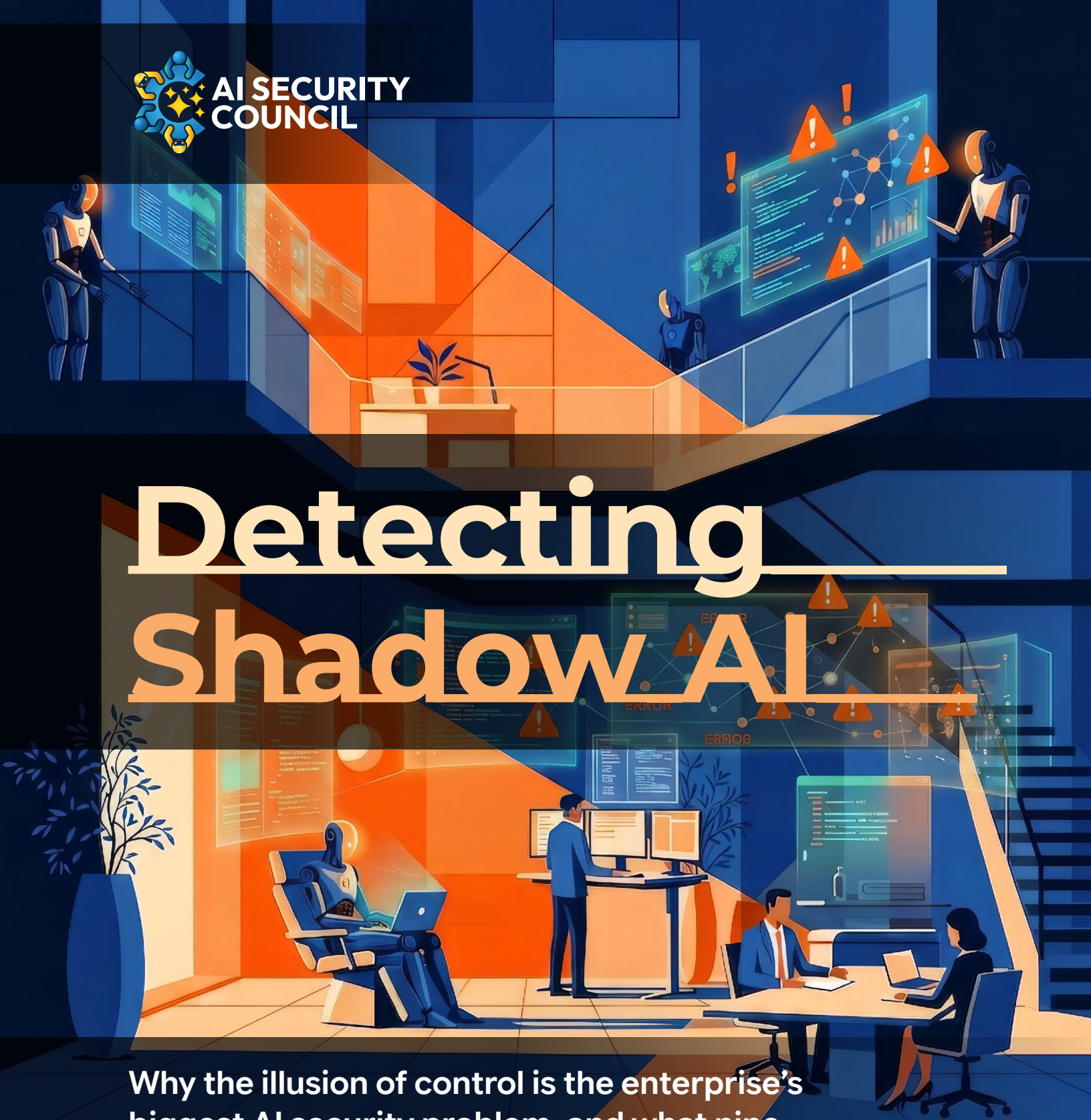
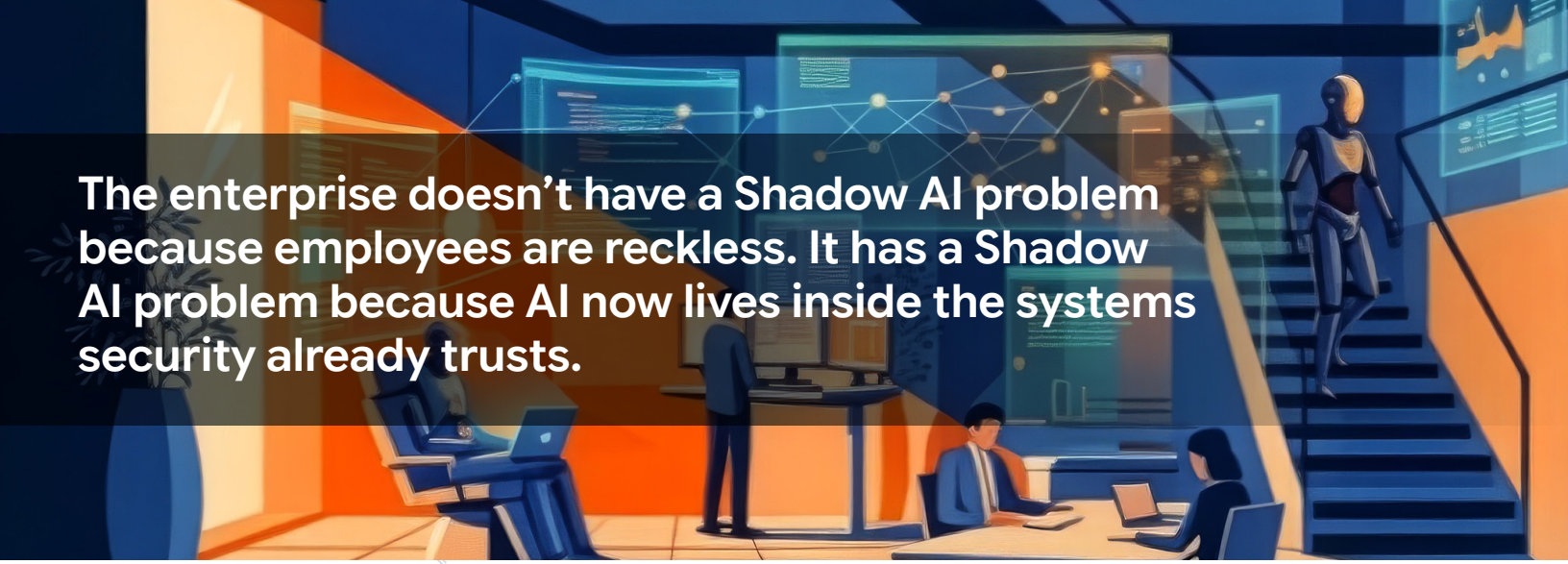
AI SECURITY COUNCIL

