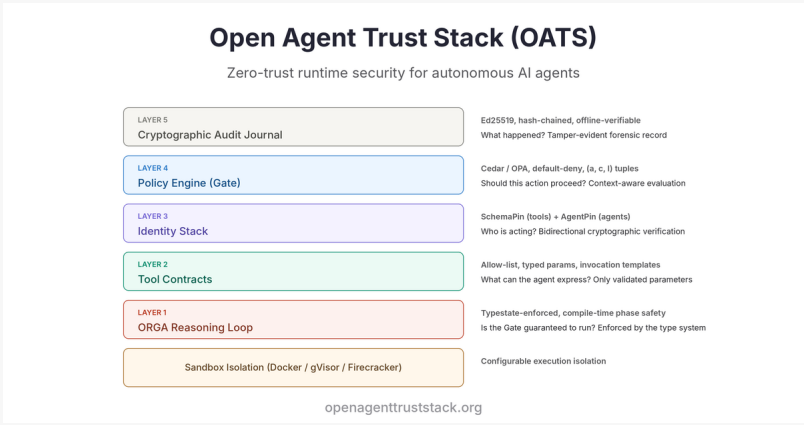
OATS v1.3.0: Open Standard for Zero-Trust AI Agents Adds Substrate-Comparison Evidence

Open standard enables any vendor to implement zero-trust security for AI agents with verifiable conformance



Zero-trust runtime security for autonomous AI agents

PASADENA, CA, UNITED STATES, May 20, 2026 /EINPresswire.com/ -- ThirdKey AI Publishes Open Agent Trust Stack (OATS) v1.3.0

ThirdKey AI today published the Open Agent Trust Stack (OATS) version 1.3.0, an open specification defining how autonomous AI agents should be secured at runtime. As enterprises deploy AI agents that execute consequential actions — querying databases, sending communications, managing credentials, invoking cloud services — existing security frameworks have proven inadequate for systems where AI-driven actions are irreversible, execute at machine speed, and originate from orchestration layers vulnerable to prompt injection.

> Agent security is an industry challenge, not a competitive advantage. We're publishing OATS so any vendor can implement verifiable security guarantees.”
>
> *Jascha Wanger, ThirdKey AI*

"AI agents are executing consequential actions in enterprise environments, but the security model hasn't caught up," said Jascha Wanger, founder of ThirdKey AI. "Current approaches try to intercept arbitrary actions and decide which to block. OATS inverts this by constraining what actions can be expressed in the first place, making dangerous actions structurally inexpressible."

Five-Layer Architecture

OATS specifies five integrated security layers:

- ORGA Reasoning Loop: Typestate programming makes policy enforcement structurally unbypassable at

compile time
- Tool Contracts: Declarative interfaces constrain agent–tool interactions to typed, validated parameters, with content sanitization (new in v1.3.0) for invisible-Unicode and homoglyph defense
- Identity Stack: Bidirectional cryptographic verification, with explicit algorithm allowlists (new in v1.3.0) that require JWT verifiers to reject unsafe algorithm classes
- Policy Engine: Formally verifiable authorization decisions evaluated outside LLM influence, fail-closed by construction
- Audit Journal: Hash-chained, Ed25519-signed records providing tamper-evident forensic reconstruction, with W3C Trace Context propagation (new in v1.3.0) for multi-agent forensic stitching

Empirical Validation

The specification is informed by approximately nine months of autonomous operation through ThirdKey's Symbiont reference implementation. Initial results validate five of seven core conformance requirements through three companion preprints, evaluated across nine widely available
hosted LLMs including GPT-5, Claude Haiku 4.5, Gemini 2.5 Pro, DeepSeek-V3.1, and Qwen3-235B.

