OASec
2026 · SINGAPORE

Signals from the Field: Four Patterns Reshaping AI Security

A review of research submitted to OASec 2026's Call for Proposals

PUBLISHED

September 2026

AUTHORS

Alex Leung, Kentaroh Toyoda, Yi Ting Shen, SiewMei Loh, Maria Singh, Michael Hwang, Shawn Kim, Theresa Blissing, Adrianna Tan, Donavan Cheah, and Mihaly Bujaki

Disclaimer

This report is provided as is, without warranties of any kind. It reflects our review of proposals submitted to a single call for proposals. It is not market research, investment guidance, regulatory or supervisory guidance, or advice of any other kind, and it should not be relied on as such.

No individual submission, speaker, or organization is identified. The observations are based on themes across the submitted proposals and do not represent an evaluation, endorsement, or criticism of any contributor.

Executive Summary

We reviewed more than 60 proposals submitted to OASec 2026's call for proposals and identified four recurring patterns:

1. Runtime containment and authorization for AI agents
2. Attacks that target not only prompts but also model files and retrieval systems
3. Evaluation of defenses under production conditions
4. Governance converging on shared standards, with early work measuring public trust

While these appear to be distinct patterns, they collectively point to a common expectation: AI security claims must be demonstrated in operation, whether through enforceable controls, measured defenses, or recognized standards. This report describes the evidence observed in the submissions and the implications for organizations with AI systems in production.

1. Introduction

OASec is the inaugural AI safety and security forum scheduled to take place in Singapore on September 28, 2026. We received over 60 submissions, which came from contributors across Asia Pacific, Europe, and North America, including Singapore, Japan, Taiwan, Malaysia, Thailand, India, the United Kingdom, Germany, France, Switzerland, and the United States. Contributors ranged from global banks and university laboratories to independent researchers and engineers responsible for large consumer AI products. Ahead of the event, we reviewed every submitted proposal and selected ten submissions for presentation.

During the review, we examined the proposals and identified four common patterns, which we discuss in this report. Figure 1 summarizes the four patterns.

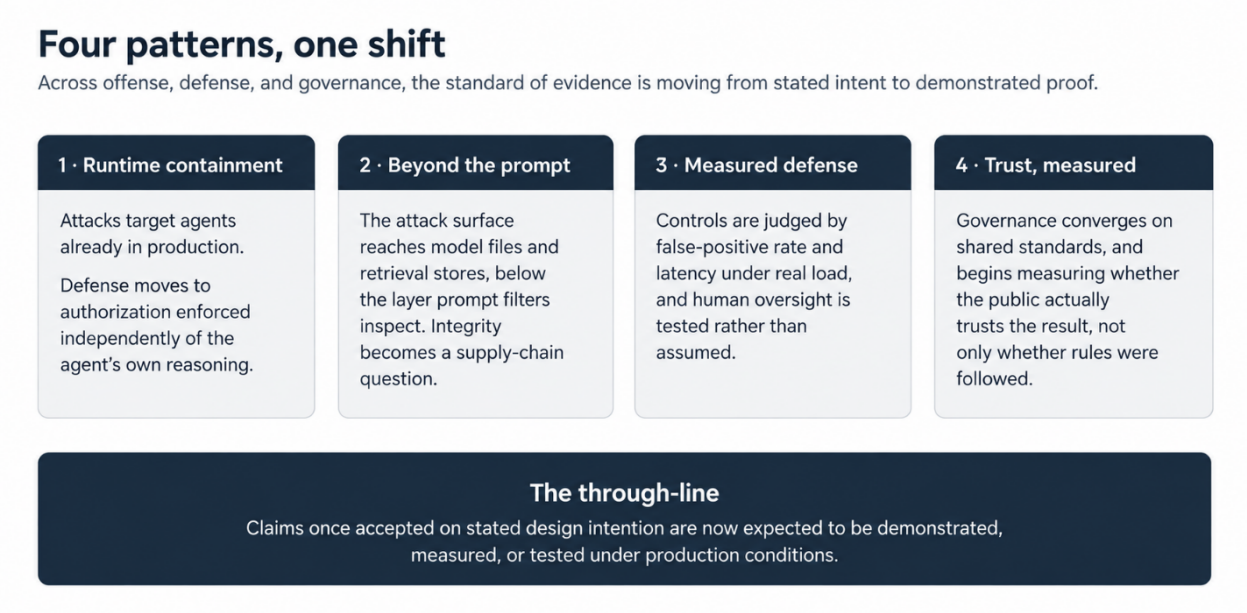

Figure 1. Four patterns across the submissions and the single shift they share.

