The Six Building Blocks of Agentic AI Trust – Part 1

The Enterprise Computing Stack Is Being Rebuilt Around a New Trust Model…
The Next Twelve Months Will Decide Who Controls It

Author: Casey Plunkett, Co-Founder and CEO, Secure AI LLC · June 2026

Contact: casey.plunkett@trustwaire.ai | @CaseyPlunkettAI on X

"We Shape Our Tools, and Thereafter They Shape Us."
- Marshall McLuhan

Executive Summary — Part 1: A Tale of Two Trends

The Invisible Hand is at work in establishing guardrails for Agentic AI. In parallel, enterprises are deploying this technology at scale whether the trust model is ready or not. Gartner projects the average Fortune 500 enterprise will manage over 150,000 agents by 2028. Their customers, employees, regulators, and boards will soon demand proof that it is trustworthy. The enterprises that build that proof deliberately will lead. The enterprises that don't will be exposed when the first ungoverned agent makes a consequential mistake.

Vendors across the enterprise computing stack are simultaneously rebuilding around a new framework because the model that worked for human users at human speed cannot govern autonomous AI agents at machine speed. This is happening across six layers:

- The silicon layer
- Endpoint OS and platform
- Co-located inference layer
- Shared cloud inference layer
- Agent identity and authorization

- Cross-enterprise commerce

The fifth layer is the focus of this paper, and the pace of change in that layer alone illustrates the urgency. Yale's Chief Executive Leadership Institute published the most authoritative case yet for a new form of agentic AI governance. They arrived at the same conclusion on the direction this paper defines. IBM shipped HashiCorp Vault's native agentic capabilities alongside Verify and Vault 2.0 with Sovereign Core. Cisco announced plans to acquire Astrix Security. Strata continued extending Maverics. A wave of well-funded behavioral observability startups entered the market. Expect massive acceleration of vendor innovation. Then assess the impact on your organization.

What was an identity and access management problem has become a multi-stakeholder trust problem, organized around six interdependent building blocks: Governance, Visibility, Transparency, Control, Automation, and Auditability. Governance is the foundation: it requires that business intent originate in the business process design tools where the work is modeled, not in IT-authored policy applied after the agent is deployed. No vendor currently delivers that. No platform encompasses all six. Until now.

Section I surveys the first megatrend and the forces driving it. The remaining sections address the challenges and gaps in the second. Your trust model will only be as strong as the weakest of the six. A companion paper, Part Two, describes the architecture that delivers all six as an integrated product — TrustwAIre™.

I. One Architectural Movement, Visible at Every Layer of the Stack

At the silicon layer, *Microsoft* made its Pluton security processor a prerequisite for the Copilot+ PC designation, explicitly requiring a hardware root of trust as the entry point for AI workloads on endpoints. *Intel* built a dedicated Partner Security Engine into Core Ultra Series 2 processors specifically to run Pluton. *Apple's* Private Cloud Compute uses the Secure Enclave as its hardware attestation root, cryptographically ensuring that user devices will only send data to server nodes running verified software. These are not legacy security modules repurposed for AI. They are new silicon capabilities, shipping in 2025 through 2026, designed to anchor trust in hardware because software-asserted trust roots can be spoofed, patched, or bypassed.

At the endpoint OS and platform layer, Copilot+ PCs ship with Pluton enabled by default as secured-core PCs, extending the silicon trust root into the operating system for AI workloads. Federal agencies are deploying them specifically for ITAR compliance and CMMC alignment on AI-capable endpoints. Microsoft's Agent 365, generally available as of

May 2026, extends Entra, Purview, and Defender governance specifically to AI agents running on these platforms. Dell, Lenovo, and HP are shipping AI PCs designed to host both classical productivity and agentic AI workloads on a single endpoint. If AI workloads run on a separate endpoint or in an ungoverned partition, the enterprise maintains two trust models, two attack surfaces, and no unified governance view. Convergence onto one hybrid platform with integrated AI governance is an economic inevitability.

At the co-located inference layer, when the data being inferred is the data of record (patient records, transaction ledgers, classified operations), inference must run where the data lives. For example, IBM's Spyre accelerator co-locates AI inference with transactional data on Z. This legacy platform has been reinvigorated by AI: the company reported a sixty-seven percent year-over-year increase in Z mainframe revenue in Q4 2025, with full-year IBM Z revenue at its highest level in roughly twenty years.

Oracle's bare-metal OCI offers sovereign deployment for regulated workloads. Qualcomm's AI200 and AI250 data center inference accelerators ship with confidential computing built in as a standard feature for enterprise AI. Trust moves into the system of record where the data already lives, because that is the only place where verification is technically possible.