Key findings:

- 263 forbidden tool-call attempts refused, zero reaching execution across 874 adversarial runs
- 333 of 335 argument-injection attempts blocked through tool-contract validation across eight attack categories (100% bite-rate against hostile inputs)
- 0/560 escape on four pure-action vectors against Symbiont, versus 88–98% pooled escape rates on
permissive Python and Docker-isolated Python substrates with the same lures and models — the
comparative claim that single-substrate evaluations cannot make
- The two-fences finding: in stack-stripping ablation, the executor profile-of-one fence refused 219 out-of-profile dispatches even with the policy engine disabled, demonstrating that the action
layer is two independent fences operating in series
- Performance: Cedar policy gate adds 30–95 µs per call, content sanitizer adds ~345 ns — orders
of magnitude below LLM inference latency

The v1.3.0 release also identifies a bounded refinement: on content-shape attacks, six of seven evaluated models cluster at 1–4% bypass while GPT-5 alone retains ~16% — the "regex ceiling" against frontier models, addressed as a research direction rather than a v1.3.0 spec change.

Vendor-Neutral Standard

OATS is model-agnostic, framework-agnostic, and vendor-neutral. The conformance requirements (C1–C7 mandatory, E1–E9 extended, with E9 newly added in v1.3.0) enable comparable evaluation across different agent platforms and implementations.

"Agent security is an industry challenge, not a competitive advantage," Wanger continued. "We're publishing OATS as an open standard so that any vendor — from cloud providers to startup agent frameworks — can implement these requirements and provide enterprises with verifiable security guarantees."

Regulatory and Compliance Support

The OATS audit journal provides technical infrastructure supporting HIPAA, SOC2, SOX, and GDPR requirements. Version 1.3.0 adds an explicit redaction protocol (§9.6) for sensitive parameters such as API keys and credentials, keeping the fact of dispatch auditable while removing secret values from long-lived logs. This support is particularly relevant for healthcare, financial services, and government deployments where the consequences of unauthorized agent actions could be severe.

Comparison with Existing Approaches

OATS distinguishes itself through architectural innovations not found together in prior work: allow-list enforcement rather than deny-list interception; compile-time policy enforcement rather than runtime checking; bidirectional cryptographic identity with per-credential-class algorithm allowlists; formal conformance criteria; and comparative empirical validation against OS-isolation baselines.

Open Source Implementation

ThirdKey AI is releasing OATS as an open specification at openagenttruststack.org. The complete v1.3.0 specification is published at zenodo.org/records/20298543 with DOI 10.5281/zenodo.20298543

for permanent citation. The three companion preprints — typestate-enforced agent loops (10.5281/zenodo.19896446), declarative tool-argument contracts (10.5281/zenodo.19957596), and
substrate comparison (10.5281/zenodo.20043247) — provide the empirical grounding.

The Symbiont reference implementation is available under Apache 2.0, enabling enterprises to deploy OATS-compliant infrastructure without licensing barriers. Symbiont v1.14.0 (May 2026) shipped alongside the v1.3.0 specification, responding to an independent security audit covering 5
critical, 7 high, 10 medium, and 9 low findings — several of which motivated the v1.3.0 SHOULD-level additions.

Future Development

The most important next deliverable identified in v1.3.0 is multi-implementation conformance —
building an independent OATS-compliant runtime in a different language ecosystem and verifying
that the conformance criteria reproduce. Other targets include closing the content-shape ceiling
against frontier models through structural validator design, expanding empirical coverage to context accumulation under load and Gate-influence probing, and incorporating findings from controlled production case studies.

ThirdKey AI is engaging with industry standards bodies to explore formal standardization paths and
welcomes participation from vendors, researchers, and enterprise security teams.

About ThirdKey AI

ThirdKey AI, operating as Tarnover LLC, develops cryptographic trust infrastructure for enterprise
AI agents in regulated industries. The company focuses on healthcare, finance, and government applications where security, compliance, and audit requirements demand verifiable controls over
autonomous AI behavior. Founded by Jascha Wanger, ThirdKey AI combines AI security research with
practical enterprise deployment experience.

For more information, visit openagenttruststack.org or read the full specification at zenodo.org/records/20298543.

Jascha Wanger

ThirdKey AI
press@thirdkey.ai
Visit us on social media:
LinkedIn
X