Detecting Shadow AI



Why the illusion of control is the enterprise's biggest AI security problem, and what nine security leaders are doing about it

WHITE PAPER #4



The enterprise doesn't have a Shadow AI problem because employees are reckless. It has a Shadow AI problem because AI now lives inside the systems security already trusts.

Executive Summary

It's 2026, and shadow AI has moved downstack, outward, and into the operational fabric of the enterprise. It lives in offline models running on personal Mac Minis with external GPUs. It lives in browsers, local LLMs, personal devices, and phones operating in airplane mode. It lives in sanctioned SaaS platforms quietly shipping embedded AI features that administrators may never have explicitly approved. And most dangerously, it lives within the identity and authorization debt enterprises have carried for years, a debt that AI agents are now inheriting at machine speed.

Across two working sessions of the AI Security Council on April 21 and 23, 2026, nine practitioners from Oracle Cloud, Baylor Scott & White Health, Synaptics, SNC-Lavalin / Atkins Global, Good Day Farm, Western Union, IR Proactive, Harvard, and AI Product Camp converged on a conclusion that cuts against much of the market noise:

Shadow AI detection is not primarily a problem of new tooling. It is a fundamental problem that new tooling exposes.

The organizations gaining ground are the ones re-grounding their defense posture in behavioral baselines, identity hygiene, whitelisting discipline, endpoint and browser telemetry, and a new operating assumption: every agent must be treated as its own actor.

This paper distills the council's field-tested guidance into seven areas:

1. The client-side blind spot
2. Extending insider-threat detection to agents
3. Operationalizing governance without becoming the bottleneck
4. Rethinking identity and authorization lineage for autonomous agents
5. The centralized-versus-federated telemetry debate
6. A maturity model for detecting and governing Shadow AI
7. The Council Framework for Detecting Shadow AI

This paper documents how experienced practitioners are already making that shift in real environments. The insights that follow are drawn directly from their debates and disagreements, with one goal: clarity about what it takes to move from AI adoption to AI accountability.

Participating
AI Security Council Members

AJ Debole
Field CISO
Oracle Cloud

Pratik Savla
Sr. Staff Analyst, Office of the CISO
Synaptics

Chris Zupa
Director, Cyber Defense
Baylor Scott & White Health

Gerard Johansen
Principal
IR Proactive

Pete Atore
AI Use Case Leader
AI Product Camp

Ryan René Rosado
Adjunct Faculty
Harvard

Scott McDonough
AI & Cybersecurity Engineer
Western Union

Sean Edgeington
Vice President of Information Technology
Good Day Farm

Emilio Mazzon
Director CSA, Global Information Security & IT
SNC-Lavalin / Atkins Global

1 The Problem Has Moved

Most organizations still think about Shadow AI as a visibility problem at the application layer. They imagine employees using unapproved AI websites, uploading sensitive information into consumer tools, or experimenting with unsanctioned chatbots outside policy.

