Yuvraj Pradhan

AI Security Researcher and Engineer, Founder of Vyrion

yuvraj0227@gmail.com | +91 97777 69399
yuvrajpradhan.com | linkedin.com/in/yuvraj-pradhan-27uv
github.com/Uvi-12 | Google Scholar

RESEARCH PROFILE

Technical AI security researcher investigating whether human-oversight and authority policies remain enforceable once agentic systems execute them. First author of an accepted ACM AISec 2026 paper on approval provenance; evaluated 43 version-pinned systems, coordinated disclosure across 29 vendors and projects, and built verification controls for approval integrity and instruction authority.

AI SECURITY AND GOVERNANCE RESEARCH

Approval-Provenance Failure *Independent Research, First Author* 2026

- Defined an authorization-integrity failure in which a system's representation of human approval can diverge from the decision actually made by the human, allowing approval to be forged, rebound, replayed, suppressed, or laundered before execution.
- Identified a policy-enforcement gap in human-in-the-loop systems: an organization can require an authorized human decision while the runtime accepts mutable workflow state as sufficient evidence that the required approval occurred.
- Evaluated 43 version-pinned systems under a defined state-writer adversary: 29 exposed, 12 resistant negative controls, and two not applicable.
- Coordinated disclosure with 29 vendors and projects through CERT/CC VINCE, Microsoft MSRC, Google AI VRP, AWS VDP, ASF Security, and GitHub Security Advisories. Reporter for CERT/CC case VU#550098; Flowise confirmed GHSA-2hjv-g5fh-6w4j as a Medium-severity CWE-345 vulnerability.

Instruction-Authority Divergence in AI Coding Agents *First Author* 2026

- Defined the gap between the instruction policy a reviewer believes has been approved and the authority policy an AI coding agent actually resolves at runtime through discovery, precedence, or import rules.
- Ran controlled evaluations on two coding agents: the reviewed root instruction held in 200/200 control runs, a dependency-supplied instruction won in 200/200 precedence runs, and import carriage reached first-party output in 86–99% of trials.

PUBLICATIONS

PEER REVIEWED

1. **Y. Pradhan**, A. Singh, A. Kumari, A. Pradhan, and D. Singh. "Ghost Approvals: Approval-Provenance Failure in Agentic AI Frameworks." *19th ACM Workshop on Artificial Intelligence and Security*, co-located with ACM CCS 2026. Accepted, 2026.
2. **Y. Pradhan**, A. Singh, N. Kulkarni, A. Priyadarshi, O. Rajale, and K. Mohite. "Nano Large Language Model of 125 Million-Parameter for STEM Education." *Discover Computing*, vol. 29, article 161, 2026. doi.org/10.1007/s10791-026-10033-z

UNDER REVIEW

3. "Nobody Approved This: Approval-Provenance Failure Across Agentic AI, Workflows, and CI/CD." Under review.
4. "Who Witnesses the Human in the Loop? The Approval Record the System Can Write for Itself." Under review.
5. "The Agent Obeys the Wrong Principal: Instruction-Authority Divergence in Coding Agents." Under review.

RESEARCH SOFTWARE AND VERIFICATION TOOLS

Vyrion *Founder · Python, Node.js, Ed25519, RFC 8785, Apache-2.0* 2026

- Built a fail-closed mechanism for enforcing human-approval policy through an Ed25519-signed, single-use authorization Seal rather than trusting mutable workflow state as evidence of approval.
- Binds approval to the approver, action, arguments, run, checkpoint, tenant, audience, policy, expiry, and nonce, making the authorization decision independently verifiable at execution time.
- Released two live-certified framework adapters and thirteen integration-ready adapters with cross-language verification, in-toto attestations, SARIF output, and reversible installation. vyrionlabs.com · github.com/Uvi-12/Vyrion · pypi.org/project/vyrion

Ovrin *Python, SPDX 3.0, GitHub Actions, Apache-2.0* 2026

- Built an offline governance and measurement tool for coding-agent instruction files that identifies instruction origin, CODEOWNERS coverage, per-agent reachability, and supply-chain paths through which unreviewed instructions may affect agent behavior.
- Powered a measurement study of 2,470 repositories, finding instruction files in 38.1% of repositories and a file-level review gap of 30.5%. github.com/Uvi-12/Ovrin

SELECTED TALKS AND PRESENTATIONS

Vyrion: Exploiting and Hardening Approval Integrity in AI Agents and Workflows *Black Hat MEA 2026, Arsenal, Riyadh, Saudi Arabia* — Dec 2026, Accepted

Ghost Approvals: Approval-Provenance Failure in Agentic AI Frameworks *ACM AISec 2026, co-located with ACM CCS, The Hague, Netherlands* — Nov 2026, Accepted

Ghost Approvals: Decision-Provenance Integrity Failure in Agentic AI Frameworks *DeepSec 2026, Vienna, Austria* — Nov 2026, Accepted

Ghost Approvals: The Agent Says a Human Approved It. No Human Did. *AI Agent Security Summit by Zenity, London, United Kingdom* — Oct 2026

Nobody Approved This: Hijacking Approval Authority *AltSecCon, Bengaluru, India* — Oct 2026

The Stdio Deadlock Nobody Warned Us About *MCP Dev Summit, Bengaluru, India; featured in the Agentic AI Foundation event recap* — Jun 2026

Who Do You Trust? Securing Multi-Agent MCP Systems in Healthcare *MCP Dev Summit, Mumbai, India* — Jun 2026

The Sticky Session Bottleneck: Why MCP Agent Orchestration Breaks When You Scale *MCP Dev Summit, Mumbai, India* — Jun 2026

Lightweight Small Language Models for Low-Resource Devices *ICRTEBTM 2026; presented research paper* — 2026

OPEN-SOURCE CONTRIBUTIONS

- **17+ merged documentation pull requests** to Hugging Face Transformers (5), NumPy (4), PyTorch (2), and CPython, pandas, LangChain, and scikit-learn (6+), including the Audio Classification and ASR pipeline pages, the AMSGrad equation in the Adam/AdamW docs, and Formatter initialisation in Python's logging documentation.

EDUCATION

MIT ADT University *B.Tech., Computer Science and Engineering, AI and Edge Computing · CGPA 8.26/10* — Expected Jul 2027

ADDITIONAL RESEARCH AND ENGINEERING

Multi-Agent Consensus System for Cancer Detection — 2025

- Built a LangGraph-based system coordinating five specialized agents for diagnostic review and consensus; the need to bind prescriptions, medicines, and test reports to a doctor's approval motivated the approval-provenance research program.

Enterprise Multimodal Retrieval System — 2026

- Built a text-and-image retrieval pipeline using LangChain, Tesseract OCR, and ChromaDB, tested across more than 10,000 pages with query latency below 1.2 seconds.

RESEARCH AND TECHNICAL EXPERTISE

AI governance and security: human oversight, authorization provenance, instruction authority, verification mechanisms, auditability, AI software supply chains, coordinated disclosure, threat modeling, trust-boundary analysis
Security and assurance: CWE, CVSS, Ed25519, RFC 8785, SHA-256 commitments, nonce and replay defense, SBOM, SPDX 3.0, in-toto attestations
Agentic systems: LangGraph, LlamaIndex, CrewAI, OpenAI Agents SDK, Google ADK, AutoGen, Haystack, Strands, Agno, Semantic Kernel, DBOS, Apache Burr, Apache Airflow, Genkit, Vercel AI SDK, MCP
Engineering: Python, JavaScript, Node.js, C, C++, SQL, Bash, PyTorch, Hugging Face, Docker, AWS, Google Cloud, pytest, GitHub Actions, Git