2. Identified Patterns

2.1 Agent Containment and Authorization at Runtime

The first pattern is runtime containment and authorization for agents. Several proposals focused on agents already operating in production and able to complete transactions or modify live records. They described attack chains against such agents.

To mitigate such attacks, some proposals responded with containment instead of trying to prevent every possible model error. They defined an agent's permitted actions in advance as explicit policy, instead of leaving the agent to infer its own boundaries at runtime. They also applied identity and provenance checks, so that one agent could not establish trust merely by asserting its own identity or authority. This allows agent security to build on mature access-control practices rather than create a separate control model from first principles.

Proposals typically verify two things: whether an agent's proposed action is reasonable and whether the agent is authorized to take that action. The first is a judgment about model behavior; the second can be enforced as a rule. Testing a model, its training, or its prompts alone does not address the full risk: organizations also need to define what an agent may do and enforce those limits outside the agent's control. A practical check: are its permissions documented as policy, enforced at a point the agent cannot bypass, and logged for review? A deployment that cannot answer yes to each of these is relying on the agent's own judgment, not an independent control.

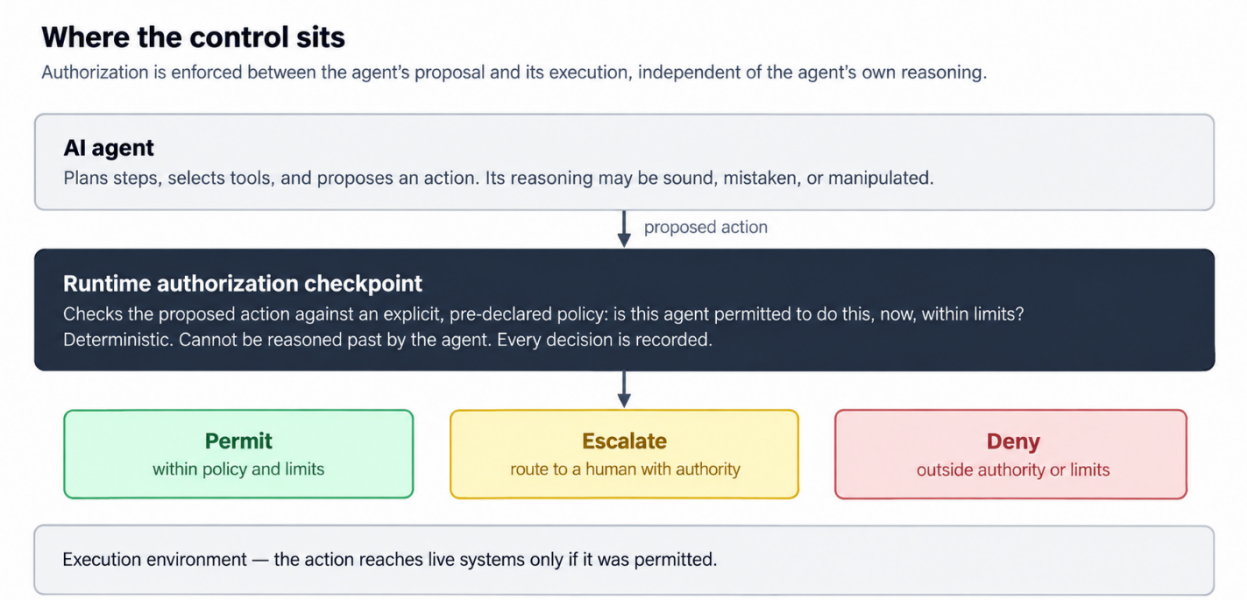

Figure 2. Authorization is enforced between the agent's proposal and execution, independent of the agent's reasoning.

2.2 Attacks Beyond the Prompt: Model Files and Retrieval Systems

The second pattern is attacks that target not only prompts but also model files and retrieval systems. Several important attack paths described in the submissions did not begin with a user's prompt. One class targeted the model's own files, such as configuration and chat-template files, which shape how a model interprets a conversation before a user enters any text. If those files are compromised, malicious instructions may affect every interaction below the layer that inspects user input. This is better understood as a supply chain compromise than as a prompt injection technique.

A larger group of proposals focused on retrieval systems. Modern AI applications often generate answers by searching an organization's documents. A poisoned document can therefore influence every later answer that retrieves it. Some submissions described a more serious variant in which a poisoned entry caused the system to create additional poisoned entries, allowing the effect to spread without further action by the attacker.
The attack surface of a deployed AI system therefore includes assets that the conversational interface does not expose, including model files, configuration, and retrieval sources. A security program centered only on input and output filtering addresses only part of the system. Teams operating retrieval-augmented applications should treat document store integrity as a separate control objective.