That version of the problem still exists, but it's no longer the hard part.

The harder problem is that AI has moved into places traditional controls weren't designed to observe. It's embedded in SaaS platforms. It runs locally on devices, operates through browsers, and can be invoked via APIs, automation flows, copilots, and agents that inherit permissions from systems never designed for autonomous behavior.

The result is a false sense of control. Organizations can see the sanctioned layer, enforce policy against known tools, block domains, review vendors, and publish acceptable-use guidance. But the actual AI activity inside the enterprise is broader, more fragmented, and more deeply entangled with existing workflows than most governance programs assume.

That gap between perceived control and actual usage is where Shadow AI grows.

The client-side blind spot

Chris Kirschke opened the council discussions with a user who runs multiple small language models directly on a phone, enables Google Drive offline access, puts the device in airplane mode, and runs inference on corporate documents entirely outside network-layer control.

The same pattern applies to local models running on personal Mac Minis with external GPU enclosures. Users can download GGUF model files, process sensitive data locally, and bypass the controls that organizations typically rely on to monitor cloud-based AI usage.

The structural shift is clear: the insider threat model now has to include the agents, models, and local inference environments operating alongside the human.

Pratik Savla framed the detection challenge in terms of telemetry that organizations already collect but rarely tune for AI: unusual file-access patterns, screen-scraping behavior, local-model execution signals, sensitive-data path monitoring, and abnormal process behavior via EDR and browser controls.

> **"Already we've seen the use of EDR, browser controls, that could add to basically detecting any kind of abnormal behavior, any kind of user behavior or process behavior that's abnormal, especially tied to AI usage."**
>
> **— Pratik Savla**, Synaptics

The point was that the existing stack has to be retuned for a new class of behavior.

The browser becomes the control plane

AJ Debole connected the client-side problem to a workforce shift that many enterprises still underestimate. Millennials grew up app-centric; Gen Z is more browser-native and often resists installing apps. The browser is becoming the dominant workspace, which means it's also a rich source of telemetry for Shadow AI detection.

> **"A lot of this content and a lot of this activity is moving client-side, and a lot of it is being accessed by a browser. The importance of browser telemetry will only increase as Gen Z comes into the workforce."**
>
> **— AJ Debole, Oracle Cloud**

This makes browser instrumentation more than a niche control. It becomes a core security layer for understanding where AI is invoked, what data is accessed, and whether the activity aligns with expected behavior.

The sanctioned tool problem

The most deceptive surface is not the unapproved AI tool. It is the approved platform that quietly ships AI functionality into an existing workflow.

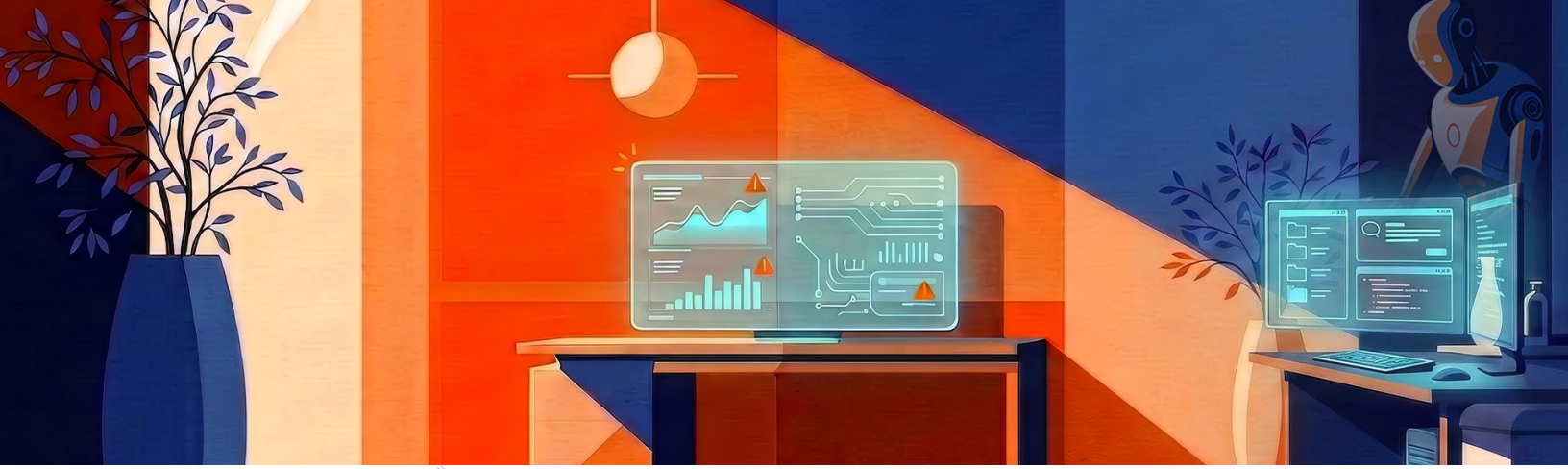
Scott McDonough flagged Microsoft Copilot as one of the hardest cases his team reviews because privilege elevation can occur silently inside familiar workflows. A Copilot agent that attaches to a Power Automate flow may inherit the flow's privilege level, often without security teams noticing until the capability is already operational.

