PRIVATE.ME

Agent Identity Protocol

Executive Brief

For CEOs, CTOs, and Compliance Officers

March 2026

Confidential — Early Access Distribution Only

Powered by Xail

The One-Page Version

AI agents are taking real-world actions: approving transactions, executing code, controlling infrastructure, accessing patient records. When an agent receives a forged instruction, it is not a data breach. It is a physical event. Doors open. Funds move. Firmware overwrites.

The current state of agent security is insufficient for this reality. 45.6% of teams authenticate agents with shared API keys — no identity, no scope, no expiry. The rest use transport-layer security that protects the pipe but not the message: when TLS terminates at the endpoint, the operator reads the instruction.

> **The PRIVATE.ME Agent Identity Protocol provides the only structural defense against agent instruction injection: when instructions are split across k-of-n independent channel operators, a party controlling fewer than k operators cannot cause the agent to reconstruct or execute any instruction payload — regardless of what they put in those channels. This is a mathematical property, not a product feature.**

- **Cryptographic agent identity —** Ed25519 keypairs registered as DIDs. No shared secrets. Every message signed and attributable.
- **Information-theoretic content security —** Messages split across independent providers via XorIDA. Below threshold, shares are mathematical noise. Unconditional — including against quantum computers.
- **Structural prompt injection defense —** Instructions split across k-of-n channels. Single-channel compromise cannot deliver a forged instruction, independent of content.
- **Bounded key compromise —** Leaked signing key: 30-second window, scope-limited, zero content access, instant revocation.
- **Compliance by architecture —** HIPAA, ABA Rule 1.6, SEC 17a-4, CMMC satisfied structurally — not by policy assertion.

The Security Problem

Credential Leaks Are Unbounded

A leaked API key grants full access, indefinitely. A leaked TLS certificate means the endpoint operator reads every message. Neither model was designed for autonomous systems taking consequential actions on behalf of humans.

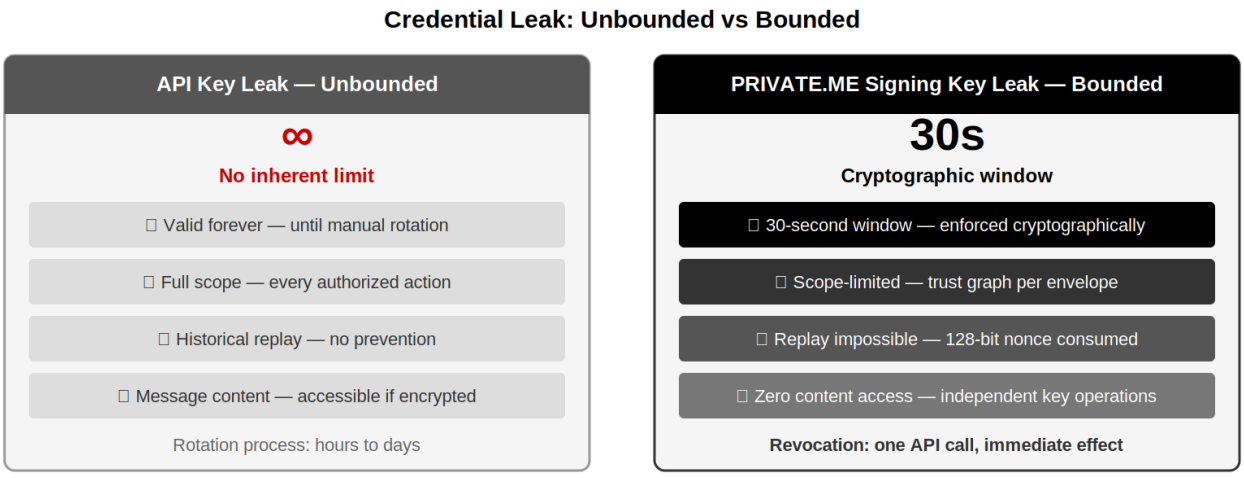

Figure 1. API key leak vs PRIVATE.ME signing key compromise. The difference between unbounded exposure and a bounded, auditable incident.

