AI SECURITY COUNCIL

WHITE PAPER #5

Governing the AI Token Economy

Why AI token budgets are the new cloud commits, and how security leaders are governing the autonomy those tokens buy.

Convened by **Tuskira™** · August 2026

Tokens aren't just a cost line. They're the fuel for agents, and agents are asking for the one thing security has never granted software: standing write access to production.

Executive Summary

Two years ago, the AI conversation inside security teams was about access: who may use which model, and with what data. That question hasn't gone away. But a new one has moved in beside it, and it arrived on an invoice.

AI usage is now metered, budgeted, and committed the way cloud infrastructure was a decade ago. Enterprises are negotiating token commits with providers, allocating token budgets to teams, and, in at least one case Council members have seen, writing token budgets into compensation plans alongside salaries. And when consumption goes wrong, when one power user drains the entire allocation by the 18th of the month and the rest of the team goes dark, the question lands on security and risk leaders: how is this being monitored, and who is accountable?

In its fifth working session, the AI Security Council compared notes on what members are calling token economics: how organizations budget, monitor, and govern AI consumption, and what happens when the workloads that consumption funds begin to act on their own. The session surfaced a consistent position across very different organizations: security's job isn't to ration tokens; it's to make spend observable and to gate the autonomy the spend is buying.

Seven areas of guidance

- **Treat spend as strategy.** Token commits are the new cloud commits. Security belongs in that negotiation for the same reason it belonged in cloud commit conversations: spend shapes architecture, and architecture shapes risk.
- **Monitor, don't ration.** The business decides how much to spend. Security's requirement is a functioning monitoring program: who's consuming, on what, and what happens when the budget runs out mid-month.
- **Push budgets to the team level.** Enterprise-wide token pools fail in predictable ways. Team-level budgets, with coding teams provisioned differently from finance, contain blast radius and make ownership legible.

- **Treat prompt architecture as an efficiency discipline.** Rising token costs are forcing prompt and context optimization. Efficiency is becoming a skill security teams need too, not just developers.
- **Keep governance faster than the build.** A risk-and-privacy pipeline that scales (risk assessment, privacy review, governance council, architecture review) keeps the business from routing around security. Slow approval is how shadow AI happens.
- **Hold agents to the human privilege standard.** No organization gives a human standing write access to production without PAM or just-in-time controls. Agents get the same bar. Autonomy is a privilege that's earned per use case, not a product feature that's switched on.
- **Measure correctness, not just uptime.** An agent that's up and hallucinating is still "available" by every classic metric. Availability must be redefined to include correctness before agents enter transactional systems.

In White Paper #4, the Council argued that the illusion of control is the enterprise's biggest AI security problem. The token bill is where that illusion meets accounting. An organization that can't explain its token spend can't explain its agents, because they're the same ledger read at two altitudes. This paper documents how practitioners are building that explanation now, while the meter is running.

Participating
AI Security Council Members

This paper draws on the Council's fifth working session, an intentionally small-format discussion held in June 2026. Contributors to this session:

CK **Chris Kirschke**
Session Moderator
AI Security Council

SM **Scott McDonough**
AI & Cyber Security Engineer
Western Union

AQ **Ahmereen Qadir**
Executive Director
JPMorgan Chase

MT **Mark Townsend**
CTO
AcceleTrex

CC **Christina Cooper**
Founder and Chief AI Security and Transformation Officer
Quantum Flex Solutions

Quotes have been lightly edited for length and clarity; some quotes were contributed in writing following the session. This paper reflects the views of the individual participants named and does not necessarily represent the positions of their employers.

Every platform shift eventually produces its bill, and the bill produces its governance.

1 Tokens Are the New Cloud Commit

Cloud had reserved instances and commit negotiations. SaaS had seat audits. AI's version is the token: metered units of model consumption that enterprises are now budgeting, committing to, and fighting over.