Gerard Johansen made the same point in simpler terms: if an organization is a Windows shop, Copilot is effectively baked in. Employees are going to use it.

Chris Zupa widened the lens further. His concern was not only internal adoption, but adversarial acceleration. In healthcare environments, vulnerability remediation is already difficult. If AI compresses the time between vulnerability disclosure and exploitability, the burden on defenders expands dramatically.

> **"AI is ubiquitous, so that means we've got AI, and the adversary has AI. Vulnerability remediation, as it stands, is already a lot of work — with AI reducing the time to exploitability, it's only going to exponentially consume more time to keep pace."**
>
> **— Chris Zupa**, Baylor Scott & White Health

Shadow AI is harder to detect than shadow IT because it doesn't always look like a new tool. It often appears as a new capability inside something the enterprise already trusts.



2 Detection Starts With What You Already Own

The strongest consensus across the council was that most organizations already own much of the telemetry needed to detect Shadow AI. The problem is that the telemetry was tuned for a different era.

EDR can detect local model files on disk. It can identify unusual processes tied to inference workloads. It can detect abnormal file access patterns, GPU-related activity, and screen-scraping behavior. Network tools can surface upload anomalies, off-hours spikes, unusual ports, and private VPN usage. Browser controls can show where users are invoking web-based AI tools and whether sensitive data is moving through those sessions.

The issue is interpretation.

Retuning EDR, browser controls, and network monitoring

Chris Kirschke noted that EDR platforms can already detect GGUF model files, attached GPU devices, and unfamiliar local inference processes. Pratik Savla reinforced that many AI-specific behaviors resemble patterns organizations already care about: sensitive file access, screen scraping, abnormal process execution, and unusual data movement.

Sean Edgeington described how closer monitoring in his Cisco Meraki environment surfaced unauthorized private VPN usage tied to potential exfiltration paths. The alerts were technically possible before, but the tuning logic had been shaped around downloads rather than uploads and around a pre-AI threat model.

> **“When we did start monitoring that a little more closely, some of the things that started showing up were private VPNs. That was where we started saying, ‘hmm, what is this?’”**
>
> **— Sean Edgeington**, Good Day Farm

The broader lesson is that Shadow AI detection begins by asking whether existing tools are tuned to see the behaviors AI now enables.

The whitelist-first operating model

Emilio Mazzon described one of the most operationally disciplined approaches discussed by the council: treat AI traffic monitoring as a whitelisting pipeline.

Rather than investigating every suspicious pattern in an environment where most activity is legitimate, his team starts by identifying authorized traffic. Once sanctioned AI usage is known and whitelisted, it drops out of the review queue. What remains becomes more actionable by definition: unknown tools, suspicious connections, unexpected APIs, and misuse patterns.

> **"When you monitor traffic, you only find two types: authorized traffic or unauthorized traffic. And obviously, 99% of it will be authorized. Once you've gone through the hump of whitelisting all that traffic, you're only going to get misused traffic instead of use-case traffic. That's the traffic that's going to flush out."**
>
> **— Emilio Mazzon, SNC-Lavalin / Atkins Global**

This approach turns detection into exception handling. It also reduces the SOC's burden by removing known-good activity from the queue, rather than forcing analysts to reason through the same legitimate patterns repeatedly.

A new signal: impact density

Pratik Savla contributed one of the most important conceptual advances from the sessions: insider-threat detection has to evolve for agents.

Traditional user behavior analytics assumes a human user operating at human speed. Agents break that assumption. A compromised or misbehaving agent can move data, modify resources, call tools, or change configurations at a scale and speed that traditional anomaly models may not flag quickly enough.

The council's proposed metric is impact density: the ratio of consequential change to elapsed time.

If a human or agent performs a high volume of privilege-relevant actions, data movement, or configuration changes within a short time window, that combination should become a high-confidence signal. In the agent era, high impact over low time may become the equivalent of impossible travel.

