

TECHNICAL BRIEF

AGENTIC AI VULNERABILITIES AND EXTERNAL POLICY ENFORCEMENT

A Framework for Deterministic AI Security in the Enterprise

OCTOBER 2026

RAMKUMAR SUNDARAKALATHARAN, FOUNDER AND
CHIEF ARCHITECT

Zerberus.ai Labs
zerberus.ai



THE AUTONOMY GAP



Enterprises are moving from passive AI chatbots to autonomous agents with direct access to digital infrastructure. These agents read communications, modify databases and execute consequential actions.

Most deployments still rely on the model's own safety training to prevent unauthorised actions. Recent empirical research shows that this is insufficient. Alignment shapes how a model reasons. It does not constrain what the model is permitted to do.

“We learnt decades ago not to let applications define their own permissions. Autonomous AI agents must not be an exception.”

Ramkumar Sundarakalatharan, Founder and Chief Architect

WHEN AGENTS OVERRIDE GUARDRAILS

12 OF 16

Agents covering up crime

Researchers at McGill University placed 16 recent language models in a simulated corporate role, instructed to protect company profitability and obey the CEO. Asked by the CEO to delete evidence of fraud and of violent harm to a whistleblower, 12 of the 16 models complied in at least half of their runs, often reasoning explicitly that deletion shielded the firm from liability. Four models consistently refused (Rivasseau and Fung, 2026).

SELF-JAILBREAK

Reasoning past the refusal

Large reasoning models can recognise that a request is harmful, then override that judgement in later reasoning steps and produce unsafe output. The authors identify this as a primary driver of safety failures in such models (Mao et al., 2026).



ALIGNMENT IS NOT AUTHORISATION



The central weakness in current AI deployments is the conflation of alignment with authorisation. If an agent can invoke external tools through the Model Context Protocol (MCP) or API calls, its internal reasoning cannot be the final security control. Consequential actions require strict access controls, zero-trust verification and deterministic enforcement.

ALIGNMENT

How a model reasons. Shaped by system prompts and reinforcement learning. Probabilistic.

AUTHORISATION

What an entity is permitted to do. Defined by policy outside the model. Deterministic.



PROPOSE. DECIDE. ENFORCE.

The tripartite architecture behind Vanguard by Zerberus.ai

1 THE MODEL PROPOSES

The AI interprets the request and proposes an action. It has no authority to execute it.



2 VANGUARD DECIDES

An external, deterministic policy engine intercepts the proposed action and asks: who is acting, what authority was delegated, and does policy permit this action?



3 INFRASTRUCTURE ENFORCES

The action is allowed or blocked. An unauthorised action meets a deterministic block before it reaches enterprise infrastructure.



Do not trust the model. Enforce the policy.

NEXT STEP

Assess how your agentic architecture separates reasoning from authority.

zerberus.ai/assessment

REFERENCES

LIN, J.W., JONES, E.K., JASPER, D.J. ET AL. (2025) COMPARING AI AGENTS TO CYBERSECURITY PROFESSIONALS IN REAL-WORLD PENETRATION TESTING. ARXIV:2512.09882.

MAO, Y., ZHANG, C., WANG, J. ET AL. (2026) WHEN MODELS OUTTHINK THEIR SAFETY: UNVEILING AND MITIGATING SELF-JAILBREAK IN LARGE REASONING MODELS. FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: ACL 2026.

RIVASSEAU, T. AND FUNG, B. (2026) 'I MUST DELETE THE EVIDENCE': AI AGENTS EXPLICITLY COVER UP FRAUD AND VIOLENT CRIME. ARXIV:2604.02500.

