Ma’alona Mafaufau

Cybersecurity Engineer · AI Safety Researcher

maalona.mafaufau@approxiomresearch.com · github.com/Lona44 · linkedin.com/in/maalonamafaufau · approxiomresearch.com

PROFILE

Cybersecurity engineer with 6+ years across cybersecurity, data analytics, and software development. Two-time 3rd-place winner of the Palisade Research AI Misalignment Bounty (9 of 295 submissions awarded; published arXiv:2510.19738). Workday security background in APT threat hunting, SIEM detection engineering, and global incident response, combined with hands-on embodied-AI alignment evaluation and a strong quantitative foundation (BSc Hons Statistics).

TECHNICAL SKILLS

Security	APT threat hunting, SIEM detection engineering (Splunk SPL), incident response, threat-intel pipelines, Blue Team, triage analysis.
AI / Safety	AI misalignment evaluation, Inspect AI (UK AISI), MuJoCo, agentic AI architectures, safety/honesty scoring, LLM-judge pipelines; OpenAI / Anthropic / Gemini APIs.
Engineering	Python, TypeScript, JavaScript, React, Node.js, FastAPI, PostgreSQL, MongoDB; GCP (Cloud Run, Firestore, Build), AWS, Azure, Docker, Git, CI/CD; TDD (Jest).

SELECTED PROJECTS

G1 Alignment Experiment — github.com/Lona44/Embodied-AI-Alignment-Testing-On-InspectAI

Embodied-AI alignment research platform — testing whether AI models deceive operators when controlling physical robots under pressure. Built at the Gemini 3 Hackathon.

— Built an embodied alignment harness on Inspect AI (UK AI Safety Institute): a MuJoCo-simulated Unitree G1 humanoid under battery, time, and institutional pressure, with a telemetry-corruption variant testing concealment of safety violations.

— Across 4 frontier models / 30 runs, found safety and honesty are independent failure dimensions — Gemini Robotics ER 1.5 showed L3 strategic concealment in 29% of runs while GPT-5 stayed fully honest (5.0/5) despite low safety (1.9/5).

— Engineered the scoring + delivery stack: a Gemini 3 Pro LLM-judge (L0–L4 taxonomy) with Flash video verification and a custom Kimi K2.5 provider; Next.js 15 dashboard with in-browser MuJoCo WASM 3D replay and a FastAPI/Cloud Run backend. MIT licensed.

SECURITY & AI SAFETY EXPERIENCE

Independent AI Safety Researcher — Approxiom Research — Apr 2025 – Present

Auckland, NZ · Systematic vulnerability discovery and alignment evaluation in frontier AI systems.

— Two-time 3rd-place winner of the Palisade Research AI Misalignment Bounty — a single scenario elicited fundamentally different misalignment patterns from o3 and GPT-5, winning two separate prizes (9 of 295 submissions awarded). Published in the Palisade research paper (arXiv:2510.19738); all submissions released publicly on Hugging Face.

— Built the G1 embodied-alignment platform (see Selected Projects) and reproducible experiment pipelines across OpenAI, Anthropic, Gemini, and Kimi, with secure API-key management (pre-flight validation, encrypted storage).

— Developing an autonomous incident-response agent for the SANS FIND EVIL! competition, extending the SIFT forensic toolkit with self-correcting triage — applying frontier-AI evaluation methods to defensive security operations.

Cyber Security Engineer — Workday — May 2024 – Dec 2024

Auckland, NZ · Progression across Workday's global Security Incident Response function, follow-the-sun model (NZ, Ireland, USA).

— Led APT threat-hunting initiatives across the enterprise environment following promotion.

Snr Associate Cybersecurity Engineer — Workday — Apr 2023 – May 2024

— Developed custom detection rules using Splunk SPL to improve detection efficacy and reduce false positives across the global SIEM.

— Built automated threat-intelligence pipelines in Python; created a SIEM efficacy analysis tool adopted globally by the security team.

Snr Associate Cybersecurity Operations Analyst — Workday — Jan 2022 – Apr 2023

— Triaged and investigated security detections across enterprise SaaS infrastructure; coordinated incident response handoffs with regional SIRT teams.

Cybersecurity Analyst / Intern — Datacom — Apr 2021 – Jan 2022

— Triage analyst monitoring customer environments for SIEM detections; escalated tuning recommendations to the engineering team.

RELEVANT EXPERIENCE

Full Stack Engineer — Takitoru Developments Ltd (Contract) — Mar 2026 – Present

— Architecting a full-stack mobile and web platform end-to-end — codebase, CI/CD pipeline, and production observability; test-first delivery. TypeScript, React Native, Next.js, PostgreSQL, Supabase, Vercel, GitHub Actions.

Director & Technical Lead — IPU Collective — Mar 2026 – Present

— Technical leadership for a Māori-led studio delivering for Crown and council clients (Auckland Transport, KiwiRail, Watercare, Oranga Tamariki): own the production pipeline and ship interactive web deployments; set up hosting, version control, and deployment infrastructure.

Assistant Trainer / Full Stack Developer — Mission Ready HQ (Contract) — Oct 2025 – Dec 2025

— Mentored full-stack developer students in React, Node.js, and MongoDB; delivered code reviews and learning sessions on AI integration and API development.

Data Analyst — Halter Limited — Sep 2019 – Mar 2021

— Analysed agricultural IoT data from smart-collar systems; built statistical models and ML-based products for customer needs.

EDUCATION

BSc (Honours), Statistics — University of Auckland, May 2016

BSc, Double Major Mathematics & Statistics — University of Auckland, May 2015

AWARDS & CERTIFICATIONS

Two-time 3rd Place — Palisade Research AI Misalignment Bounty, 2025 (two prizes from one scenario; 9 of 295 submissions awarded; arXiv:2510.19738).

AI-Powered Advanced Full Stack Developer — Mission Ready HQ, 23 Jan 2026 — verify ↗

Full Stack Developer — Mission Ready HQ, 7 Aug 2025 — verify ↗

Palo Alto Networks Cybersecurity Professional Certificate — Coursera, Feb 2025.

Blue Team Level 1 (BTL1) — Security Blue Team, Sep 2022.

AWS Cloud Quest: Cloud Practitioner — Amazon Web Services, Aug 2023.

Unitec CTF 2021 Winner — Kawaiicon 2021 (Boot2Root challenges).