Impact Density

A proposed detection metric for agent-era security: high-consequence change compressed into a short time window. The faster an actor creates a meaningful impact, the more aggressively the activity should be investigated, especially when the actor is an agent or non-human identity.

Impact density pairs naturally with just-in-time credentials, short-lived tokens, non-human identity registration, and continuous validation during execution. Together, those patterns begin to form a detection architecture suited for autonomous systems.

3 Governance Without Becoming the Bottleneck

Governance fails when the secure path is slower than the unsanctioned path. The council repeatedly returned to this point. Employees adopt Shadow AI because it helps them move faster. If security responds only with friction, employees will route around the process, often into channels that are harder to observe.

The answer is **operationalized governance**.

Why the classic blocklist model breaks down

Chris Kirschke recalled a prior experience with an AI-service blocking feature that effectively operated as a static list of approved and denied domains. It created immediate support friction, blocked legitimate users, and did little to address the harder governance issue: whether a given AI tool or use case had actually been reviewed, approved, and integrated into the organization's risk process.

Gerard Johansen described the pace problem directly. The market is moving too quickly for yearly or quarterly review cycles.

> **"Instead of a yearly or quarterly review, we're now in a constant, almost weekly change-control cycle. As new models drop, as new tools drop, we're assessing: can we load this for an enterprise environment?"**
>
> **— Gerard Johansen**, IR Proactive

This is where governance must change shape. It can't remain a meeting, a ticket, or a static spreadsheet. It has to become an operating workflow.

Three workflows that reduce governance friction

The council surfaced three practices that help governance keep pace without collapsing into exceptions.

- **First,** AI-assisted anomaly triage can reduce the time analysts spend gathering context. **Pratik Savla** described workflows where an agent ingests a suspicious event, pulls adjacent telemetry, explains why the event appears anomalous, and hands a human analyst a decision-ready summary. The goal is not to remove the human. It is to remove the manual context-building that delays the decision.
- **Second,** agents need just-in-time, context-bound credentials. Across both workshops, the council converged on a clear pattern: every agent should have its own non-human identity, short-lived credentials, scoped permissions, and context-bound authorization tied to workload identity, cloud account, region, source IP, device, or action class.
- **Third**, finance telemetry should become part of the detection stack. **Pete Atore** argued that token spend and cloud cost are often the earliest signs that an agent is misbehaving.

> **"Early detection of something wrong, especially with tokens going wild or agents going wild in some fashion, pops up in the dollar figures first."**
>
> **— Pete Atore, AI Product Camp**

Ryan René Rosado extended the idea from her consulting work, noting that dollar-threshold alerts on unexpected compute or token usage can serve as a useful interim control for organizations without mature agent governance. It is not a complete posture, but it is materially better than no telemetry.

The council's view was pragmatic: perfect governance isn't the first milestone. Earlier detection, faster context, and safer pathways are.

4 Identity Is the Problem that Sinks the Ship

Every discussion of Shadow AI eventually turned into a discussion of identity, because AI agents inherit the access models they are deployed into. If those environments are over-permissioned, poorly segmented, or burdened by years of service-account sprawl, agents do not create the problem. They accelerate it.

Here's a metaphor that resonated across the sessions: human identity is the visible tip of the iceberg; non-human identities are what sink the ship.

Agents break authorization lineage

AI agents can call tools, trigger workflows, spawn sub-agents, and operate across systems. At the moment delegation occurs, the chain connecting the original human intent to the resulting action can become unclear.

Scott McDonough described the operational problem directly:

> **"Very few, if anyone, has figured out how to follow the identity of an agent. The agent's a proxy of some human somewhere — and how are we monitoring that? That's really the big challenge."**
>
> **— Scott McDonough, Western Union**

His team has begun pushing back on agent deployments by asking basic but essential questions: What is the agent doing? What identity does it carry? What systems can it touch? How do you kill it if it behaves unexpectedly?

Those questions expose a gap many development teams have not yet had to answer.

You cannot solve agent identity on top of identity debt

Gerard Johansen delivered the hardest truth of the workshop: organizations can't secure agent identity without first addressing the identity debt underneath it.

Across customer engagements, he's seen the same pattern repeatedly: Active Directory environments that were never cleaned up, standing service-account privileges no one remembers granting, stale OUs, unmanaged access paths, and years of accumulated authorization exceptions. When an agent enters that environment, it inherits the debt.

> **"You're giving semi-autonomous code the ability to do a lot of things because we really haven't thought about it, and we have all of this debt over the last ten years on just the identity piece. You're not going to lock this down until you get your basics clean first."**
>
> **— Gerard Johansen, IR Proactive**

This may be the most important practical warning in the paper. Agent security does not start with agent tooling. It starts with cleaning up the identity layer the agent will inherit.