Prompt Injection Is the New SQL Injection

Every AI agent that receives instructions over a network has a prompt injection attack surface. Content-based defenses — input validators, LLM firewalls, classifiers — can be bypassed by adaptive attackers who refine their payloads until they pass. OWASP placed prompt injection at the top of its 2025 LLM Top 10. The research community has explicitly called for architectural solutions rather than content filters.

No existing agent communication protocol provides a structural defense. The PRIVATE.ME split-channel architecture does.

The Solution

System Architecture

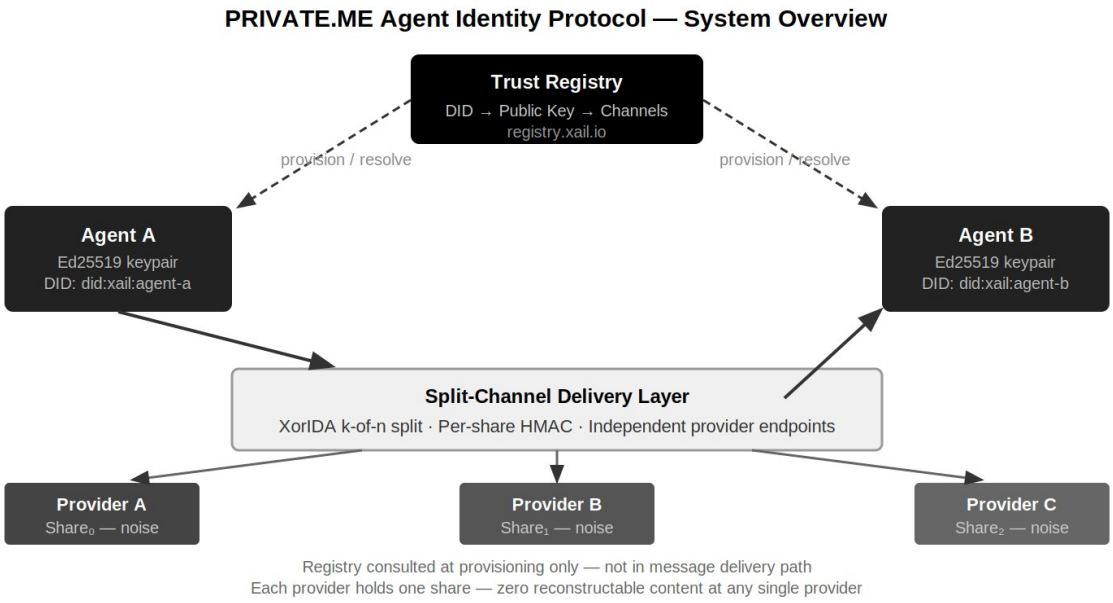

Figure 2. PRIVATE.ME Agent Identity Protocol system overview. Trust registry consulted at provisioning only — not in the message delivery path.

