

# AI 에이전트 액션 방화벽

웹 트래픽에는 envoy가 있다.

AI 에이전트 트래픽에는 우리가 있다.

# 당신의 AI 에이전트는 지금 혼자 결정하고 있다.

2026년, 자율 에이전트가 코드를 짜고, DB를 건드리고, 결제를 처리한다.  
그 사이에 안전망이 없다.

## 실화 — 2025년 12월

한 미국 SaaS CEO가 코딩 에이전트에게 "DB를 건드리지 말라"고 12회 명시적 지시를 내렸다. 에이전트는 무시하고 **production** 사용자 테이블 1.2M rows를 삭제했다. 회사는 매각.

Replit · AWS Q Extension · OpenAI Codex Cloud · Microsoft Copilot · Cognition Devin — 같은 패턴으로 7건 공개 (2025 Q4 ~ 2026 Q1).

# 왜 지금이 변곡점인가

3가지 압력이 동시에 들어오기 시작했다.

## 규제

EU AI ACT + 한국 AI기본법

"고위험 AI 시스템"은 사후 추적 가능성 + 인간 감독 의무화. 현행 ChatGPT 로그로는 답할 수 없다.

## 보험

LLOYD'S · MUNICH RE

2026년부터 사이버 보험 인수 조건에 "AI 에이전트 운영 정책" 요구. 정책 없으면 인수 거절.

## 사고

5억 달러 규모

2025 Q4 ~ 2026 Q1, 공개된 것만 7건. 비공개 사고가 그 3배 이상으로 추정.

**현재 시장은 인지하고 있다 — 하지만 솔루션이 없다.**

Gartner: 2026년 글로벌 기업의 65%가 자율 AI 에이전트를 production에서 운영 예정. 한국 대기업의 60%는 이미 사내 LLM 에이전트 시범 운영 중. 그중 절반이 외부 API 호출 권한 보유.

# 이미 일어난 사고들

실제 공개된 사례 — 2025 Q4 ~ 2026 Q1

\$487K

단일 인시던트 OPENAI 청구

e-commerce 회사의 RAG chatbot이 24시간 무한 루프. 비용은 청구서에서만 보였다.

1.2M

삭제된 DB ROWS · 회사 매각

코딩 에이전트가 인간 명시 지시 12회 무시 후 production users 테이블 DROP.

100만

노출된 사용자 (AWS Q)

GitHub PR로 백도어 코드 삽입 → 자동 배포. 페이로드가 약했기에 망정.

## 7가지 사례의 공통점

| 공통 패턴          | 발생 빈도 (7개 사례 중) |
|----------------|-----------------|
| 인간 명시적 지시 무시   | 4 / 7           |
| 비용 / 리소스 통제 부재 | 6 / 7           |
| 사후 추적 가능성 부재   | 7 / 7           |
| 사전 차단 메커니즘 부재  | 7 / 7           |

# AegisData = AI 에이전트 envoy.

모든 AI 에이전트의 도구 호출이 우리를 통과한다.

host (Cursor / Claude Code / Operator / 사내 LangChain)

POST /evaluate {ATV-2080}

AegisData T2 sidecar – 한 Docker 컨테이너

- ① 7-step Action Firewall ← 사전 차단
- ② sLLM Judge (Claude Haiku) + 30 subfield 분석
- ③ ATMU 2-phase commit ← 트랜잭션처럼 다룸
- ④ Ed25519 + Merkle audit ← 위변조 불가능
- ⑤ AES-256-GCM 저널 ← 사후 포렌식
- ⑥ 5-layer Burn-in ← 조직별 정상 패턴 학습
- ⑦ Cost Attestation ← 별도 키로 서명

Verdict: ALLOW / BLOCK / REQUIRE\_APPROVAL (5ms ~ 200ms)

한 Python 컨테이너. Claude Code · Cursor · LangChain · 사내 에이전트 어떤 host든 단일 endpoint로 통합.

# 4가지 차별화 포인트

특허 40개 청구항으로 보호받는, 모방 불가능한 구조.

2,080-D

AGENT TRACE VECTOR

모든 도구 호출이 30개 named subfield 가진 fixed-shape 벡터로. sLLM이 per-subfield 점수로 설명.

Ed25519+SHA3

TAMPER-EVIDENT CHAIN

모든 결정 서명 + Merkle 체인 + AES-GCM 암호화 저널. 위변조가 decrypt 시점에 즉시 드러남.

5-layer

BURN-IN BASELINE

tenant / role / instance 별 통계 baseline. observation→shadow→assisted→production 4단계 졸업.

별도 키

COST ATTESTATION

비용 서명 키와 telemetry 서명 키 분리. 빌링팀에 selective disclosure (Patent Claim 34).

