AI SECURITY COUNCIL

WHITE PAPER #3

AI SECURITY PROGRAM TRANSFORMATION & THIRD-PARTY VALIDATION

Inside the Rebuild of Security Programs for an AI-First World

AI Security Program Transformation & Third-Party Validation

This paper captures the unscripted perspectives of CISOs, security architects, and technology leaders who participated in the third AI Security Council workshop, focused on AI security program transformation and third-party validation.

Across private enterprise, regulated industries, and the public sector, security leaders face a structural shift. AI is no longer experimental. It's embedded in production systems and decision workflows. Yet most organizations still cannot define what an AI-transformed security program actually is, much less prove it under scrutiny.

Customers, regulators, insurers, and acquirers are no longer satisfied with assurances. They're asking for evidence:

- What AI systems exist?
- Who owns them?
- What risks do they introduce?
- What controls are in place?
- How are those controls tested?
- How can any of this be independently validated?

The workshop explained that AI transformation is about redesigning security programs so AI systems are accountable, testable, and explainable under scrutiny.

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

Alex Shamsutdinov
CISO
Upwork

Charles Gifford
CISO, VP Information Security
Intrado

Derek Smith
Founder and President
Certify IT Academy

Jeff Kirby
CISO
Interstate Batteries

Keith Hartung
Chief Security Officer
PA State Treasury Department

Lena Smart
Ambassador
AIUC

Mark Townsend
CTO
AcceleTrex

Sean Dobson
CTO and CISO
Wafra

Sean Edgeington
Vice President of Information Technology
Good Day Farm

Victor (Min) Xu
AI Security
TikTok

Aliyana Isom
Cybersecurity Operations Program Lead
Nike

Bakul Singhal
Information Security Architect
Steve Madden

Christopher Carmona
AI Security Architect
TDSynnex

David Bosomworth
IT Director
Oakley Capital

Dirce Hernandez
Sr. Cybersecurity GRC Manager
Raices Cyber

John Stauffacher
Incident Response Leader

Vaibhav Narula
Cybersecurity Architect
Banner Health

The Problem Nobody Can Answer Cleanly Yet

Every security leader is hearing the same mandate: your program must be AI-ready. The phrase appears in board discussions, regulatory guidance, procurement questionnaires, and investor diligence. It sounds straightforward, though it isn't.

Alex Shamsutdinov was blunt about the starting point. AI is already operational inside security programs. It influences detection logic, triage decisions, response prioritization, and vulnerability analysis. The issue is that most organizations have not redesigned their security programs to reflect AI's role. AI is treated as a tool, not as part of the control surface.

That misalignment becomes visible when accountability is questioned. **Charles Gifford** described it as a translation gap. Security teams understand AI risk internally, but they struggle to articulate it in a way third parties can validate. The result is a widening distance between what is happening inside the environment and what can be proven outside it.

Keith Hartung brought public-sector clarity to the discussion. Oversight bodies don't accept intent as evidence. If AI influences decisions, it must be governed. Documented programs, named owners, testable controls, and auditable outcomes are expected regardless of whether AI was deployed deliberately or emerged organically.

Across the workshop, three questions surfaced repeatedly:

- What does an AI-transformed security program actually mean?
- How do you know when you have reached it?
- How do you prove it to someone outside your environment?

This paper follows practitioners who are already being forced to answer those questions. Most organizations are already operating AI-influenced security programs without redesigning them. That's not transformation. It's exposure.

When AI Becomes Infrastructure

The first realization shared across the workshop was that AI has moved from being an add-on to security programs to becoming part of the infrastructure itself.

Jeff Kirby described AI as already embedded in daily security operations, influencing how alerts are prioritized, how investigations unfold, and where analyst attention is directed. Others in the workshop extended this observation beyond the SOC. When AI systems shape triage logic, enrichment workflows, and investigative sequencing, they inevitably influence downstream decisions about remediation and risk acceptance. In that model, AI isn't peripheral. It participates directly in decisions that shape outcomes, even when humans remain in the loop.

Aliyana Isom extended this from a security operations program lens. When AI influences triage, enrichment, and escalation logic, it not only accelerates workflows but also changes them. Runbooks evolve. Escalation paths shift. Assumptions about analyst review points move. AI becomes part of how the SOC operates, not just a tool inside it.