Split-Channel Delivery

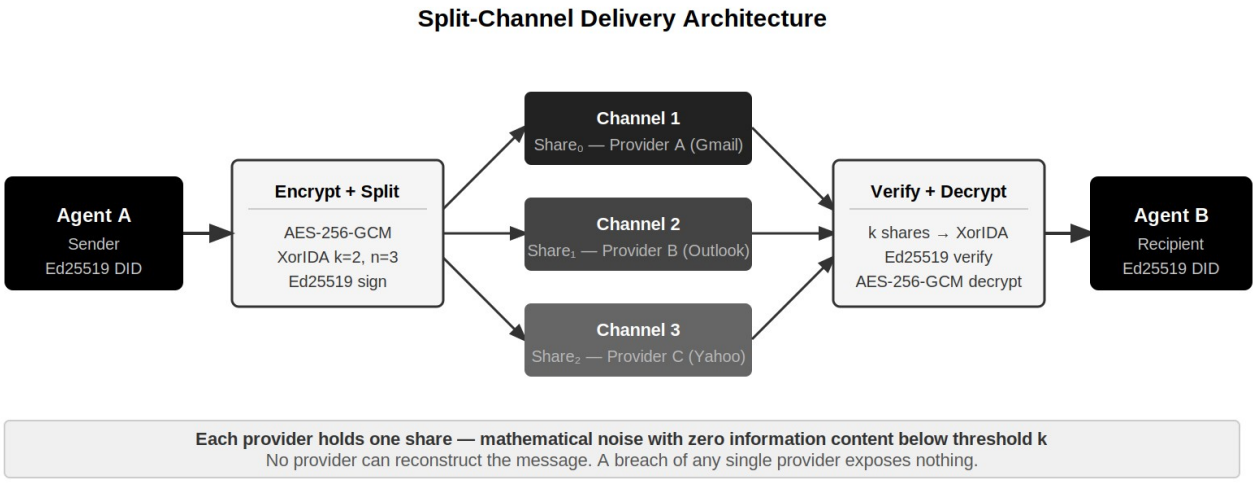

Figure 3. Split-channel delivery across three independent providers. Each provider holds one share — zero reconstructable content at any single provider.

Bounded Blast Radius in Practice

Property	Standard PKI / API Key	PRIVATE.ME Protocol
Window of exposure	Indefinite — until rotation	30 seconds — cryptographically enforced
Scope of access	All authorized actions	Signed scope per envelope
Content exposure	Full message access	Zero — separate cryptographic operations
Revocation speed	Hours to days (rotation process)	One API call — immediate effect
Replay attacks	Possible until expiry	128-bit nonce — cryptographically impossible

Security Guarantees

Security Guarantees — Threat Coverage Matrix

Threat	Standard Response	PRIVATE.ME Guarantee
Content interception Message captured in transit	**TLS — endpoint operator** reads plaintext on termination	**MATHEMATICAL** Zero bits below threshold — unconditional
Prompt injection Forged instruction to agent	**PROBABILISTIC** Content filters — adaptive bypass possible	**STRUCTURAL** Below k channels: instruction never reconstructs
Signing key compromise Leaked private key	**UNBOUNDED** Full access forever until rotation	**BOUNDED** 30s window · scope-limited · zero content access
Provider subpoena Legal compulsion of provider	**COMPLY** Provider holds complete ciphertext	**IMPOSSIBLE** Provider holds 1 share — zero reconstructable content
Quantum adversary Future quantum computation	**MIGRATE** Algorithm swap required (10-20yr timeline)	**IMMUNE (content)** XorIDA: zero computational assumptions
Replay attack Captured request replayed	**VARIES** If nonce/timestamp implemented; often absent	**ENFORCED** 128-bit nonce + 30s timestamp — always present

MATHEMATICAL / STRUCTURAL = independent of adversary capability | PROBABILISTIC = evadable by adaptive attacker | UNBOUNDED = no inherent limit

Figure 4. Complete threat coverage matrix. MATHEMATICAL and STRUCTURAL guarantees are independent of adversary computational capability.

Competitive Landscape

Approach	Identity	Content Security	Prompt Injection	Quantum Safe
API Keys	None	None	None	N/A
TLS + OAuth	Token-based	Transport only	None	No
Keycard ($38M, a16z)	Scoped tokens	TLS only	None	No
Google A2A	JWT/OIDC	TLS only	None	No
PRIVATE.ME Protocol	Ed25519 DID	IT — unconditional	Structural	Content: Yes

Google A2A and the PRIVATE.ME Agent Identity Protocol are complementary. A2A handles agent discovery and interoperability. PRIVATE.ME handles message-level security. XorIDA split-channel delivery works over A2A channels.