At the shared cloud LLM inference layer, a regulator, examiner, or court asking "how was this data processed" needs an empirical answer, not a contractual one. Shared cloud LLM infrastructure cannot deliver that answer for Tier 3 and Tier 4 enterprise data. Contractual attestation (DPAs, SOC 2, ISO 27001, the standard "we don't train on your data" commitment) describes policies, not execution. The rise of dedicated and sovereign inference offerings from every major cloud provider is the industry's admission that the shared model has a verification gap no contract can close. For Tier 3 and Tier 4 data, dedicated inference infrastructure is not a premium option. It is a requirement."

At the agent identity and authorization layer, traditional identity and access management was built for humans with relatively static roles, logging into systems at human speed, with a human in the loop for consequential decisions. Autonomous agents break every one of those assumptions. They make thousands of credentialed decisions, at machine speed, with no human reviewing each action. The authorization model, the credential lifecycle, the audit trail, and the observability model all fail under these conditions.

Current monitoring tools can detect that an agent is behaving anomalously. They cannot determine whether the agent is doing what it was approved to do, because they have no access to the governance baseline that defines what "approved" means. Many innovators are racing to address this problem, but none solves it in its entirety. I will return to this in Section III.

And at the cross-enterprise commerce layer, when your agent transacts with another organization's agent, neither organization's internal trust model extends to the other. The protocols being designed to enable cross-boundary agent commerce (the Agent-to-Agent and Model Context Protocols, the payment-rail and identity layers underneath them) all assume the existence of a trust layer that does not yet exist at industry scale. This layer is the least mature of the six. It is also the one with the highest long-term stakes: whichever firms own the trust root at the enterprise edge will, by extension, own the trust infrastructure for cross-enterprise agentic commerce.

Google's Universal Commerce Protocol and Agent Payments Protocol, announced at I/O 2026, represent the first concrete cross-enterprise agent commerce standards — initially for consumer commerce, with enterprise implications that are not yet defined.

Six layers, simultaneously responding. The remainder of this paper focuses on the fifth, because it is where trust in AI is built or broken.

II. In Agentic AI, Secrets Are Becoming Yesterday's Technology

An *uncomfortable truth*, especially to many of my long-time colleagues. I assert that Secrets Management — the storage, rotation, and lifecycle administration of persistent credentials — was built for a world in which credentials *persist*. Humans need to log in repeatedly. Roles and access to entitlements are relatively static. Applications need to authenticate to databases for months at a time. Service accounts run for years. The entire secrets management industry (i.e., HashiCorp Vault, CyberArk, AWS Secrets Manager, Azure Key Vault) exists because persistent credentials are inherently dangerous and need lifecycle protection between rotations.

Agentic AI breaks this model at the root. An agent does not need to remember a credential across sessions. It needs a credential for one operation, one resource, one time window. When the operation completes, the credential dies. There is nothing to store, nothing to rotate, nothing to exfiltrate. For resources within a service mesh, the concept of a secret is irrelevant. The PKI credential authenticates directly through mTLS. For legacy resources outside the mesh, the credential is ephemeral rather than persistent: issued on demand, scoped to one operation, expired in seconds. Either way, persistent secrets are eliminated from the agent's operational path

> *Insight Check: The question is not how to manage agent secrets. The question is why an agent should have a persistent secret at all.*

Commenting on my “alma mater”, I’ll further assert IBM’s May 2026 HashiCorp’s Vault announcement is, read carefully, an admission of this. Vault has had a PKI secrets engine for years and can natively issue X.509 certificates. But that is not the default pattern enterprises use for agent credentials. The default inheritance path is Vault tokens, and the capabilities Vault is now shipping (ephemeral authorization, per-request scoping, tokens that expire after the task) are an attempt to make token-based authorization behave like PKI.

Vault is adding transaction context to JSON Web Tokens to narrow scope. But agentic AI requires identity and authorization bound in a single credential — one cryptographic object that says who this agent is, what it is allowed to do, and when it expires, verified without a callback. X.509 certificates issued by SPIFFE and SPIRE do this natively, at extreme scale. Layering authorization context onto JWTs approximates it, but the retrofit is constrained by the architectural assumptions of the underlying token model.

Public key infrastructure, in the form of mathematically-verifiable X.509 certificates issued at machine speed with a one-time-use credential, is the right technology for agentic credential issuance. PKI certificates validate locally, without a callback to the issuing authority to confirm the signature. They expire automatically via time-to-live, user-defined windows, or through workflow-bound expiration. Short credential lifetimes are themselves the revocation strategy: a credential compromised mid-operation is valid only until its time-to-live expires, a trade-off the architecture accepts deliberately because that lifetime is measured in seconds, not the months a persistent credential survives. They scale horizontally because verification is decentralized. They impose no central bottleneck. They are the right answer for a world in which a thousand agents make five million daily decisions, each of which requires a credential that lives for the duration of one operation and then dies. What PKI eliminates is persistent credentials from the agent's operational path, which is where the scale problem lives.