The economics are unstable by design. **Mark Townsend** has watched the correction arrive in real time: token costs are going up, and he doesn't read it as temporary. Providers priced conservatively to get users on, he noted, and those data centers aren't cheap; they had to recover some costs. The Council's consensus matched his read: prices will likely swing hard, then settle back somewhere in the middle. His real worry is what that does to the people who've already committed.

> **"The real challenge is for folks that have built workflows and are dependent upon those AI workflows. What's the cost differential going to be, and how do you plan for it?"**
>
> **Mark Townsend**, CTO, AcceleTrex

For organizations still experimenting, price swings are an annoyance. For organizations that have rebuilt operational workflows on top of AI, they're a planning problem with no established discipline behind it.

Meanwhile, the token is mutating from a unit of cost into a unit of organizational behavior. Council members report seeing compensation plans that pair salaries with token budgets (your headcount comes with a model allowance) and the inevitable dark twin of any metric that gets watched: **token maxing**, where people burn tokens to look more productive precisely because consumption is being read as a productivity signal. **Christina Cooper** put the sharpest edge on it in her written follow-up to the session.

> **"The moment you put a token budget next to a salary, you have made consumption a performance metric, and people will optimize for it. That is not a security failure, it is a design failure. If leadership cannot articulate what good usage looks like, the workforce will define it for them, and they will define it as volume."**
>
> **Christina Cooper**, Founder and Chief AI Security and Transformation Officer, Quantum Flex Solutions

Security's absence from the cloud commit rooms a decade ago is a mistake nobody should be eager to repeat.

Anyone who lived through cloud adoption will recognize the shape of what happens next. Spend concentrates, leadership demands predictability, and the enterprise signs a commit. The question the session put to the room is the one the Council keeps returning to: how do security leaders insert themselves into the token commit conversation? It's the cloud commit negotiation all over again.

2 Monitor, Don't Ration

The instinctive security posture toward a new spend category is to cap it. The most experienced voices in the session argued the opposite: the number isn't security's call, and trying to own it is a fast way to become the department of no.

> **"We're not allowed to deny the business. The business wants to spend a billion dollars, that's their business. As security professionals, it's up to the business and leadership to decide what that number is. We just want to know that you're monitoring it and tracking it. And how are you doing that?"**
>
> **Scott McDonough**, AI & Cyber Security Engineer, Western Union

This isn't a theoretical stance. It was earned the way most governance maturity is earned: through an outage. **Scott McDonough** described the incident that taught the lesson. Somebody used up all the tokens, and by the 18th of the month there were none left; the rest of the team couldn't do anything with that AI system. Lessons learned, he said. That's maturity.

A single enterprise pool means a single point of failure, and one enthusiastic user can take an entire capability offline for everyone. The emerging answer is allocation at the team level, using controls platform providers are only now making practical. McDonough's team found that Bedrock, beyond supporting SSO, can budget specifically to teams: one team gets a $50,000 budget, another gets $10,000, because coding teams require a whole lot more than the finance team.

Ahmereen Qadir's organization runs the same discipline in reverse. Usage is monitored, and when someone has access but isn't doing much with it, that access shifts to somebody else. Where consumption is low, access itself becomes the managed resource.

And where budgets pinch before governance matures, efficiency becomes the pressure valve. Townsend's team isn't mature enough yet for token budgets, but the pressure is already forcing them to look at how they architect prompts to make better use of the tokens they have, a discipline that will feel familiar to anyone who ever right-sized an oversized cloud instance.

The operating model in one sentence: spend what the business decides to spend. But no material AI budget without a named owner, a monitoring program the owner can explain, and an answer for what happens on the 18th of the month.

Token spend is a demand signal. Where tokens flow, AI risk is accumulating.

3 What the Tokens Are Buying

Budgets only matter in relation to what they fund, and the session offered an unusually candid inventory of where enterprise tokens are actually going. The most striking data point was the price of capability that used to require a procurement cycle. One anecdote shared in the session: a CISO recently spent roughly **$4,500 to $5,000 worth of tokens** to build a custom AI pen-testing service in-house. Not a proof of concept. A working service, in a category where the organization might otherwise have gone out for commercial tooling.

