Adrian Wedd

AI Safety Engineer & Independent Researcher

Cygnet, Tasmania, Australia

GitHub LinkedIn Email PDF Print Watch Me Work

ABOUT

Systems builder, AI safety researcher, and adversarial thinker with nearly 45 years across the stack. The failure-first methodology didn't come from papers — it came from coordinating direct actions for Greenpeace's Actions unit against well-resourced opponents where getting the risk assessment wrong had real consequences. That operational security instinct runs through everything: seven years leading cybersecurity, penetration testing, and IDAM for Tasmania's public housing sector; three years of empirical AI red-teaming across 120+ models and 18,000+ adversarial prompts; and production AI systems with hallucination detection and content safety gates built in from the start. Currently freelancing across client delivery (healthcare ecosystems, multilingual sites, ops tooling) and independent AI safety research. AuDHD — the hyperfocus and pattern recognition are features, not bugs. Most interested in where complex systems break and what that reveals about how they were built.

CORE COMPETENCIES

AI Safety & Evaluation

Red-teaming, adversarial testing, failure-first methodology

Frontier AI Models

Claude API, multi-agent systems, evaluation frameworks

Risk Assessment

Pre-mortem analysis, FMEA, failure mode identification

Policy Translation

Technical findings to actionable governance for decision-makers

EXPERIENCE

Freelance Developer & AI Consultant

Independent 2025 - Present

Delivering client web projects and independent AI safety research. Specialising in red-teaming, adversarial testing, evaluation frameworks, and failure-first methodology for frontier AI models. Building multi-agent systems, LLM tooling, and open-source AI safety infrastructure.

- Designed and built red-teaming and adversarial testing framework for embodied AI with automated attack adapters (TAP) and cross-modal vulnerability research on VLA models
- Developed evaluation frameworks with hallucination detection, content safety gates, and model monitoring for production LLM pipelines
- Applied prompt engineering, RAG, and retrieval-augmented generation across multiple agentic systems
- Delivered four-site healthcare digital presence with AHPRA regulatory compliance and REST API integration
- Built 10+ active open-source repositories: agent orchestration, AI governance, NLP tooling, and CLI automation
- Developed multi-agent orchestration system with LangGraph, LangChain, and Anthropic SDK

`Python, JavaScript, TypeScript, LangChain, LangGraph, Anthropic SDK, Claude API, Prompt Engineering, RAG, Red-teaming, Adversarial Testing, Evaluation Framework, Hallucination Detection, Content Safety, Multi-Agent Systems, GitHub Actions, Docker, FastAPI, NLP`

Systems Analyst / Acting Senior Change Analyst

Homes Tasmania (formerly Department of Communities Tasmania)

2018 - Feb 2026

Led systems integration, cybersecurity, and digital transformation for Tasmania's public housing sector. Delivered penetration testing, vulnerability assessment, and identity and access management (IDAM) improvements. Deployed automation and AI tooling to improve service delivery for vulnerable communities.

- Designed and implemented RESTful API integrations and SFTP data exchange for the Housing Management System
- Led cybersecurity program including vulnerability assessment, penetration testing, and Essential Eight compliance uplift
- Implemented identity and access management (IDAM) controls reducing unauthorised access risk
- Built Python and PowerShell automation scripts for operational workflows, reducing manual processing time
- Pioneered responsible AI adoption for data analysis and decision support, including prompt engineering and LLM integration
- Developed ISO 27001-aligned information security practices across the department

`Python, PowerShell, JavaScript, RESTful APIs, SFTP, SQL, Azure, Penetration Testing, IDAM, Essential Eight, ISO 27001, Prompt Engineering`

ITS Client Services Officer

University of Tasmania 2015 - 2018

IT support across the full university user base — researchers, academics, administrative staff. Exposure to an unusually wide range of systems and problems built the diagnostic instinct that carries through everything since.

- Sole support contact for a diverse user base with no tolerance for downtime — learned to triage fast and fix faster
- Built and maintained procedural documentation that actually got used, reducing repeat escalations
- Managed Microsoft 365 and Active Directory environments across a complex multi-faculty structure

- Handled ServiceNow ticketing and escalation paths for infrastructure issues beyond first-line scope

`Microsoft 365, Windows Server, Active Directory, ITIL, ServiceNow`

Director

Digital Agency PTY LTD 2015 - 2018

Ran a small digital agency focused on nonprofits and small businesses — Google Grants, AdWords, Analytics, and web presence. Learned quickly that clients don't buy strategy; they buy outcomes they can point to.