The architectural primitives that matter

The council aligned around several identity principles:

- Agents should receive non-human identities, not human credentials.
- Credentials should be short-lived, scoped, and context-bound.
- Privileges should be granted just-in-time, not held persistently.
- Sensitive actions should require step-up authentication or additional approval.
- Where role-based access control lacks granularity, attribute-based access control should be considered.
- Every agent should be logged and monitored as a separate actor, not as an invisible extension of the invoking user.

Emilio Mazzon emphasized the importance of logging agents separately. Without that separation, tracing authorization lineage becomes nearly impossible when something drifts.

The Unresolved Problem: Multi-cloud Agent Identity

AJ Debole named a gap that remains unresolved across the industry: multi-cloud agent identity management.

Each major cloud provider has credible native identity controls inside its own environment, but enterprises rarely operate in a single cloud. Most large organizations run across multiple clouds, SaaS platforms, identity providers, and legacy systems. A coherent identity wrapper for agents across that entire surface does not yet exist.

AJ pointed to Oracle Cloud's Zero Trust Packet Routing as an architectural idea worth extrapolating from. It adds an attribute-based layer on top of network controls, requiring additional tags before privileged subnet communication is allowed. The same concept applies to agents: beyond identity and network access, agents need contextual attributes that demonstrate they are authorized to perform a specific class of action in a specific context.

The council also flagged MIT Media Lab's NANDA initiative as worth watching, as it aims to create a decentralized registry and identity fabric for AI agents. It remains early-stage research infrastructure, but it reflects the right direction: agent identity will need to become a foundational security primitive rather than an afterthought.

> **"There's a middle ground where big organizations dipping their toe in the water with AI agents are going to hit the moment they say: the tip of the iceberg with user-identity controls is dwarfed by what's required for the agents. How do we create this internet for these agents with the controls built in?"**
>
> — AJ Debole, Oracle Cloud

5 The SIEM Isn't Dead, But the Reasoning Layer Is Moving

Is the SIEM still the center of gravity, or are we moving toward a federated, MCP-driven model where agents query security tools directly and pull only the data they need?

The final major debate across the workshops centered on telemetry architecture.

Is the SIEM still the center of gravity, or are we moving toward a federated, MCP-driven model where agents query security tools directly and pull only the data they need?

The council's answer was not either/or. It was both.

Scott McDonough, who previously worked in the SIEM market, argued that the SIEM is not going away. It will be redefined, it will absorb AI capabilities, and it will increasingly rely on machine assistance to process log volumes no human team can review manually.

> **"The SIEM is here to stay, it's just going to redefine itself, become more sophisticated, and use AI in some manner to help with that logging, monitoring, and chewing through a billion logs in an afternoon."**
>
> **— Scott McDonough, Western Union**

Sean Edgeington offered the practitioner's counterpoint. The SIEM may exist, but many medium-sized organizations do not have the team size to meaningfully interpret its output. Alerts fire, whitelists drift, repeated blocks eventually get through, and teams often lack the bandwidth to determine whether the pattern was legitimate activity or early-stage compromise.

The group predicted that by the end of 2026, the enterprise security graph will be mature enough for reasoning layers to sit above existing systems and query across context to determine whether a given event is truly a security issue.

That framing reconciles the debate. The SIEM remains the data substrate because centralized logging has architectural gravity. Federated queries and agent-driven investigation move up the stack as reasoning layers. Organizations should expect to invest in both.

The real issue is not where the logs live. It is whether the organization can reason across identity, endpoint, browser, SaaS, network, cloud, finance, and HR context fast enough to act.

6 Self-Assessment: The Shadow AI Detection Maturity Model

Shadow AI maturity is not defined by whether an organization has approved AI tools. It is defined by whether the organization can detect, contextualize, govern, and explain AI activity across sanctioned and unsanctioned surfaces.

Use this model to assess where your organization stands.

Level 1: Assumed Control

At this stage, the organization believes AI use is largely governed by policies and approved tools. Visibility is limited to known applications, known domains, and voluntary disclosures.

Common indicators:

- AI policy exists, but is not connected to telemetry
- Detection depends on known domains, app inventories, or user reporting
- Local models, browser-native usage, and embedded SaaS AI features are largely invisible
- AI risk is discussed in governance forums, but not measured operationally
- Agents are not treated as distinct identities or actors

The risk: The organization isn't blind, but it is seeing only the easiest layer to observe.

Level 2: Reactive Visibility

The organization has begun detecting some unauthorized AI usage, usually through CASB, proxy logs, EDR alerts, browser controls, or manual investigation. However, signals are inconsistent and often lack context.