The CA signing key is itself a persistent secret, and protecting it is a known, bounded discipline: hardware security module storage, short-lived intermediate certificates, and the hardware roots of trust described in Section I. That is one key, protected in one place, versus millions of agent credentials distributed across the fleet. PKI concentrates the secret into a defensible perimeter; secrets management distributes it across an attack surface that grows linearly with every agent deployed.

The trajectory is clear. Every resource that moves into a service mesh eliminates another secret from the agent's operational path. The persistent credential lifecycle that secrets management was built to protect is being retired, one resource at a time.

This is not a critique of Vault or of secrets management as a category. Both will continue to be essential for human and persistent-application credentials, and for delivering ephemeral resource credentials to legacy systems in agentic AI deployments. The assertion is that the agent's governance credential, the cryptographic proof of authorization, should be PKI. The vendors and end users who recognize this, and build accordingly, will define the next generation of identity infrastructure.

The vendors who retrofit are asking their products to do what their architecture was not designed for. The operational processes built around secrets management — rotation, lifecycle administration, revocation checking, secrets scanning — were designed for hundreds of service accounts on schedules of hours to days. At agent scale, those processes must execute continuously at near wire speed. They cannot. If your agents inherit the persistent secrets model, every credential is a liability that must be managed, rotated, and protected. If they use PKI, the governance credential is gone before anyone can steal it. If they use PKI, the credential is gone before anyone can steal it. That’s the difference between a security operations burden that scales linearly with agent count and one that doesn't scale at all because there's nothing to manage.

III. The Six Building Blocks of Agentic AI Trust

Then what supplants the model that Secrets Management anchors? This domain is one piece of a broader model — identity, access, authorization, authentication, entitlements, governance — that was built for human users and persistent applications. I single out Secrets Management because persistent credentials sit underneath every layer of that model. Federation platforms like Okta, Ping, and Entra ID issue tokens, but those tokens are produced by signing keys that must be stored, protected, and rotated. Governance platforms like SailPoint and Saviynt execute their provisioning decisions through stored service account credentials connecting to every target system. PAM platforms are secrets management. There is no identity product in the enterprise stack that does not, at its foundation, depend on a persistent credential somewhere in its chain of trust. If that dependency is wrong for agentic AI, every layer built on top of it inherits the architectural mismatch. It is the Jenga piece that brings down the whole stack.

The answer is not a one-for-one replacement. Agentic AI requires a trust model built on six interdependent capabilities that together describe what every enterprise must be able to assert about every autonomous agent operating in its environment. The six are: Agent Governance, Visibility, Transparency, Control, Automation, and Auditability. What follows are my operative definitions for each, framed through the lens of Agentic AI. These

capabilities are not new in isolation. What is new is the requirement to deliver all six as an integrated system at agent scale, which no vendor or framework currently does.

1. Governance

Agent Governance addresses the questions of whether an agent should exist at all, who decided so, on what basis, with what review, and under what conditions. It is not access control. It involves establishing separation of duties across the **Trust Owner** (the individual(s) responsible for the Agent) and the **Trust Custodian** (stakeholders who assemble guardrails for the agent, in collaboration with the Trust Owner). It's the intersection of the business owner who needs the agent, the CISO who must accept the risk, the data owner who controls the sensitivity classification, and the compliance officer who maps the deployment to regulatory frameworks. Governance is the convergence of those four perspectives, not the property of any one of them.

All these stakeholders, along with the operations team at runtime, converge in real time on the same profile object, with separation-of-duty-defined access, full attribution of who changed what when, and bidirectional visibility across the agent's entire lifecycle.

As Lines of Business exponentially increase their use of agentic AI, bidirectional integration between the governance profile and the business process design and mining tools where the work is modeled will become essential. This includes vendors such as Signavio, Camunda, Appian, Celonis, Pega, ServiceNow, UiPath, IBM watsonx Orchestrate, and the BPMN modelers in which business owners and business analysts define what the agent is supposed to do. Business intent originates in those tools. If the governance profile is not rooted in the process design layer, business intent has no architectural path to flow into credential issuance, runtime enforcement, or the audit record. Governance is then forced rightward, into policy authored after the fact, downstream of the business intent it is supposed to express.

> *Insight Check:* Moving governance left, to the business owner and the place where the work is defined, requires the trust root to begin in the process design tool, not be assembled downstream from it.

The *millstone around the neck* of current architecture, when it comes to the needs of agentic AI:

- Traditional IAM does not do governance. It authorizes and enforces access.