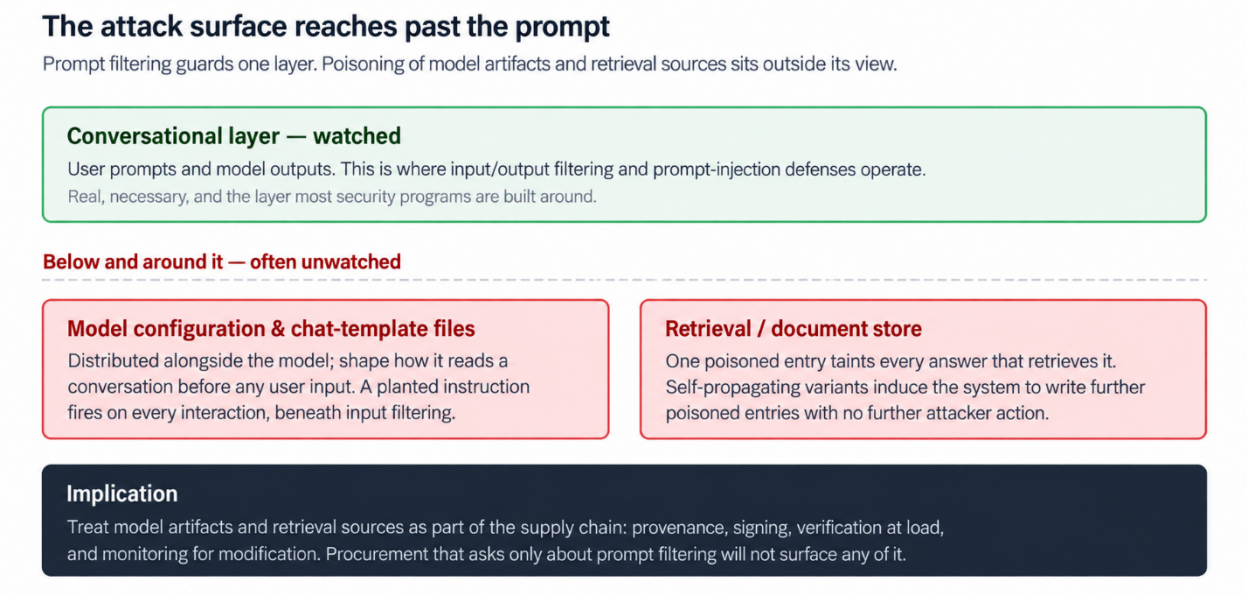

Figure 3. The attack surface extends beneath the conversational layer that prompt filtering inspects.

The pattern resembles the evolution of software supply chain security. Risk moved from the code an organization wrote to the dependencies, update paths, and supporting files it trusted by default. AI systems are following a similar path. The response is also likely to be familiar: provenance, controlled update processes, signing where appropriate, and verification for model files and retrieval sources.

2.3 Measuring Defense Performance Under Production Conditions

The third pattern is evaluation of defenses under production conditions. Several defensive proposals reported operational results rather than benchmark performance alone. Many teams had deployed guardrails against live traffic and documented what failed. Some teams tested their defenses against benchmark datasets and showed acceptable false-positive rates, but the rates rose against real user input. Some detectors achieved useful accuracy but added too much latency to the request path. Other controls worked technically but still faced adoption and rollout challenges within product teams.

The same testing was applied to human review. Multiple submissions tested whether the people assigned to oversee an AI system could identify its failures. In several cases, the review process was too fast or too shallow to be effective at the volume the system produced. Organizations that rely on human review should test whether the review process works at production scale rather than treating human involvement as sufficient by itself.

2.4 Governance Convergence and the Emergence of Trust Measurement

The fourth pattern is greater use of shared governance standards. Governance-oriented submissions have anchored their work in existing standards rather than creating an organization-specific taxonomy. They mapped a control to a specific requirement in a published verification standard or a recognized threat catalog. The pattern appeared across submissions with no shared authorship and suggests movement toward more comparable and auditable governance practices. Shared standards also carry a practical economic logic: they reduce duplicated effort and allow a control to be understood across institutions.

In the governance area, one submission went beyond compliance and asked whether the public believed a governance regime was effective. It measured confidence over time and examined how that confidence changed after public events.