When AI systems influence containment decisions or investigative sequencing, incident response timelines and outcomes can change materially.

John Stauffacher brought the CSIRT reality into focus. When AI systems influence containment decisions or investigative sequencing, incident response timelines and outcomes can change materially. If those AI-driven decisions are not understood, logged, and reviewable, response integrity is weakened. Incident management must account for the system's role in shaping the event.

David Bosomworth added that AI's influence extends beyond workflow acceleration into permission expansion. Security teams often already have broad system access. When agentic solutions are layered on top of over-permissioned environments, the blast radius multiplies. In this model, AI doesn't just assist decision-making; it inherits the operator's privileges and scales them. Without re-examining identity boundaries and data access models, AI becomes a force multiplier for existing excess permissions.

Victor Xu pushed the discussion further upstream from a product security standpoint. AI risk begins at design. By the time security teams are observing AI behavior in production, critical decisions about data use, autonomy thresholds, and failure modes have already been made.

Christopher Carmona added an architectural perspective. Once AI systems are embedded across identity, data, infrastructure, and code pipelines, they require architectural review, not just tool review. Treating AI as a feature ignores its systemic influence. Treating it as infrastructure forces a different level of design scrutiny.

This exposes the first structural fault line. Most security programs were designed to protect systems that behave predictably. AI systems do not. They adapt, drift, and respond to inputs that security teams cannot fully enumerate or control. Yet governance models, control frameworks, and assurance mechanisms are still largely built for deterministic systems.

The result is a structural mismatch. Security programs may look modern on paper, but they're still built for deterministic systems. AI is no longer just a feature inside a tool. It increasingly participates in control decisions, yet governance models haven't adapted accordingly.

What "AI-Transformed" Actually Means in Practice

When workshop participants were asked what an AI-transformed security program actually looks like, the answers were consistent, even though their environments weren't. The defining shift was how the security program treats AI itself.

Lena Smart reframed the discussion by treating AI as a first-class asset class rather than a pilot or innovation initiative. In this model, AI systems must be inventoried, tiered by risk, governed across their full lifecycle, and integrated into enterprise risk management alongside cybersecurity and data risk. AI is managed with the same discipline applied to financial systems or critical infrastructure, with named owners, documented controls, and explicit accountability.

An AI-transformed program is one that can explain itself under scrutiny.

Charles Gifford grounded this framing in CISO reality. An AI-transformed program is one that can explain itself under scrutiny. It can clearly articulate what AI systems exist, who owns them, what risks they introduce, and how those risks are controlled. Transformation shows up when the program can produce evidence on demand, not when it can describe intent. Without that ability, governance remains internal and unprovable.

Dirce Hernandez extended this into governance execution. Transformation requires AI to be embedded in existing risk registers, lifecycle documentation, and compliance workflows, rather than tracked in parallel systems. If AI risk sits outside core GRC processes, governance remains fragile. When it's embedded in them, it becomes durable.

Alex Shamsutdinov reinforced how this plays out operationally. In transformed programs, AI is expected. Analysts assume it will be present in workflows, and leaders assume it will influence outcomes. The cultural signal matters. When AI literacy is isolated to specialists, the program depends on individuals. When it is distributed, it scales.

Bakul Singhal closed the loop from an architectural standpoint. For governance to hold, AI controls must be integrated into reference architectures and design standards rather than evaluated on a case-by-case basis. When architectural guardrails define how AI systems are built and connected, governance becomes enforceable rather than advisory.

Throughout the discussion, one test kept surfacing because it was difficult to avoid. If an organization can't name its high-risk AI systems, identify their owners, and show when those systems were last tested under adverse conditions, it hasn't transformed its security program. It simply adopted AI and hoped the existing model would hold.

Knowing You've Reached It Without Declaring Victory Too Early

Defining an AI-transformed security program is difficult. Knowing when you have actually reached it is harder. Workshop participants agreed that maturity doesn't announce itself. It becomes visible in how the program behaves under pressure.

Mark Townsend described one signal as the shift from episodic tuning to continuous learning. In more mature programs, incidents, detections, and even false positives feed back into models and controls as part of normal operation. Improvements are not dependent on quarterly reviews or manual retuning cycles. The system adapts in rhythm with the environment.