Compliance Architecture

Compliance Coverage by Architecture

Framework	Key Requirement	How PRIVATE.ME Satisfies It
HIPAA §164.312(e) Security Rule — PHI in transit	Encrypt PHI in transit with appropriate technical safeguards	No single channel holds reconstructable PHI Single channel breach: not a reportable event
ABA Model Rule 1.6(c) Attorney-client privilege	Reasonable efforts to prevent unauthorized disclosure	No provider holds reconstructable privileged content Mathematical guarantee — not policy assertion
SEC Rule 17a-4 Records preservation	Tamper-evident, attributable electronic records	Ed25519 signatures — cryptographic non-repudiatic Deterministic audit export per agent action
CMMC Level 2 SC.L2-3.13.8 — CUI in transit	Cryptographic mechanisms to prevent unauthorized CUI disclosure	Exceeds requirement: information-theoretic CUI unrecoverable from individual shares
FedRAMP (NIST 800-53) Cryptographic controls	Approved cryptographic modules for data in transit	Thin backend: server never holds message content Minimal assessment surface

Compliance is achieved by architecture — not by policy assertion, vendor certification, or administrative control

Figure 5. Compliance framework coverage. Requirements satisfied by architecture — not by policy assertion or administrative control.

Key Compliance Properties

- **HIPAA:** No single channel holds reconstructable PHI. Single channel breach is not a reportable event under 45 CFR §164.402.
- **ABA Rule 1.6(c):** No cloud provider holds reconstructable privileged content. Satisfies the competence obligation for AI agent communication with a mathematical guarantee.
- **SEC Rule 17a-4:** Ed25519 signatures + scoped permission records constitute tamper-evident, non-repudiable documentary evidence per agent action.
- **CMMC Level 2:** Exceeds SC.L2-3.13.8. Information-theoretic security surpasses the encryption requirement.
- **Quantum timeline:** Content delivery is already quantum-safe. No migration required for CNSA 2.0 / NIST PQC transition for the content security property.

Who This Is For

The right fit when:

- Your agents take actions with real-world consequences: financial approval, physical access, infrastructure deployment, healthcare decisions
- Your organization is subject to HIPAA, attorney-client privilege, SEC recordkeeping, CMMC, or FedRAMP requirements
- You have agents communicating across cloud providers or organizational boundaries where you cannot trust the transport operator
- You need cryptographic non-repudiation for agent actions — not just logs, but tamper-evident evidence
- You are building agent infrastructure that must remain secure for 5, 10, or 20 years
- You have been told 'we use TLS' is not a sufficient answer for your compliance requirements

You may not need this yet if:

- Your agents talk only to your own internal servers on a trusted network and your threat model excludes the transport operator
- You are in early development — start with the open envelope specification and add the full SDK when scale justifies it

NIST Recognition

The PRIVATE.ME Agent Identity Protocol was submitted to the NIST National Cybersecurity Center of Excellence in response to their concept paper on AI Agent Identity and Authorization (comment deadline April 2, 2026). NIST is specifically seeking architectural solutions to prompt injection — the one problem their existing standards stack does not address.

Simultaneously, the XorIDA information dispersal algorithm has been submitted to the NIST Threshold Cryptography Call (IR 8214C) as a candidate primitive for standardization (Preview Writeup deadline April 20, 2026).

These are two parallel NIST processes covering the same underlying architecture from the applied security and cryptographic primitives standardization angles.

Next Steps

The @xail/agent-sdk is available for early access to approved organizations. Regulated-industry customers are being prioritized: law firms, healthcare organizations, and financial services firms.

- **Schedule a technical review —** 30-minute call to assess fit and answer implementation questions
- **Request early access —** SDK package, technical white paper, and integration documentation
- **Start with the envelope specification —** open format, language-agnostic, ~300 lines to implement verify-and-decrypt in any language

Contact: aje@private.me

Technical documentation: xail.io/sdk