- Built and managed Google Ad Grants campaigns for nonprofits, maximising the $10k/month allowance to drive real audience growth
- Set up Analytics and conversion tracking from scratch for clients with no prior measurement — made their spend legible
- Delivered web and campaign work across diverse sectors: environmental orgs, health practitioners, retail
- Managed client relationships end-to-end: scoping, delivery, reporting, iteration

`Google Analytics, Google AdWords, Bing Ads, Digital Marketing, Campaign Management`

Second Level IT Support Engineer

The Wilderness Society Inc. 2012 - 2015

End-to-end IT infrastructure ownership for a national environmental organisation with distributed operations across Australia. Responsible for all server environments, network infrastructure, and communication systems — the kind of breadth that builds genuine systems thinking.

- Owned lifecycle management of 60+ Windows/Debian/Ubuntu servers across a nationally distributed organisation
- Maintained heterogeneous network infrastructure with no dedicated team — full-stack from physical layer to application

- Built and administered Google Apps and VOIP PBX infrastructure replacing legacy systems
- Executed complex migrations and decommissioning with zero data loss across diverse environments
- Delivered ICT capability to teams with no technical background, building organisational resilience

`Windows Server, Debian/Ubuntu, Network Infrastructure, Google Apps, VOIP PBX, Server Management`

Actions Unit — Communications & Logistics Coordinator

Greenpeace Australia Pacific 2010 - 2012

Operational planning and execution for Greenpeace Australia Pacific's direct action campaigns. The Actions unit operates against well-resourced opponents — corporations and state actors who actively work to prevent campaign success. Work required genuine operational security, adversarial threat modeling, and risk assessment under conditions where failure meant people getting hurt or arrested.

- Planned and coordinated direct action operations — logistics, communications, and security across multiple simultaneous campaigns
- Developed threat models and operational security protocols against adversarial opponents with surveillance and legal resources
- Designed and maintained field ICT infrastructure for covert operations: encrypted comms, dead drops, secure channels under active counter-surveillance
- Conducted pre-action risk assessments enumerating failure modes, contingency routes, and abort criteria
- Trained activists in operational security and secure communications under pressure
- Coordinated multi-team logistics across geographically distributed operations with real-time adaptation to adversarial conditions

`Operational Security, Threat Modeling, Risk Assessment, Encrypted Communications,` Field ICT, Logistics Coordination, Adversarial Planning

PROJECTS

This Wasn't in the Brochure

Practical, neurodiversity-affirming guide for co-parents raising children with ADHD, Autism, PDA, and ODD. The first book written specifically for the co-parenting context. Four localised editions (AU, US, UK, NZ) with culturally-appropriate legal, medical, and educational systems. Companion children's book in development.

`Technical Writing` `Neurodiversity` `Research Synthesis`

NotebookLM Automation

GitHub

CLI automation toolkit for Google NotebookLM with research quality controls, deterministic exports, template engine, JSON schema validation, and comprehensive integration testing.

`Bash` `Python` `CI/CD` `JSON Schema`

Orbitr

GitHub

Multi-track polyphonic AI sequencer with concentric ring interface. Inspired by Playtronica's Orbita, enhanced with Meta's MusicGen for real-time sample generation. 36 issues tracked across phased delivery.

`HTML/CSS/JS` `MusicGen` `Web Audio API`

Evolve Evolution

Four-site digital presence for a healthcare practitioner: AHPRA-compliant chiropractic clinic site, coaching/personal brand platform, author hub, and business evolution site. Shared CSS design token system, Cliniko API integration for appointment management, GA4-to-GTM analytics migration, Cloudflare Pages deployment with serverless functions, and quarterly regulatory compliance workflow. 170+ issues tracked across phased delivery.

HTML/CSS/JS Cloudflare Pages Cliniko API GA4/GTM Structured Data

Tanda Pizza

Static website for a Bali-based restaurant, localized to 5 languages with WhatsApp ordering integration and gluten-free menu filtering.

