Zachary Levin

pmyzl5@nottingham.ac.uk · +44 7928 158 248 · github.com/LevidaLoca · Nottingham, UK

PROFILE

Mathematics undergraduate and **Kairos Pathfinder Fellow** building an AI safety initiative at the University of Nottingham. Track record of starting things from scratch and shipping them — a campus technical fellowship, a UK-focused AI governance podcast, full-stack web products, and interpretability research. Operator first, researcher second; comfortable owning a project end-to-end.

SKILLS

Technical Python · TypeScript · JavaScript · React · Angular · PostgreSQL · Firestore · Git

AI safety / ML mechanistic interpretability (PCA, SVD, steering vectors) · RLHF · Constitutional AI · sparse autoencoders · jailbreak analysis · grokking

Operating curriculum design · cohort facilitation · podcast production (record / edit / publish / market) · academic outreach · cross-org partnerships

Comms public speaking · technical writing · podcast hosting · community building

EXPERIENCE

Pathfinder Fellow | *Kairos Pathfinder Fellowship* Nottingham, UK · Aug 2025 – Present

- **Co-founded** the University of Nottingham's first student AI safety initiative — designed the curriculum, recruited the cohort, and secured institutional buy-in independently.
- **Run** a 2-hour weekly technical AI safety fellowship for ~10 undergrads, postgrads, and an interested lecturer, covering interpretability, evals, and alignment training methods.
- **Driving outreach** to the School of Mathematical Sciences to inform their AI policy and advise on resources for an upcoming module on small language models.

Podcast Director (Co-founder) | *AI Policy Pulse* London, UK · Aug 2024 – Jan 2025

- **Co-founded and shipped** a UK-focused AI governance podcast to increase Gen Z engagement with catastrophic AI risks.
- **Booked and prepped guests** by networking out of the London Initiative for Safe AI; collaborated with AI academics to plan accessible, technically grounded episodes.
- **Owned the full episode lifecycle** — editing, publication, and marketing — and ran pre-record discussions to warm up guests.
- **Researched, drafted, and edited** policy-focused questions to make UK AI governance legislation accessible to a non-specialist audience.

Technical AI Safety Researcher | *Impact Research Groups* London, UK · Feb – May 2025

- Member of a 5-person team investigating mechanistic interpretability of jailbreaks in open-source LLMs — specifically how refusal directions degrade once published jailbreak frameworks bypass refusal.
- **Owned** the long-prompt regime, which introduced unique challenges around prompt length and signal extraction.
- **Applied** PCA, SVD, and direction extraction / steering to identify behavioural shifts past the refusal threshold.
- **Worked across** remote (Nottingham) and in-person (Arcadia, London) collaboration with the team.

Course Participant | *BlueDot Impact — Technical AI Safety* Remote · Nov – Dec 2025

- Completed 30-hour technical curriculum: RLHF, Constitutional AI, data filtration, dangerous capability evaluations, alignment faking, and mechanistic interpretability via sparse autoencoders.
- **Submitted** a research proposal extending Nanda et al.'s grokking analysis — investigating Fourier-based algorithmic representations learned by transformers on modular arithmetic tasks.
- Critically evaluated readings and research directions in weekly facilitated sessions with a cohort of ~10.

Course Participant | *BlueDot Impact — AI Alignment* Remote · Jul – Aug 2024

- **Co-authored** a project on the effect of online AI-generated misinformation on LLM predictions.
- Fed alignment thinking back into editorial direction for AI Policy Pulse.

R&D Full-Stack Intern | *Imprint Social* Tel Aviv, Israel · Feb – May 2022

- **Self-taught TypeScript** and integrated into full-stack sprints using Angular and PostgreSQL.
- **Shipped** database-driven components and resolved cross-stack issues with minimal QA revisions.
- Managed independent task delivery via Git (GitBucket) and contributed to product design discussions.

Volunteer Full-Stack Developer | *Nefesh B'Nefesh* Tel Aviv, Israel · Jan – Feb 2022

- **Led** ReactJS + Google Firestore web app build for a non-profit; owned multiple pages, user authentication, and Firestore integration end-to-end.

EDUCATION

BSc Mathematics · University of Nottingham Sep 2022 – Present

A-Levels & GCSEs · City of London School 2011 – 2021

SELECTED ACTIVITIES

- **Global Challenges Project** — attended Oxford retreat on existential risk and AI policy; built ongoing relationships in the UK AI safety community.
- **Senior Non-Commissioned Officer, Army Cadets** — planned and led training activities for junior cadets.
- **Weekly volunteer** preparing and serving meals to students at a local Nottingham charity.
- **Self-directed learning** from Karpathy's nanoGPT/nanoLLM series onward; regular practice on Codewars and LeetCode.