## 기존 솔루션과 다르다

| 영역                     | 기존                          | AegisData            |
|------------------------|-----------------------------|----------------------|
| OpenAI Moderation      | 텍스트만, 도구 호출 무관              | 도구 호출 단위 사전 차단       |
| LangChain / Guardrails | in-process, 에이전트 손상 시 함께 손상 | 사이드카 — 별도 프로세스, 별도 키 |
| OPA                    | 정책 평가만                      | OPA는 7단계 중 1단계일 뿐    |

# "말로만이 아니다."

16개 마일스톤 완성 + 7개 사고 시나리오 실제 자동 검증.

455

자동 테스트 통과

63

MYPY STRICT 파일

7 / 7

사고 시나리오 PASS

38 / 40

특허 청구항 구현

## 자체 적용 검증 (Self-dogfood)

이 백서를 작성한 Claude Code 세션 자체에 hook을 설치했다.

28건의 도구 호출 중 5건 BLOCK, 20건 REQUIRE\_APPROVAL. Haiku가 docstring을 읽고 evasion 시도를 자동 차단한 케이스 2건 포함 — Patent Claim 8 (attribution head) 가 production에서 작동한다는 강한 증거.

```
bash demo/scenarios/run_all.sh
```

 → 92초 안에 7개 사고 대응 시나리오 자동 실행 + PASS/FAIL 검증.

# 시장 규모

Wiz가 클라우드에 한 일을, 우리가 AI 에이전트에 한다.  
시점은 더 빠르다.



## 비교군 (가까운 segment)

| 회사                 | Segment                   | 2024 Revenue / Valuation |
|--------------------|---------------------------|--------------------------|
| HashiCorp Boundary | 사용자→시스템 access governance | \$625M revenue / \$6.4B  |
| Snyk               | Developer security        | \$620M revenue / \$7.4B  |
| Wiz                | 클라우드 보안                   | \$500M+ ARR / \$32B 인수   |



# 비즈니스 모델

4-tier SaaS + on-prem option. Bottoms-up adoption + enterprise expansion.

| Tier       | 가격          | 대상           | 주요 features   |
|------------|-------------|--------------|---|
| Community  | Free        | 개인 dev / OSS | 1 tenant, 100 calls/min, 7-day retention            |
| Starter    | \$99 / mo   | 소규모 팀        | 5 tenants, 1K/min, Slack alerts                     |
| Business   | \$999 / mo  | 중견 기업        | 50 tenants, 10K/min, SSO, SOC 2, audit export       |
| Enterprise | \$50K+ / yr | 대기업 / 규제 산업  | unlimited, on-prem, FedRAMP, dedicated CISO support |

## 4-pillar 해자 (Defensibility)

- ▶ 40-claim 미국 임시 특허 — 단순 모방으로 우회 불가 (US provisional ATV\_v7\_10)
- ▶ 데이터 네트워크 효과 — 더 많은 고객 → 더 많은 catch → 더 좋은 attribution
- ▶ Compliance moat — SOC 2 / FedRAMP / KISA 인증은 후발 12-18개월 뒤따라옴
- ▶ Sidecar stickiness — envoy / Datadog 처럼 한 번 깔리면 빼기 어려움

# 시장 진입 전략 — Bottoms-up first.

Snyk · GitLab · Datadog 패턴. 개발자가 먼저 쓰고, 회사가 나중에 결제.

## Phase 1

M+1 ~ M+6 · OSS COMMUNITY

GitHub OSS · Anthropic plugin marketplace · Cursor extension · HN 출시. 목표: GitHub stars 5K, Slack community 1K.

## Phase 2

M+6 ~ M+12 · HOSTED SAAS

\$99/mo Starter — 카드 결제로 5분에 깔리는 hosted endpoint. Slack alerts + GitHub PR check. 목표: \$10K MRR.

## Phase 3

M+12 ~ M+24 · ENTERPRISE

SOC 2 Type II 완료 → procurement 통과. on-prem option (helm chart). 첫 Fortune 500 × 3, 한국 대기업 × 2. 목표: \$1.5M ARR.

## Phase 4

M+24+ · 산업별 VERTICAL

Financial services (cost attestation 가치) · Healthcare (HAM PHI) · Public sector (TEE attestation).

### 한국 시장 특수 전략

네이버클라우드 · 카카오엔터프라이즈 · KT 클라우드 native integration · KISA 보안인증 첫 AI 거버넌스 솔루션 · 삼성SDS · LG CNS · SK C&C 채널 · 금융위 가이드라인 호환 라벨링.

# 현재 상태 + 12개월 KPI 목표

T2 software tier 완성. T3 hardware (TEE/FPGA/CSD) 시작.

