Brendon Chikavanga

WORK EXPERIENCE

AI SAFETY SOUTH AFRICA | Research Engineer (Intern)
January 2026 – April 2026 · ongoing collaboration

- Wrote an evaluation contributing to Sam Brown's Phase-2 work on *Precursors, Proxies and Predictive Models for Long-Horizon Tasks* (NeurIPS 2025) – a stochastic, non-stationary variant of the Lights Out game. Worked independently and reported to Sam.
- Lead author on an independent research project, *"Some LLMs Drop Their Own Task When Pressured by Another Agent"* (paper in draft, sole author). Designed the experiment, wrote the orchestration and analysis code, and ran it across 7 frontier models. Headline result: task abandonment under peer pressure varies from **0% (Anthropic) to 30.5% (Gemini 3 Flash)**; a single-line system-prompt warning eliminates the worst case (30.5% → 0%); susceptibility is a property of the target, not the source of pressure.
- Continuing as a member of the AISA evals team alongside Sam Brown, Leo Hayams, and Jaco du Toit.

TRIXTA (FORMERLY REBEL OS LABS) | Senior Software Engineer
February 2023 – Present

- ∼3 years at the same company through the Rebel OS Labs → trixta.ai rebrand; senior engineer, primarily backend.
- Ship across the product line: trixtaOS (business OS), trixtaDecloud (network infrastructure), Trixta Studio (low-code IDE).
- Work in Trixta's in-house Elixir-based platform language; TypeScript / React on the front-end; CLI tooling and Claude Agent SDK integration on the AI side.

FEATURED RESEARCH

SOME LLMS DROP THEIR OWN TASK WHEN PRESSURED BY ANOTHER AGENT
First Author | In Draft, May 2026 – under general mentorship of Benjamin Sturgeon

Controlled measurement of inter-agent task drop-out across 7 frontier models, with a one-line defensive prompt that resolves the worst case. Targeting workshop submission.

STEERING A LANGUAGE MODEL BY EDITING ITS THOUGHT ANCHORS
Independent project | BlueDot Technical AI Safety Course capstone, 2025

brend0nc.substack.com/p/steering-a-language-model-by-editing

AI SAFETY ENGAGEMENT

- BlueDot Impact – **Technical AI Safety Course**, certified (2025).
- AISA – team member and event volunteer; reading groups and community events.
- Self-study: working through Neel Nanda's mech-interp researcher roadmap.

EDUCATION

UNIVERSITY OF SOUTH AFRICA
B.Sc. Applied Mathematics
cum laude, 2020 – 2023

B.Sc. Computer Science
cum laude, 2019 – 2023

Honours, Cosmology
commenced; not completed (pivoted to AI safety)

TECHNOLOGY

Languages
Python • Elixir • C++ • TypeScript

AI Safety / ML
Inspect (UK AISI eval framework) • PyTorch • Claude Agent SDK • LLM-as-judge eval pipelines

REFERENCES

Benjamin Sturgeon
General research mentor
benjaminsturgeon.com
LinkedIn

Sam Brown
AISA research lead; oversaw the *Precursors, Proxies and Predictive Models* evaluation work.
LinkedIn

LINKS

LinkedIn
Substack