Specialization

Coursera

2023

Who didn't follow Andrew's course? His courses are a blessing to everyone who wants to learn AI.

Quantum Courses

IBM Quantum Learning

2023 - Present

If I become a teacher one day, I am very likely to take inspiration from the brillant John Watrous. His courses are comprehensible, interesting, complete, advanced, concise, just perfect. They are a very good complement to theorical quantum books.

Kubernetes Administrator

Udemy

2023

Even though I unfortunately did not have the time to finish the course, it was a super hands-on course where I learned the basis of Kubernetes and the cloud in general, which clearly saves my life today.