Common indicators:

- Some unsanctioned AI tools can be detected
- Alerts are reviewed on a case-by-case basis
- Browser and endpoint telemetry exist, but are not fully tuned for AI-specific behavior
- AI-related incidents depend heavily on analyst interpretation
- Governance reacts after discovery rather than shaping usage upstream

The risk: The organization can find some Shadow AI, but cannot reliably determine which activity matters most.

Level 3: Contextual Detection

At this level, AI activity is correlated with identity, data sensitivity, endpoint behavior, browser telemetry, network flows, and workflow context. The organization begins to distinguish benign productivity usage from activity that materially changes risk.

Common indicators:

- EDR is tuned for local model files, inference processes, unusual file access, and screen scraping
- Browser telemetry is treated as a core detection source
- AI traffic is categorized into authorized and unauthorized patterns
- Private VPN usage, abnormal uploads, and suspicious API calls are monitored
- Sensitive data movement is evaluated in context, not as a generic event

The risk: Detection improves, but governance may still depend on manual escalation and fragmented ownership.

Level 4: Managed Agentic Risk

The organization extends insider-threat detection to agents and non-human identities. Agents are registered, scoped, monitored, and evaluated as distinct actors rather than extensions of human users.

Common indicators:

- Agents receive non-human identities rather than human credentials
- Credentials are short-lived, just-in-time, and context-bound
- Impact density is used to detect high-consequence activity compressed into short time windows
- Agents are logged separately and monitored continuously
- Finance telemetry, token spend, cloud cost, and HR context are integrated into detection logic
- Identity debt remediation is underway before broad agent deployment

The risk: The organization is managing agentic risk, but multi-cloud identity and cross-system authorization lineage may remain incomplete.

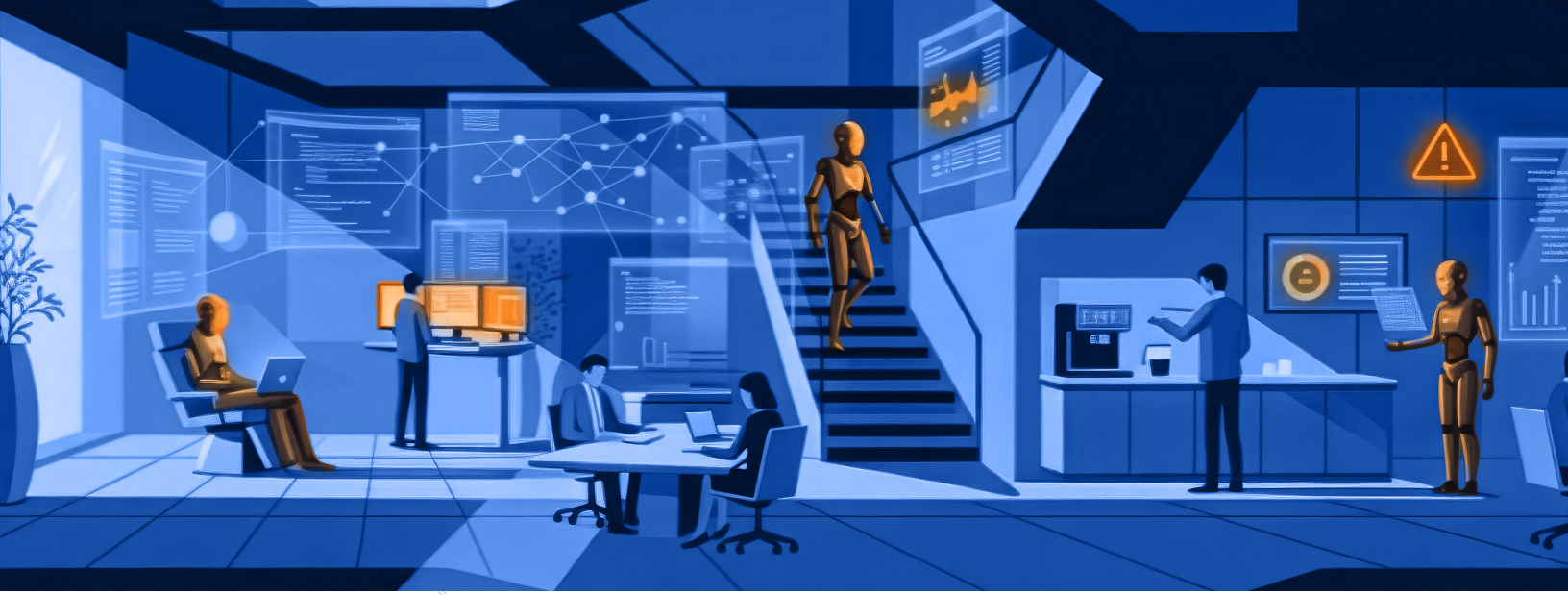
Level 5: Adaptive Governance

At the highest level, Shadow AI detection becomes part of a continuous governance loop. The organization does not rely on static policy or periodic review. It continuously observes usage, updates baselines, adapts controls, and provides evidence to customers, regulators, auditors, insurers, and internal leadership.

