Boyd Kane

Research Scientist - AI Safety

+27 67 322 7353
UK & South African Citizenship
Berkeley, California
boydrkane@gmail.com

boydkane.com
boydrkane
beyarkay
Anonymous Feedback

Experience

MATS 9.0 with Alex Turner & Alex Cloud

Jan 2026 - Apr 2026
Berkeley, California

- The project by my coauthor Jo Jiao & I was accepted as one of 9 MATS Symposium Spotlight talks out of a cohort of 100 fellows, and will be continued during the MATS 9 Extension in London to prepare the project for submission to NeurIPS.
- LLMs behave differently in evaluations than they do when we're not watching them. It's possible that an LLM would behave misaligned in some very narrow range of scenarios (e.g. when it could exfiltrate its weights *and* nobody's watching *and* it has internet access *and* ...).
- Alex Turner (Google DeepMind) and Alex Cloud (Anthropic) are my mentors as I work on a project to distinguish between LLMs that would exhibit this behaviour and LLMs that won't.
- This project uses finetuning as a method of evaluation: by measuring how easily an untrusted LLM is able to learn some misaligned behaviour, we can get an estimate for how likely the unfinetuned LLM is to behave misaligned.

AI Safety South Africa

Jun 2024 - present
Expanding AI Safety in RSA

AI Safety | Community building | Organisation

- AI Safety South Africa (AISSA) runs weekly paper discussions and meetups to discuss the latest AI safety research.
- Through regular attendance and engagement I am one of the most active members and regularly contribute discussion points.
- With input and funding from AISSA, I have started another chapter in Stellenbosch, a local university town.
- AI Safety Stellenbosch has similar goals to AISSA, but hosting events in Stellenbosch (about 40 minutes drive from Cape Town) enables the discussions to reach further and engage new people.

CubeSpace Satellites

Mar 2024 - present
Satellite Embedded Systems Engineer

Embedded C | MISRA | Self-directed innovation

- At CubeSpace, I develop and maintain all firmware and software required to control satellites in orbit, as well as various auxillary ground-based software systems.
- I developed several key systems used in orbit, including the NAND-flash file system for storing firmware on board and aspects of the GNSS parser.
- I lead the effort to reimagine our ground supply equipment to make it more powerful for both our customers and our production department. This involved writing all the bare-metal firmware for an STM32F7-based product that allows the user to send/receive telecommands and telemetry to any CubeSpace product across a wide range of communications protocols. Significant complexity in this product was introduced because of the years of custom solutions and mutually exclusive hardware options that I brought under one easy-to-use device.

Condor Camp South Africa

Oct 2024
10-day AI Safety Retreat

Mech. Interp. | ARENA | AI Safety

- I was accepted as a participant at the South African Condor Camp, a 10-day retreat hosted at Zevenwacht wine estate near Stellenbosch, South Africa and funded by Open Philanthropy.
- The retreat featured AI technical alignment & AI governance master classes, 1-on-1 with various international speakers, project sprints for working on AIS related work, and guided debates with other participants about policy and technical AIS challenges

Other experience:

AWS SDE Intern Dec 2022 - Feb 2023 *Amazon Web Services* — Error discovery | High availability

AWS SDE Intern Dec 2021 - Feb 2022 *Amazon Web Services* — Microservices | High availability

Deep Learning Indaba X Jul 2025 *Poster Judge and Speaker* — Deep Learning Conference | AI Safety | Speaker

Talk about the economics of AI Jun 2025 *Speaker to 70+ prominent ZA business people* — Public Speaking | AI

Cooperative-AI in-person course Apr 2025 - Jun 2025 *Participant (12-week course) with essay output* — AI Safety | Course

Projects

Factorion

Apr 2025 - present
RL agent to beat the videogame Factorio

RL PPO

- Training an RL agent (PPO) to autonomously design factories in a Factorio-inspired grid world by placing belts, inserters, and machines to maximise throughput
- Multi-component reward function and curriculum learning to combat reward sparsity and reduce reward hacking
- Constraint validation system enforcing safe exploration across a combinatorial action space

ListenToAnything.com

Oct 2024 - present
Website-to-podcast converter

Rust Text To Speech Profitable

- Listen To Anything allows users to convert any blogpost, article, or PDF into a podcast episode which gets automatically uploaded as a podcast episode to a custom podcast feed.
- They can then subscribe to this feed with their podcast app, and listen to their reading list.
- I created the website after seeing 1) my reading list grow, not shrink and 2) that I never had enough quality podcasts to listen to.
- I use a high-quality text-to-speech algorithm so the voice is *very* good.
- Give it a go! It takes 2 minutes and I promise you'll be impressed by how easy the voice is to listen to.

Eskom Calendar

Jul 2022 - May 2024
Open source software

Rust CI/CD Web Servers FOSS

- Eskom Calendar provides free and open-source loadshedding schedules to everyone.
- It has received national media attention, over 60k weekly downloads at peak, and more than 190 stars on GitHub.
- The name of the project comes from the free internet calendars which put loadshedding events in your digital calendar, so that you can schedule meetings and work shifts around the ever-changing loadshedding schedules.
- I have written a detailed writeup on my blog: https://boydkane.com/projects/eskom-calendar

Other projects:

Attractors Nov 2021 - May 2022 *Strange attractors as art* Rust Computer Art Interactive Chaos

Conway's Game of Life in Hardware, without any code Jun 2025 - present PCB design Hardware digital logic

Education

MSc Computer Science (by dissertation)

Feb 2023 - Nov 2023
Stellenbosch University, ZA

- I graduated with distinction
- I built gloves that let you type without a keyboard. Hardware sensors measure the acceleration of your fingertips. Machine learning converts those measurements into keystrokes
- 51 different gestures could be recognised. Everything was designed and built by me: the hardware, the dataset, the machine learning code, and the software to control the device
- The thesis was completed in one year. You can watch the device being used here

BSc Hons. in Computer Science

Jan 2022 - Dec 2022
Stellenbosch University, ZA

- I graduated Cum Laude.
- My honours thesis provided the core of what would become my Master's thesis.
- I also took several courses: Functional Programming, Advanced Algorithms, Machine Learning, Artificial Intelligence, and Data Science

BSc in Computer Science and Mathematical Statistics

Jan 2019 - Dec 2021
University of Cape Town, ZA

Math. Stats. Comp. Sci.

- I took courses in the theory of algorithms, the theory of computation, probability theory, statistical inference, the theoretical basis for machine learning, Spanish, and business finance.
- I was twice awarded a merit-based scholarship valued at R20 000 (about 25% of a year's tuition), first in 2020, then again in 2021.