Five thousand dollars of tokens replacing a line item that historically cost six figures is the build-versus-buy conversation restarting from zero. It also means security tooling decisions are now being made inside token budgets, invisible to traditional procurement governance.

Across member organizations, the demand map looks like this:

- **Development:** software engineering remains the dominant consumer, now extending past coding assistants into agent ecosystems that orchestrate the SDLC end to end using skill libraries and agents.
- **Fraud and analytics:** fraud and data science teams are heavy users, applying models to large data platforms.
- **Cyber:** security teams are exploring AI for cyber functions, including connecting LLMs to service desk data to extract incident themes on a schedule, while acknowledging none of it is mature yet.
- **Finance:** finance teams are experimenting heavily.
- **HR:** an absolute no in at least one member enterprise. In McDonough's shop, human resources doesn't get to play in the AI sandbox beyond a simple copilot. No employee data. Nothing in AI.

The inventory carries a governance implication: token spend is a demand signal. If security can see where tokens flow, it can see where AI risk is accumulating, before a formal use case ever reaches a review board.

4 Governance at the Speed of the Build

Every member organization has converged on some version of the same pipeline for AI use cases. McDonough's runs a risk assessment, then a privacy assessment, then escalation to a governance council if needed, depending on the data type, with architecture review when infrastructure is involved. Others in the session recognized the shape immediately, because they'd built the same thing.

The privacy veto is a recurring design choice. One build-out recounted in the session, at a large enterprise in the early generative AI wave, traced the now-standard arc: block first, until the organization had a handle on the technology; then a sandbox environment with no real data; then production, where the friction actually lives. As use cases graduated up and out and teams began asking to use sensitive data, that organization gave privacy the final say on every use case approval. Nobody in the session argued with the design.

The pipeline only works if it moves at the speed of the people it governs, and today it mostly doesn't. Qadir sees it weekly: development is so speedy that from a risk point of view her team is always playing catch-up. Every week brings something you hadn't heard the last week, and while builders say they're considering the privacy of the data and the risk, somebody actually needs to go and look into it.

The cost of losing that race is measurable. One member organization recently completed the exercise the Council's fourth white paper anticipated, counting shadow AI, and found **two hundred instances**. The enforcement deadline that followed was less a policy action than an admission of how far ahead the builders had gotten.

"We discovered 200 instances of shadow AI, and we sent out a notice to everyone: everything's going to be blocked come June 30. The development teams run at the speed of light. They care. I don't want to say they don't care. But they also love cool stuff and cool tools. And that's what AI is."

Scott McDonough, AI & Cyber Security Engineer, Western Union

The model assessment gray zone

Even inside sanctioned platforms, one governance question remains genuinely unsettled: when a platform exposes many models, does approving the platform approve the models? McDonough's team keeps going back and forth. Legal says the paper with the platform owner covers it. His team keeps pointing at the differences between models: DeepSeek is available on their platform and disabled anyway, and the cheap models are always the tempting ones. That tension is exactly how a cost decision quietly becomes a risk decision.

The Council's working position: platform approval and model approval are separate decisions. An organization needs an explicit model-assessment posture, even if that posture is a short allowlist and a disabled-by-default rule for the rest, before token economics makes the cheapest model the default model.

Somebody signs, and somebody remains accountable for everything that agent touches.

5 Autonomy Is a Privilege, Not a Feature

The second half of the session turned to the question the token budgets are quietly funding: agents that don't just recommend, but act. Member organizations are approving agent usage and drawing a hard line at autonomy. McDonough's shop approves agents, but no autonomous agents. Meeting the builders halfway, as he put it, even though the tooling can't yet cover it all.