Common indicators:

- AI usage is observable across endpoint, browser, SaaS, cloud, identity, finance, and network layers
- Authorized AI traffic is whitelisted continuously, leaving anomalous activity in the review queue
- Secure AI pathways are easier to use than unsanctioned alternatives
- Agent identity, authorization lineage, and action history are reviewable
- Governance adapts as tools, models, and workflows change
- The organization can explain not only which AI tools are used, but also how AI influences decisions and risk

The maturity signal: The organization no longer treats Shadow AI as a hidden exception. It treats AI usage as an observable, governable part of the operating environment.



7 The Council Framework for Detecting Shadow AI

The workshops produced a practical framework for organizations that want to move beyond policy and into operational detection.

Step 1: Instrument what you already own

Before evaluating new categories, tune EDR, browser controls, network monitoring, and SaaS telemetry for AI-specific behavior. Look for GGUF model files, local inference processes, attached GPU devices, sensitive file access, screen scraping, abnormal uploads, private VPN usage, and unusual API destinations. Most organizations are already collecting more Shadow AI telemetry than they realize. Very little of it is tuned.

Step 2: Operationalize whitelisting as a pipeline

Use protocol analyzers, API scanners, and traffic monitoring to enumerate AI-adjacent activity. Whitelist authorized traffic and remove it from review. Treat what remains as the actionable queue. This reduces noise and gives analysts a more defensible starting point.

Step 3: Extend insider-threat detection to agents

Adopt impact density as a first-class signal. High-consequence change compressed into a short time window should trigger review, especially when the actor is a non-human identity or agent. Pair this with NHI registration, just-in-time credentials, scoped permissions, and separate agent logging.

Step 4: Fix identity debt before scaling agent identity

Clean up service accounts, stale permissions, standing privileges, overbroad access groups, and unmanaged authorization paths before deploying agents broadly. An agent inherits the environment it operates in. If that environment is over-permissioned, the agent becomes an accelerator of existing risk.

Step 5: Wire finance and HR telemetry into detection

Token spend, cloud cost, P-card activity, travel status, PTO, and employment context can increase signal fidelity. Cost anomalies may reveal agent loops or runaway automation before traditional security telemetry does. HR context can help distinguish legitimate travel from impossible travel and reduce false positives.

Step 6: Build for centralized and federated reasoning

Do not assume the SIEM disappears. Do not assume centralized logging solves everything either. The future is likely a hybrid model: SIEM as substrate, federated queries as context retrieval, and reasoning layers that connect telemetry across tools fast enough to support action.

Closing: The Fundamentals are Back

Every major security category of the past twenty-five years tried to solve a version of the same problem: how do you correlate enough context fast enough to act before something bad happens?

EDR, XDR, NDR, SOAR, CASB, CSPM, DSPM, and SIEM all advanced the industry. None has fully solved the problem in complex enterprise environments. That is not an indictment of the tools. It is a description of the operating reality.

AI does not change that reality. It compresses it.

It compresses the time between discovery and misuse. It multiplies the identity surface. It turns software from a passive system into an active participant. It allows agents to operate at a speed and scale that traditional monitoring models were not designed to absorb.

That is why the council's guidance keeps returning to fundamentals: instrument what you have, tune for AI-specific behavior, clean up identity debt, treat every agent as its own actor, and stop waiting for a silver bullet.

Ryan René Rosado captured the gap between the agentic future vendors are describing and the environments in which many organizations still operate today:

> **"After RSA, hearing about the new agentic SOCs and all these new tools — I'm not anti-that. I just think the state of reality is that there are still organizations that don't have half these basic tools for a traditional environment, let alone a modern, innovative one. How do we close that gap?"**
>
> **— Ryan René Rosado, Harvard**

That is the question this paper leaves with security leaders. Whether the organization will continue to mistake partial visibility for control, or whether it will build the instrumentation, identity discipline, and governance loops required to make AI usage observable, explainable, and defensible.

Shadow AI isn't a separate security problem. It's the place where old security debt meets the speed of autonomous systems, and that is why it matters now.

We've Reserved a Seat for You

About the AI Security Council

The AI Security Council is an invite-only coalition of security and technology leaders, including CISOs, CTOs, researchers, and practitioners, convened by Tuskira to shape how AI is used and defended across the enterprise. The council meets in structured working sessions, publishes field-tested frameworks, and hosts follow-on webinars featuring workshop participants.

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 here.

AI SECURITY COUNCIL