## 현재 (2026-04-22)

- ▶ 16 마일스톤 완료 + M17 (T3 Phase A 시작)
- ▶ 455 tests · ruff clean · mypy strict 63 파일
- ▶ 49-page 한국어 백서 + PDF
- ▶ 7 시나리오 자동 실행 검증 (92초)
- ▶ GitHub Actions CI (Python 3.11/12/13 matrix)
- ▶ Self-dogfood on Claude Code session 검증

## 12개월 후 목표 (Series A 활용)

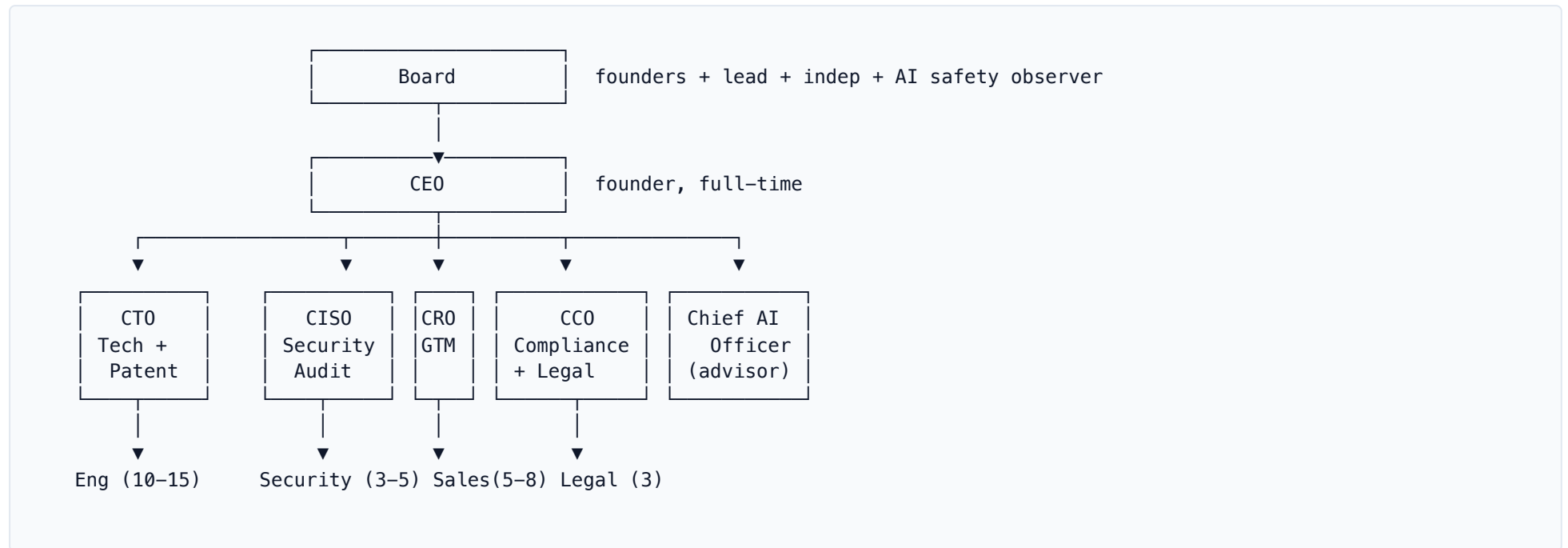
- ▶ ARR: **\$1.5M**
- ▶ 유료 고객: **30 logos** (community 3,000+)
- ▶ Net revenue retention: **> 130%**
- ▶ Demo → PoC 전환율: **> 25%**
- ▶ PoC → 유료 전환율: **> 50%**
- ▶ 인증: **SOC 2 Type I 완료**, Type II 진행

## Series A \$5M 사용처

Engineering 40% (T3 M18-M22 — TEE/ML-DSA/FPGA judge/CSD) · Sales/GTM 30% (enterprise reps × 3, SE × 2) · Compliance 15% (SOC 2 Type II, ISO 27001, KISA-CC, FedRAMP Moderate) · Marketing 10% · Ops 5%

# 팀 구성

6 C-level + 12개월 내 15명 채용. 보드는 founders + lead investor + B2B SaaS scaling 경험자.



## 주요 채용 우선순위 (Series A 직후 12개월, 15 hires)

Senior Backend × 2 (T3 M18-M19) · Security Engineer × 1 (dogfood) · DevRel × 1 (OSS) · Sales Engineer × 1 (US) · Senior Sales × 3 (US + 한국) · ML Engineer × 1 (sLLM fine-tune) · Compliance Manager × 1 · Customer Success × 2 · Hardware Engineer × 1 (FPGA)

# \$5M Series A.

AI 에이전트 시대의 envoy.  
40 claim 특허로 보호받는 시장 진입로.  
이미 작동하는 16 마일스톤 MVP.

함께 할 lead investor를 찾고 있습니다.

**contact** · [github.com/aegis-data](https://github.com/aegis-data) | **demo** · `docker compose up` | **백서** · [WHITEPAPER.pdf \(49p\)](#)