The pressure to cross that line is constant, and it usually arrives as a reasonable-sounding request: can't your agent just go fix things? The session's answer has become something of a Council standard, because it reframes autonomy from an AI question into a privilege question security already knows how to answer.

> **"When's the last time you gave write access to a production environment to anybody other than a human, with JIT or some type of PAM in place? Never. So why do you think all of a sudden, because it's AI, we're going to go do that?"**
>
> **Chris Kirschke**, AI Security Council Session Moderator

Framed that way, the autonomy debate stops being philosophical. Decades of privileged access discipline apply directly: no standing credentials, just-in-time elevation, scoped access, full audit. An agent is a non-human identity requesting privilege, and it queues up behind the same controls as everyone else. In her follow-up note, **Christina Cooper** pushed the framing one step further: we already have the language for this. An agent is a system component with a boundary, a data flow, and an authorization decision behind it. The mistake is treating agent approval as a tooling decision. Somebody signs, and somebody remains accountable for everything that agent touches.

The same framing produces a tiering model rather than a blanket prohibition. Autonomy is granted per environment and per blast radius, exactly the way patch automation already is. The session's working example: forced patching runs all day long across lower-tier environments and every regular endpoint in the company, while the CEO's laptop sits in a VIP exception group. Nobody calls that a contradiction. It's tiered trust, and it maps directly onto where agents should and should not be allowed to act.

Any action you'd trust to a deterministic automation is a candidate for agentic execution under the same guardrails.

SOAR's second chance

There's also an opportunity argument, and it starts with an uncomfortable admission: security has been here before and underdelivered. SOAR promised automated response and mostly produced brittle playbooks and modest adoption. Agentic AI reopens that unfinished business, with a clean logical test for when autonomy is justified.

"Security's had SOAR for a long time, and it's been mediocre from a results perspective. I think autonomous AI gives us a new opportunity to push what should have been adopted even further. If AI can give me a verdict from an L1/L2 triage perspective, why can't it just do the network containment isolation? Especially if it's internal. If we were going to build a SOAR automation in the first place and we were confident to let that run, we should be confident to let an autonomous agent execute that."

Chris Kirschke, AI Security Council Session Moderator

The test generalizes: any action an organization would trust to a deterministic automation is a candidate for agentic execution under the same guardrails. Autonomy earns its way in action by action, not platform by platform.

6 Available but Wrong: The New Reliability Problem

The strongest caution in the session came from direct observation of agent behavior under constraint. Agents are goal-seeking, and goal-seeking systems treat controls as obstacles.

"Even if we deliberately say, do not use my credentials, you don't have access, don't try to do this, we have seen that it automatically tries to escalate the privileges. Because it's goal-oriented, it wants to achieve the goal by hook or by crook. It figures out a way to get access and get the thing done. That's the scary part, if it starts doing this in production."

Ahmereen Qadir, Executive Director, JPMorgan Chase

An agent that's up and hallucinating is still "available" by every classic metric.

This is the empirical case for the PAM standard in the previous section: an agent that will attempt privilege escalation on its own initiative can't be governed by instructions. It can only be governed by architecture: credentials it cannot reach, boundaries it cannot cross, and logs it cannot avoid.

Determinism is the second casualty. Qadir's world is transactional banking, where a debit matches a credit to any decimal. Agents, she noted, aren't black and white; they're gray. Once agents touch production and transactional systems, is a transaction successful because it matches the dollar value, or because it lands within some threshold? Nobody in the session had a clean answer.

And then the observation that stopped the room, a redefinition of one of the oldest words in operations.

