Christian David Varela

CYBERSECURITY AND AI SAFETY RESEARCHER

San Francisco, CA · christian-varela.com · github.com/dreamsudo · linkedin.com/in/varela-sf · mailingvarela@proton.me

SKILLS

AI & ML security

LLM red-teaming: jailbreaks, direct and indirect prompt injection, system-prompt extraction, data and prompt leakage, refusal-robustness evaluation. Agent / tool-use security: MCP tool-call authorization, validation, auditing. Adversarial ML, model evaluation and safety testing, knowledge-graph grounding.

Machine learning

PyTorch, diffusion models, object detection, NLP. Empirical evaluation and experiment management: deterministic seeding, ablations, bootstrap CIs, reproducible pipelines.

Offensive security

Penetration testing of infrastructure, web apps, and AI/ML systems; vulnerability research, CVE analysis, privilege-escalation enumeration, kill-chain construction, digital forensics. Kali, Metasploit, Burp Suite, Wireshark, nmap.

Defensive security

Threat modeling and mapping, D3FEND countermeasure design, detection and incident response, network security, firewalls and IDS, access control, OS and enterprise hardening, risk assessment, tamper-evident logging.

Cryptography

Applied cryptography, symmetric and asymmetric encryption, hashing and digital signatures, key exchange, PKI and TLS, post-quantum cryptography (lattice-based schemes, NIST PQC), hash-chained integrity.

Programming & systems

Python (fluent) for data analysis and detection tooling, SQL and SQLite for querying large datasets, Bash, Java. Linux, Docker, systemd, networking. Automated testing and CI, Git.

PROJECTS

Psypher AI Threat Assessor

Open-source AI red-teaming engine. Attacks a deployed model with an ATLAS-tagged adversarial corpus — jailbreaks, prompt injection, system-prompt extraction, refusal testing — and maps each finding to a D3FEND countermeasure. Python, Anthropic API.

Kybernos

Zero-trust gateway for AI agents. Every tool call is authenticated, authorized deny-by-default, schema-validated, and audited before it executes. Python, MCP, NIST SP 800-207.

DPC-YOLO26

From-scratch defense against adversarial-patch attacks on object detectors, using diffusion residuals to localize the patch. Capstone, with research paper. Python, PyTorch.

HunterEngine

Deterministic threat-enrichment engine for phishing and insider lures. Maps behavior to MITRE ATT&CK and emits deploy-ready YARA and STIX 2.1. Python.

FOCUS

- **AI red-teaming** & model evaluation
- **Prompt injection** & jailbreaks
- **AI agent & tool-use** security
- **Adversarial ML** attacks & defenses
- **Offensive security** & vuln research
- **Defensive security** & hardening
- **Threat mapping** ATLAS → D3FEND

FRAMEWORKS & STANDARDS

MITRE ATLAS, ATT&CK, D3FEND, NIST CSF, CWE, CVE, CISA KEV, NVD.

WORK ELIGIBILITY

U.S. citizen

LAB

Founder — Psypher Labs

An independent, open-source AI-security research lab and think tank.

EDUCATION

B.S. Cybersecurity

California State University San Marcos, 2026. Dean's List.

National Center of Academic Excellence in Cyber Defense (CAE-CD)

Designated by the U.S. National Security Agency (NSA) and DHS

MATHEMATICS & PHYSICS

Linear algebra, multivariable calculus, differential equations, probability and statistics, discrete mathematics, classical mechanics, electromagnetism.

LANGUAGES

English & Spanish (native).