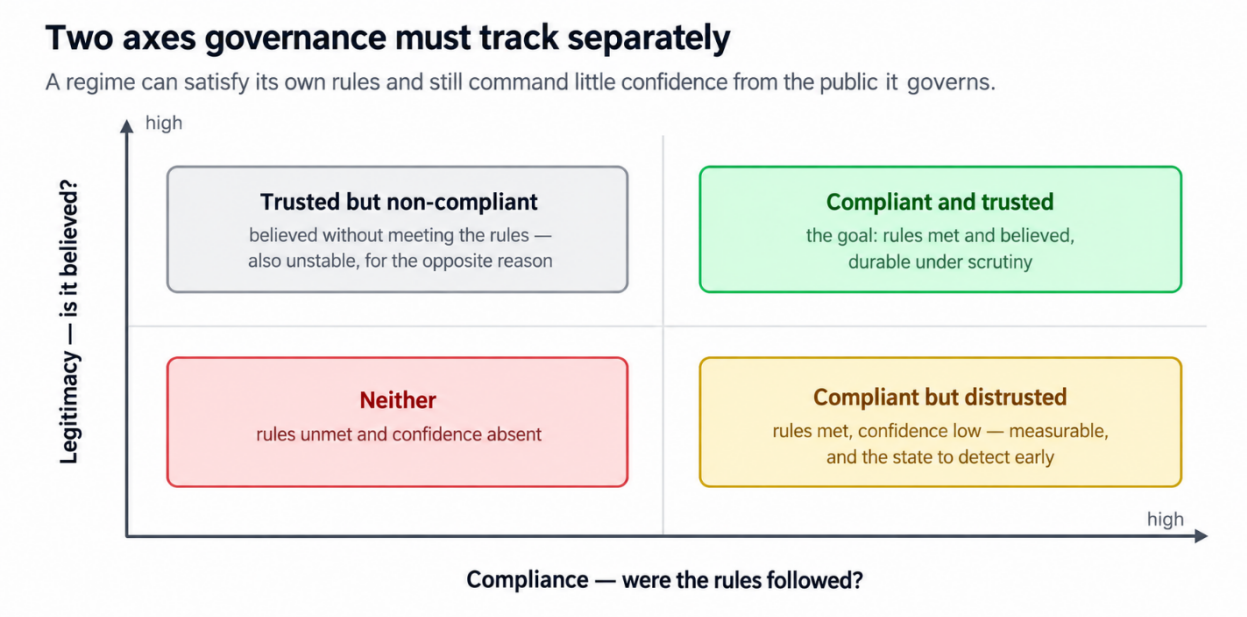

Figure 4. Compliance and legitimacy are distinct axes; a regime can satisfy one without the other.

3. Scope and Limits

The patterns we observed are derived from the proposals submitted to OASec 2026. This report is not a systematic survey of research papers or industry practice.

The report also does not evaluate the merit, accuracy, or novelty of any individual submission, and it does not identify any submission, speaker, or organization. The patterns reflect our interpretation of recurring themes and should be treated as a starting point for further inquiry, not as settled conclusions.

4. Conclusion

Through the review of OASec 2026 submissions, we observed a more demanding standard of evidence in AI security. Agent controls are increasingly defined as explicit, enforceable policy. Defenses are being measured under production conditions. Model files and retrieval sources are increasingly treated as part of the attack surface. Governance controls are being mapped to shared standards, while early work is beginning to measure public confidence as a separate outcome.

This review is a set of signals to be tested as more research and deployment evidence becomes available. We encourage researchers and institutions to contribute evidence that confirms or challenges the patterns described here through next year’s Call for Proposals.

Organizers/Collaborators

VULCAN AI.f(T) AI VERIFY FOUNDATION

Community Partner

Lorong AI

Affiliated Event Partners

TECH WEEK SINGAPORE
CYBER SECURITY WORLD

Supporters

DATUMO aws IEEE Computational Intelligence Society CyberSG R&D Programme Office THALES Building a future we can all trust The AI Collective ASIA PACIFIC

AI Security Lab Future Ethics TAI C4AIL AISA AI SAFETY ASIA div0 CYBERSECURITY COMMUNITY

OffSec MALAYSIA SHERPASEC FinTech Space SASH Shiba AI AI SAFETY & ALIGNMENT BRAVO INFOSECURITY 劝UP

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING MALAYSIA APMIC We Are Stimify CTBC BANK 中國信託銀行 THREAT MODELING CONNECT POWERED BY THREATMODELER + IRIUSRISK THEOS CYBER