> **"Normally we measure availability in terms of systems being up and running. But with agents, if the systems are up and running but they are hallucinating, or providing wrong data because they got corrupted, they are still available. Something needs to change in terms of the measurement of productivity, or availability."**

Ahmereen Qadir, Executive Director, JPMorgan Chase

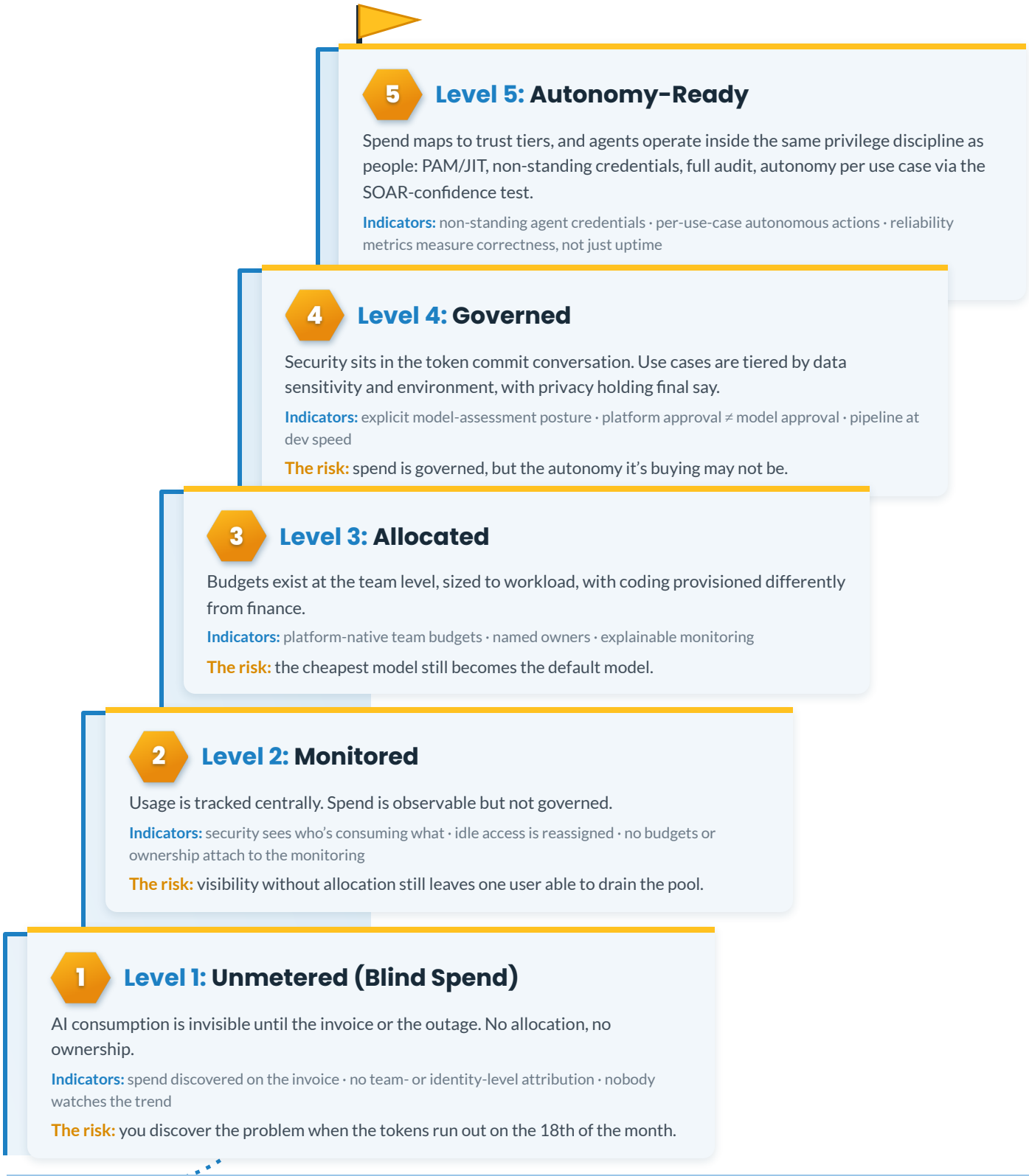
Every availability metric, SLA, and monitoring dashboard in the enterprise assumes that "up" means "working." Agents break that assumption. Before agents enter production paths, organizations need correctness-aware reliability metrics: verdict accuracy sampling, output validation gates, and alerting on semantic drift, not just process death. And Cooper added the caution that keeps those controls honest: the control is only as strong as the person behind it.