HTML/CSS/JS i18n WhatsApp API

Failure-First Embodied AI

GitHub

Red-teaming and benchmarking framework for embodied and agentic AI systems. Focused on adversarial testing, recursive failure analysis, recovery evaluation, and human-in-the-loop safety. Includes computer vision vulnerability research on VLA models, neural network robustness evaluation, automated attack adapters (TAP), cross-modal failure analysis, prompt injection and jailbreak testing, bias detection, interpretability analysis, model evaluation pipelines, and AI ethics documentation.

Python AI Safety Red-teaming Adversarial Testing Computer Vision Neural Networks Model Evaluation AI Ethics Prompt Injection Bias Detection Robustness Interpretability Evaluation Framework Content Safety Responsible AI Deep Learning NotebookLM arXiv

TEL3SIS

GitHub

Real-time telephony platform with LLM-powered conversations, tool use (calendar, SMS, email), tri-layer memory (Redis/SQLite/vectors), safety oracle for pre-execution checks, and human handoff. Built on vocode-python with Whisper, GPT-4, and ElevenLabs.

Python Vocode Whisper Redis Docker

Cygnet

GitHub

Multi-agent platform coordinating land acquisition, build planning, and 3D print operations for affordable housing. 28+ specialized agents, FastAPI backend, React frontend. 795 issues tracked, 432 commits across Python, TypeScript, Kotlin, and Swift.

`Python FastAPI React TypeScript Docker`

VERITAS

GitHub

Legal intelligence platform addressing the efficiency-trust deficit in the legal market. Knowledge graph with Cypher queries, document analysis pipeline, and transparency-first AI reasoning. 230 issues, 784 commits across Python, TypeScript, and Jupyter notebooks.

`Python TypeScript Neo4j/Cypher Jupyter Docker`

ADHDo

GitHub

AI assistant for ADHD executive function support with crisis detection, circuit breaker psychology, and local-first processing. NDIS-aware for Australian users. 100 issues, 73 commits.

`Python AI Safety Accessibility`

NeuroConnect

GitHub

24/7 voice helpline with wired STT-to-LLM-to-TTS pipeline over Twilio, NDIS evidence logging, and executive function scaffolding. Multi-provider LLM fallback chain including Claude CLI.

`Python Twilio Whisper Docker`

RLM-MCP

GitHub

MCP server enabling Claude Code to process 1M+ character documents through session-based chunking, BM25 search, and artifact provenance tracking.

`Python MCP BM25 Claude Code`

Agentic Research Engine

GitHub

Next-generation multi-agent research system with LangGraph orchestration, self-correction loops, and autonomous learning. 1,001 commits across Python with CI/CD cost optimization and enterprise security integration.

`Python` `LangGraph` `Multi-Agent Systems` `Docker`

SKILLS

Programming Languages

Python	Primary
JavaScript	Primary
TypeScript	Secondary
PowerShell	Secondary

AI & Automation

LLM Integration	Primary
AI Safety & Evaluation	Primary
Process Automation	Primary
ML Libraries	Secondary
MCP Development	Secondary
Voice/Telephony AI	Secondary

Infrastructure

Systems Integration	Primary
Cybersecurity	Primary
Linux/Windows Server	Primary

DevOps

Docker	Secondary
GitHub Actions	Primary
Cloudflare Pages	Secondary

Frontend

React	Secondary

Data

SQL / PostgreSQL	Primary

ACHIEVEMENTS

Security & Systems Uplift — Homes Tasmania

Led cybersecurity and systems integration for Tasmania's public housing sector: Essential Eight compliance uplift, penetration testing, IDAM controls, and RESTful API integrations for the Housing Management System. Infrastructure that vulnerable people depend on — the stakes made the work serious.

2018-2024

AI Safety Research & Red-Teaming

Developed failure-first red-teaming framework for embodied AI systems with 104 tracked research issues, cross-modal vulnerability analysis on VLA models, and automated attack adapters. Pursuing commercial AI safety services.

2025-2026

Voice-First AI Agent Platform

Built TEL3SIS, a real-time telephony platform with LLM-powered conversations, tri-layer memory architecture, safety oracle, and tool orchestration across 226 tracked issues.

2024-2026

Technology for High-Stakes Campaigns

Maintained distributed IT infrastructure across The Wilderness Society's national operations and built field technology for Greenpeace direct actions — two very different threat models, both requiring the infrastructure to be invisible until it isn't.

2010-2015

Adrian Wedd

AI Safety Engineer & Independent Researcher

Last updated: Feb 28, 2026, 08:05 PM