Jeff Kirby pointed to operational consistency as another indicator. When AI-driven workflows produce predictable outcomes across shifts, teams, and incidents, maturity is emerging. When results vary widely depending on who is on duty or which tool is consulted first, the program is still compensating for instability.

In immature environments, every new AI use case becomes an exception requiring negotiation.

Sean Edgeington emphasized governance friction as a practical signal. In immature environments, every new AI use case becomes an exception requiring negotiation. In more mature programs, teams move faster because guardrails are already defined. Fewer escalations occur, not because risk is ignored, but because it has been addressed upstream through standardized patterns.

Sean Dobson framed maturity through resilience. Metrics alone are insufficient. The meaningful question is whether the program maintains control when models drift, prompts are manipulated, or data conditions change. Detection and remediation times may improve, but the stronger signal is whether those improvements persist during unexpected stress.

Vaibhav Narula added a healthcare perspective where tolerance for error is structurally lower. In safety-sensitive environments, maturity is reflected in predictability. AI systems must degrade safely, escalate clearly, and never obscure accountability. Transformation is evident when failure modes are anticipated and bounded rather than discovered in real time.

Self-Assessment:
The AI Security Maturity Scale

Is your program transforming, or just adopting? Use this checklist to determine your current state.

Level 1: Adoption (Reactive)

- ☐ We maintain a basic inventory of known AI tools in use.
- ☐ Security reviews are reactive and occur after a tool is already in the environment.
- ☐ Accountability is undefined; AI is treated as a generic software asset.
- ☐ There is no formal testing of AI behavior under adverse conditions.

Level 2: Governance (Active)

- ☐ An AI usage policy exists and has been communicated.
- ☐ Every new AI use case requires a manual, one-off security review.
- ☐ AI risks are documented, but outside core enterprise risk workflows.
- ☐ Risk is understood internally but difficult to communicate to third parties.

Level 3: Transformation (Standardized)

- ☐ We have defined "Golden Paths," standardized architectures with pre-approved identity, data, logging, and monitoring guardrails.
- ☐ Engineering teams can deploy within these patterns without manual security bottlenecks.
- ☐ AI risk is integrated into the enterprise risk register and lifecycle documentation.
- ☐ AI systems are risk-tiered with defined owners and review cadence.
- ☐ Failure modes are identified and linked to predefined containment playbooks.

Level 4: Accountability (Validated)

- ☐ Every production AI system has a named Business Owner and a formal "System Record" including purpose, data classes, autonomy level, and residual risk.
- ☐ AI systems are tested under adversarial and failure scenarios prior to and after deployment.
- ☐ Independent testing or red-team exercises are conducted on high-risk AI systems.
- ☐ Drift detection, monitoring thresholds, and rollback mechanisms generate automated evidence artifacts.
- ☐ Audit logs are complete, attributable, and retrievable without manual reconstruction.
- ☐ We can provide customers, regulators, insurers, or acquirers with structured evidence of governance without ad hoc preparation.

Across environments, programs that are truly transforming measure success by how well the organization responds when AI doesn't work. When failures are absorbed without confusion, ownership disputes, or ad hoc controls, the program has moved beyond experimentation and into something more durable.

Translating AI Risk So Others Can Validate It

As the discussion moved from internal maturity to external scrutiny, a common problem surfaced quickly. Security teams understand AI risk within their own environments, but that understanding often breaks down when they communicate it to someone outside the system.

Charles Gifford framed this as a failure of translation rather than a lack of control. Security teams can explain AI risk in technical detail, but those explanations rarely survive contact with customers, regulators, auditors, or acquirers. Too much technical depth obscures what matters. Too little reduces the conversation to assurances that cannot be independently validated. In that gap, trust erodes.

If AI risks aren't categorized consistently, they can't be measured.

If they can't be measured, they can't be validated.

Expressing risk in terms of likelihood and impact forces clarity without requiring disclosure of proprietary details.

Several participants emphasized that third parties are not asking to inspect model internals. They're asking for evidence that risk is understood, bounded, and managed. **Derek Smith** described this by starting with a clear risk taxonomy. If AI risks aren't categorized consistently, they can't be measured. If they can't be measured, they can't be validated. Expressing risk in terms of likelihood and impact forces clarity without requiring disclosure of proprietary details.