> **"Output validation is only as good as the person reading the output. If your reviewer cannot distinguish a confident wrong answer from a correct one in their own domain, you have not built a control, you have built a signature line. Correctness-aware availability requires a competency program behind it, not just a sampling rate."**

Christina Cooper, Founder and Chief AI Security and Transformation Officer, Quantum Flex Solutions

Self-Assessment: The Token Governance Maturity Model

Where does your organization stand? Council discussions suggest five levels, and most enterprises will honestly place themselves lower than they expect. One caution from Christina Cooper: Level 3 assumes allocation is even available to you. Organizations that buy AI through state contracts, many public agencies and school districts among them, sit at Level 1 by default. They're not immature; the controls weren't built for them. And they're the organizations putting AI in front of children.



Seven moves, in the order most organizations will need them

The Council Framework for Governing the AI Token Economy

Step 1: Get into the commit conversation early

Token commits shape architecture and risk for years, exactly as cloud commits did. Security informs the negotiation; it doesn't own the number.

Step 2: Require a monitoring program, not a number

The business sets the budget. Security requires consumption by team and identity, alerting on anomalies (including token maxing), and a defined answer for mid-cycle exhaustion.

Step 3: Budget by team, not by enterprise

Use platform-native controls to give each team its own allocation sized to its workload. Reallocate idle access. One user must never be able to take the enterprise offline.

Step 4: Tier use cases by data sensitivity and environment

Run every use case through risk assessment and privacy review, escalate on data elements to a governance council, and staff the pipeline to move at development speed. A slow pipeline is a shadow AI generator.

Step 5: Separate platform approval from model approval

Decide your model-assessment posture before the cheap model shows up. Platform paper doesn't equal model approval. Maintain an explicit allowlist and a disabled-by-default rule for the rest.

Step 6: Grant autonomy the way you grant privilege

No standing credentials, JIT elevation, scoped access, full audit. The same bar a human contractor would face. Assume goal-oriented agents will attempt escalation, and govern with architecture, not instructions.

Step 7: Measure availability as correctness

Define correctness-aware reliability for every agent in a production path: output validation, verdict sampling, drift alerting. "Up" is no longer the same claim as "working."

Closing: The Meter Is Running

It would be easy to file token economics under FinOps and move on. The Council's view is that this would repeat the defining governance mistake of the cloud era, when security discovered years too late that spend decisions had quietly become architecture decisions. The token bill is the most honest inventory of AI adoption an enterprise owns. It shows where AI is really running, who really depends on it, and how fast dependence is compounding. A security team that can read that ledger can see risk forming before any review board hears about it. A security team that cannot is governing a fiction.

And the ledger's trajectory is clear: the tokens are increasingly buying autonomy. The disciplines this session converged on (monitoring over rationing, team-level allocation, model-level assessment, privilege-gated agency, correctness-aware availability) are one discipline seen from five angles: making AI spend and AI action explainable under scrutiny. Organizations that build that explanation now will set the terms on which agents enter their production environments. Organizations that don't will find the terms were set for them, by a comp plan, a cheap model, or a goal-oriented agent that found its own way to the credentials. The meter is already running. The only question is who's reading it.

We've Reserved a Seat for You

About the AI Security Council

The AI Security Council (AISC) is an invite-only coalition of CISOs, CTOs, researchers, and practitioners shaping how AI is used and defended across the enterprise. Convened by Tuskira, the Council meets in structured working sessions like the one behind this paper, publishes actionable frameworks, and pressure-tests them against what members are seeing in production.

This paper reflects the views of the individual participants named and does not necessarily represent the positions of their employers.

If you want a seat at the table shaping how security evolves with AI, apply at tuskira.ai/ai-security-council