Gavin Black

AI security research and engineering leader

Nashua, NH | Remote or hybrid (Boston area)

Director at Leidos, leading 15+ engineers and researchers in AI security, including a team with main-track papers at ICML and COLM. Turns their research into business terms stakeholders can act on, and stays hands-on in automated red teaming and fuzzing. PhD in cyber defense and first author of 5 IEEE papers.

Experience

Leidos · Remote

2020 – present

Since 2020 the work has been delivered to 4 Leidos business units and 1 energy customer's OT network, and written into 10+ proposals. It was selected in 2 competitive customer evaluations and shown in 4 demos and trials.

Director, Cyber Autonomy

Dec 2025 – present

15+ direct reports (engineers and researchers, some with PhDs). Own the roadmap, help hire, and demo to sector VPs and, at times, CTOs (Intelligence, Digital, Health & Civil).

- Conceived and prototyped a digital twin for pentesting and efficiency testing, and now lead its team. It turns a prompt into an emulated enterprise network in AWS and checks the connectivity graph against the prompt.
- Lead automated red teaming, using frontier models to attack environments on their own, build reduced digital twins, and generate incremental fuzzing harnesses that raise code coverage.
- Across projects, run and integrate the researchers' tools, generate the evaluation metrics, and test new ones.
- Direct research on sanitizing what agent connectors pass to models, conforming it to known-good distributions to cut refusals and downstream rework. The team publishes at COLM, ICML, and CAMLIS.
- Evaluate and locally deploy third-party decision models and LLM routers, measuring routing by tokens saved.
- Direct the team's fine-tuning (SFT) for specific tasks and its RAG and wiki knowledge stores.
- Write agent-based security tools, MCP servers (one for a custom AFL++ fuzzer), and demo UIs. Sandbox agents in Docker and Kubernetes on AWS and ship validated tools as containers, some hardened for Iron Bank.

Principal Research Scientist and Solutions Architect

2020 – Nov 2025

- Led AI security projects with 10+ researchers and engineers (2–3 direct reports) from proposal to prototype.
- Built a BERT traffic model (real traffic, injected real attacks) that routes decisions with confidence scores from embedding similarity to known-good and known-bad. Deployed in an energy customer's lab, and the basis of 2 IEEE papers (ICNC, FADE).
- Tested frontier models on fuzzing and program logic analysis, work now being fielded for an airspace customer.
- Architected an agentic red-teaming stack for enterprise AI that automates attacks and defenses and maps blast radius. Built software that turns recent CVEs into reusable red-team skills.
- Contributed to pre-LLM vulnerability models and datasets (BERT tuned on automatically decompiled code), including MCP-style tool interfaces to Ghidra and ddisasm before MCP existed.
- Researched weaknesses in tool-using LLMs, MCP, and RAG.
- Trained RL network attack models, and novelty-seeking systems that turn quantized system behavior into a surprisal reward (from the dissertation). Built pipelines that simulate attacks and defenses to test detections.
- At the Air & Space Forces Association Warfare Symposium (March 2025), built and gave live demos of several AI cyber capabilities and ran part of the Leidos demo booth with customers.

MITRE · Bedford, MA 2008 – 2020

The non-profit R&D center behind CVE and ATT&CK

Lead Cyber Systems Engineer (promoted from Senior)

- Managed 7+ direct reports from 2015 to 2020 (performance reviews, career direction, placement) and 10+ interns each summer. Ran the Bedford campus intern program, 20+ interns a year.
- Technical lead for modernizing the CVE backend, and development lead for Crown Jewel Analysis tooling.
- Led iMAS, MITRE's open-source iOS security project, and was the top contributor to its encrypted Core Data library (700+ GitHub stars, 200+ forks). Reverse engineered iOS protections in assembly.
- Security lead for next-generation Air Force platforms, adding auditing tools to the Platform One DevSecOps pipeline. Reverse engineered binaries and tested C/C++ code for vulnerabilities.

Lockheed Martin · Gaithersburg, MD 2006 – 2008

GPS Software Engineer. Wrote satellite payload encryption drivers and Kalman filter code for GPS ephemeris updates.

Publications and talks

- Balancing Security and Correctness in Code Generation · IEEE Transactions (TETCI), 2025
- Evaluating Large Language Models for Enhanced Fuzzing · IEEE Access, 2024
- From LLMs to Randomness: Analyzing Program Input Efficacy · IEEE Access, 2025
- Firewall Attack Detections and Extractions (FADE) · IEEE Data Descriptions, 2025
- Security Dataset Augmentation Invariance and Distribution Independence · IEEE ICNC workshop, 2024
- MCP Safety Audit · SPIE, 2026, co-author
- Offensive and Defense Security Dataset Augmentation · US patent application 2024/0248984 A1, pending

Talks SPIE 2026 (MCP safety findings), Military Sensing Symposium 2025 (automated AI-assisted red teaming), National Cyber Summit 2025 (Huntsville), and IEEE ICNC 2024.

First author unless noted. | Code: github.com/gavin-black-dsu

Skills

AI security	automated red teaming · prompt injection · agent and MCP security · evals and benchmarks adversarial ML
Agentic engineering	Claude Code (workflows, subagents, headless) · Codex · MCP servers · agent harnesses
Security research	fuzzing (AFL++) · reverse engineering (IDA, Ghidra, ddisasm) · pentesting (Kali) vulnerability research
ML and LLMs	PyTorch · Hugging Face · BERT · vLLM · RL (novelty search) · SFT · RAG OpenAI and Anthropic APIs
Infrastructure	AWS · Docker · Kubernetes · Iron Bank containers · Linux · FastAPI · MongoDB Git automation
Languages	Python · C/C++ · Bash

Education

PhD, Cyber Defense · Dakota State University · Dissertation: "RUFF: Resource Usage Fuzzing Framework" 2024

MS, Mathematics · University of Massachusetts Lowell · Applied and computational 2014

BS, Computer Science · Purdue University · Second major in Mathematics 2006

Public portfolio: gavin-black.com, with projects, papers in plain English, and code.