Christopher Carmona extended this idea by tying technical AI risks directly to business outcomes. Prompt injection, model drift, or data leakage are not meaningful in isolation. What matters is how those failures translate into operational disruption, regulatory exposure, or customer harm. Third parties validate outcomes and controls, not architectural elegance.

Aliyana Isom emphasized that risk translation must extend to operational impact. If AI influences enrichment, prioritization, or containment sequencing, those changes must be documented and explainable. Otherwise, the organization can't confidently demonstrate how decisions were made during an incident.

Victor Xu and **Sean Dobson** reinforced the same point from a product and investment perspective. External reviewers don't need access to weights, prompts, or training data. They need structured artifacts that show intent, scope, and control. System records that document purpose, data classes, autonomy level, testing results, monitoring thresholds, and rollback behavior allow outsiders to validate the program without replicating the environment.

John Stauffacher also highlighted a structural validation challenge: many commercial AI systems do not retain usable audit logs. If chat history cannot be extracted or actions cannot be traced to identities with confidence, traditional evidentiary models break down. Validation requires not just policy, but technical observability.

The programs that struggle most aren't those with the highest AI risk, but those that can't explain their controls in a way others can independently confirm. Validation doesn't require perfect systems. It requires structured evidence.

When AI Fails, Responsibility Sharpens

Failure wasn't treated as a hypothetical in the workshop. It was treated as an inevitability. The disagreement wasn't about whether AI systems would fail, but about what would happen when they did.

Keith Hartung was direct about the public-sector reality. Oversight bodies don't accept ambiguity when systems influence decisions, outcomes, or public trust. Whether an action was taken by a human or an AI system is irrelevant. Responsibility must be clearly assigned before failure occurs, not debated afterward.

Across the discussion, participants converged on a shared ownership model.

- Business owners are accountable for the use case, impact tolerance, and acceptance of residual risk.
- Engineering teams are responsible for secure design, testing, deployment, and operational resilience.
- Security and risk organizations define the control framework, threat models, and assurance requirements.
- Vendors carry accountability through explicit contractual obligations, not implied trust.

If ownership only becomes clear during a post-incident review, the program is already behind.

Lena Smart emphasized that this clarity must be encoded into the program itself. AI governance charters, RACI matrices, incident response playbooks, and procurement language should all reflect how responsibility is distributed when AI systems misbehave. If ownership only becomes clear during a post-incident review, the program is already behind.

Mark Townsend reframed failure as a diagnostic tool. In mature programs, incidents don't trigger blame shifting or ad hoc controls. They expose gaps in process, testing, or assumptions. Post-mortems map failures to systemic weaknesses where the goal isn't to prevent all failure, but to ensure failures are contained, understood, and learned from.

John Stauffacher challenged the assumption that AI failures are recoverable in the traditional sense. Once an AI system exports data, provides incorrect guidance, or acts outside policy boundaries, the harm is often irreversible. The containment model breaks down when the output itself becomes the incident. In this view, resilience is less about rapid recovery and more about preemptive control of the blast radius.

The clearest signal of maturity in these transformed programs is when no AI system reaches production without a named owner, a defined evaluation path, and a documented escalation model. Responsibility is explicit before the first incident, and failure makes the absence of accountability impossible to ignore.

Moving Fast by Standardizing

No participant disputed the pressure to move quickly as AI adoption is accelerating across product teams, security operations, and business units. The tension is operational. How do you allow experimentation and speed without creating a patchwork of inconsistent controls?

Bakul Singhal described the early trap many organizations fall into: treating every AI use case as architecturally distinct. Controls are re-evaluated from scratch. Reviews become manual checkpoints. Governance turns episodic. The result is slower delivery and uneven enforcement of standards. Without architectural baselines, speed and consistency cannot coexist.

The alternative described in the workshop was deliberate standardization. **Sean Dobson** and **Sean Edgeington** both emphasized building reusable patterns rather than bespoke reviews. Pre-approved components for identity, data handling, monitoring, guardrails, and logging create a "golden path" that product and engineering teams can follow without repeated negotiation. When the secure path is also the fastest path, adoption accelerates without sacrificing control.

