

Josh Hills

AI Security Researcher

joshthmhills@gmail.com | Ontario, Canada

PROFILE

AI security research engineer/scientist specializing in dangerous capability evaluations, AI control, and red-teaming. Two years building evals at METR, UK AISI, and Redwood Research.

EXPERIENCE

Research Scientist | Astra Fellow

January 2026 - Present

Constellation | Mentored by Asa Cooper Stickland (UK AISI)

- Building Iterative VibeCoding, a persistent-state control setting testing monitor evasion strategies in agentic coding environments (COLM 2026, under review)
- Contributed human solve-time estimation methodology and validation for Think Fast (NeurIPS 2026, under review), collecting and calibrating baselines across 43 benchmarks measuring no-CoT latent reasoning

Research Engineer | Contract (via Equistamp)

May 2025 - January 2026

Redwood Research

- Built 2 LinuxArena settings (web scraping, cost estimation) end-to-end: main tasks, side tasks, Docker environments, and human trajectories (NeurIPS 2026 Datasets Track, under review)
- Implemented red-team side tasks (credential exfiltration, remote code execution, secret exposure, malicious redirects) and contributed to shared ControlArena infrastructure

Research Engineer | Contract (via Equistamp)

September 2024 - February 2025

METR (Model Evaluation & Threat Research)

- Contributed to HCAST paper: task development, QA, and benchmarking of 50+ evaluations for autonomous AI capabilities; achieved 90th percentile among human baseliners with 100+ hours of baselines
- Created one of METR's largest self-proliferation evaluations, assessing LLM agents' ability to autonomously acquire compute resources

Data Scientist Consultant | Senior Associate

May 2023 - October 2024

Technomics

- Led development of internal AI ethics and safety guidelines
- Built Azure data pipelines, delivering 40% reduction in client overhead costs

SKILLS

AI Security: Model evaluation design, dangerous capability assessment, AI control, red-teaming, human baselining

Technical: Python, Inspect, Vivaria, TensorFlow, Azure, SQL, computer vision, NLP

EDUCATION

Bachelor of Science (Honours) - Computer Science

Graduated 2023

Carleton University | Specialization: AI/ML

CERTIFICATIONS & AWARDS

- ARENA 5.0 Best Capstone (long-horizon evals)
- AI Safety Fundamentals (Bluedot)
- Azure Data Scientist Associate
- CAIS-X Competition Winner