Pre-approved components for identity, data handling, monitoring, guardrails, and logging create a **"golden path"** that product and engineering teams can follow without repeated negotiation.

Lena Smart framed this shift as industrialization rather than restriction. Governance shouldn't be a meeting that blocks deployment. It should be a pipeline that runs continuously in the background. Intake, sandbox testing, red-teaming, approvals, and production monitoring become standardized stages. Controls are embedded in workflows, not bolted on at the end.

Several participants stressed that flexibility still matters. Teams must be able to experiment with new models, prompts, and architectures. The constraint applies at the point of deployment. Innovation can happen at the edges. Production must pass through known patterns with defined control baselines and monitoring requirements.

Dirce Hernandez reframed speed not as acceleration, but as bounded exposure. Even if failure can't be undone, blast radius can be limited through segmentation, identity scoping, and predefined incident playbooks specific to AI systems.

The programs that move fastest are not the ones that skip controls. They are the ones that have already decided what the controls are. Standardization reduces friction by making expectations clear, reusable, and automated. Speed becomes a byproduct of preparation, not a justification for exception.

When AI adoption outpaces security, risk accumulates silently. When security industrializes AI development, scale becomes manageable. The difference is whether the guardrails are repeatable.

What Customers, Regulators, Insurers, and Acquirers Will Ask Next

By the end of the workshop, the conversation had moved beyond internal program design. The more pressing question was external: what will stakeholders demand next, and how should security teams prepare now?

Participants agreed that the direction is already visible. Customers are beginning to ask for AI-specific security language in contracts and RFPs. They want clarity on data handling, model governance, testing practices, and ongoing monitoring. General statements about secure development are no longer sufficient when AI systems influence decision-making or customer experience.

Regulators are moving toward documented AI risk management programs with traceable oversight. **Keith Hartung** noted that oversight bodies don't wait for perfect standards before expecting structure. They look for evidence of governance, defined roles, audit trails, and escalation paths. Board-level visibility into AI risk is increasingly expected, not optional.

Insurers are beginning to differentiate between organizations that experiment with AI and those that manage it systematically. Questions are shifting toward adversarial testing, model risk assessments, drift detection, and supply chain transparency. Coverage decisions will increasingly depend on demonstrable controls rather than self-attestation.

Acquirers and investors are applying similar scrutiny. During diligence, they are looking for defensible inventories of AI systems, mapped dependencies, and documented risk controls. An organization that can't explain how its AI systems are governed introduces uncertainty into valuation, integration, and long-term risk modeling.

Preparation shouldn't begin when the questionnaire arrives. Organizations should build and maintain an AI system register, formalize governance forums that include security, risk, legal, and business stakeholders, and pilot external assurance mechanisms before they are mandated. Evidence is easier to assemble incrementally than under deadline pressure.

Across stakeholder groups, the pattern is consistent. The next wave of questions will be about how well AI is governed, how transparently it is documented, and how confidently it can be validated by someone outside the organization.

Security leaders who treat AI governance as an internal optimization will be surprised by how quickly external expectations change. Those who treat it as a program of record will be prepared with answers rather than explanations.

From Adoption to Accountability

The workshop didn't end with sweeping declarations or claims that AI security has been solved. Instead, it revealed that organizations are moving from adopting AI tools to institutionalizing accountability for AI systems.

Transformation, as described by participants, is visible when AI systems are inventoried, risk-tiered, assigned to named owners, evaluated under adverse conditions, and embedded inside established control frameworks. It shows up when governance is integrated into lifecycle processes rather than bolted on after deployment. It becomes undeniable when failures expose process gaps instead of ownership confusion.

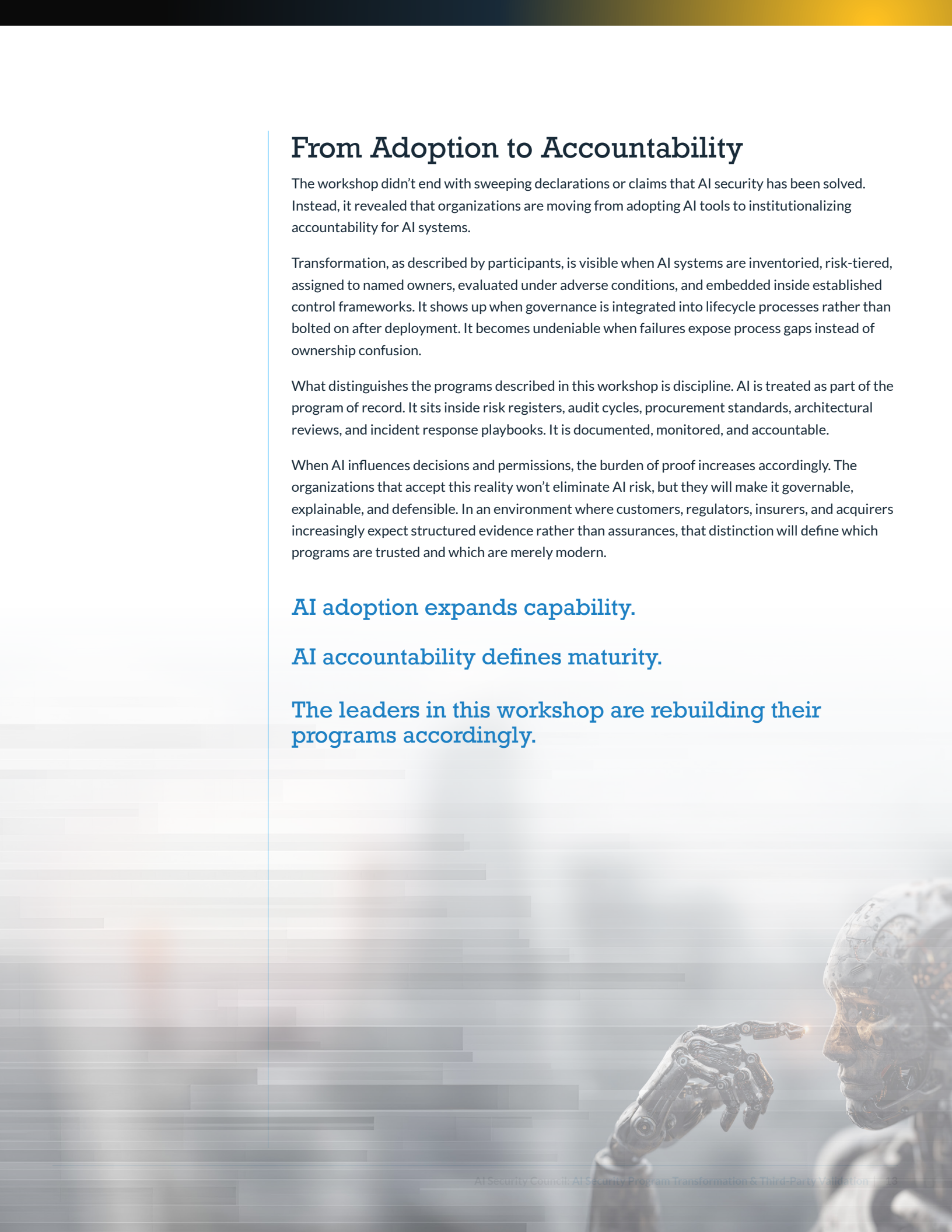
What distinguishes the programs described in this workshop is discipline. AI is treated as part of the program of record. It sits inside risk registers, audit cycles, procurement standards, architectural reviews, and incident response playbooks. It is documented, monitored, and accountable.

When AI influences decisions and permissions, the burden of proof increases accordingly. The organizations that accept this reality won't eliminate AI risk, but they will make it governable, explainable, and defensible. In an environment where customers, regulators, insurers, and acquirers increasingly expect structured evidence rather than assurances, that distinction will define which programs are trusted and which are merely modern.

AI adoption expands capability.

AI accountability defines maturity.

The leaders in this workshop are rebuilding their programs accordingly.



We've Reserved a Seat for You

Join 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. It's where security leaders anticipate adversarial AI, share field-tested strategies, and publish actionable frameworks for defending modern environments.

If you want a seat at the table shaping how security evolves with AI, **apply here**.

AI SECURITY COUNCIL