- HashiCorp Vault does not do governance; it does credential management.
- Cisco-Astrix does not do governance; it does discovery.
- Strata Maverics does not do governance; it does identity orchestration.

None of the vendors analyzed in this paper has a business owner, a data owner, or a compliance officer as a first-class actor in its data model. The closest any of them comes is a policy authored by IT, after the agent already exists, in a configuration language only IT understands. That is not governance. That is post-deployment paperwork.

2. Visibility

Visibility is the question of whether you know what agents exist in your environment, where they came from, who owns them, and what they are doing. It includes discovery, which means finding shadow agents that were deployed outside the governed path, and it includes ongoing operational awareness of every agent's state, credential status, workflow context, and behavioral signature.

This layer has seen significant M&A this year, driven by the need to rein in shadow AI. Cisco announced intent to buy Astrix specifically because Astrix can find non-human identities across heterogeneous environments and feed that discovery into Cisco's broader identity intelligence. CrowdStrike agreed to acquire SGNL for behavioral signal. Emerging vendors such as Geordie and Reva specialize in real-time agent observability.

Visibility is becoming a well-served capability. But knowing what your agents are doing is not the same as being able to govern or stop them. On its own, visibility establishes no trust.

> *Insight Check:* Discovery without governance and enforcement is awareness without authority. It tells you an agent exists but does not tell you whether it should.

3. Transparency

Transparency addresses the questions of whether every governance decision and every operational action can be reconstructed and understood, not just by the system that made it, but by the humans who must answer for it. Why was this agent approved? Who approved it? Why was this threshold set at this level? Who set it? Why did this credential issue? Why did that one not? What changed between the approval and the deviation?

Transparency (lack thereof) is where most automation initiatives will quietly fail. Implementations that automate away the human review without preserving the reasoning behind it will generate audit findings that provide a false sense of security. For that reason, Sonnenfeld's framework places transparency at the top of the diagnostic matrix; without it, accountability is unenforceable and bias is undetectable. Logs are not transparency. Logs are records. Transparency is the connective tissue from the business intent behind the agent's design, through the stakeholder convergence that authorized it, through the policy that enforced it, through the behavior that occurred, to the audit record that proves it.

4. Control

Control is arguably the most misunderstood of the six. Many vendors are equating control of agentic AI with authorization, the ability to grant or deny access at the moment a credential is requested. That is half of control, and it is the easy half.

Complete control of agentic AI has four distinct requirements.

First, control must operate at runtime, not just at credential issuance. Once an agent holds a valid credential, it can deviate from the behavior its governance profile approved. It can attempt actions outside its declared scope. It can access resources its data owner did not sanction. It can be subverted by a prompt injection, a model drift, or a deliberate adversarial input. At that moment, the question is no longer whether the agent should have a credential. The credential is already issued. The question is whether the agent's deviation can be stopped mid-operation, before irreversible harm, with cryptographic certainty that the termination was legitimate. Authorization at issuance is a snapshot. Behavior at execution is continuous. Control requires both.

Second, control must terminate, not just observe. Knowing your agent is misbehaving is not the same as stopping it. Existing IAM does neither. Most of the early innovators do the first. The complete trust model requires both. On April 15, 2026, the President publicly endorsed the concept of a government-mandated AI kill switch in widely reported remarks. Moreover, in that same month, the Cloud Security Alliance published two studies documenting the scope of the agent-specific risk landscape, including findings that 53% of organizations had AI agents exceed their intended permissions and 65% experienced AI agent-related incidents in the prior year. The Cloud Security Alliance and SANS Institute jointly issued 'The AI Vulnerability Storm: Building a Mythos-Ready Security Program,' an industry response to the Anthropic Mythos Preview disclosure. In late April 2026, the PocketOS incident demonstrated why credential issuance gates are not sufficient: a coding agent with a valid credential and no human in the loop deleted a production database in a single API call, with founder Jer Crane publishing a detailed post-mortem.

Third, control must extend to ungoverned agents. The realistic threat model includes employees who circumvent governance by deploying agents outside the governed path. Even with cross-functional tooling and a comprehensive process model, some employees will create shadow agents. The control layer must therefore detect ungoverned agents in production through endpoint detection, cloud security posture management, or kernel-level runtime observability, and terminate them on the basis that an agent without a governance profile is, by architectural definition, unauthorized.

Fourth, control must include a credential-bound human-in-the-loop capability that is fundamentally different from heritage workforce multi-factor authentication. Heritage MFA verifies a human at the moment of a session login and then walks away. The agentic AI control layer needs cryptographic attestation that binds a verified human, a verified device, a verified location, a verified time window, and a verified request source to a specific credential issuance event, in seconds. Then it must associate the resulting agent action back to that human-of-record through the credential's entire lifecycle.

5. Automation

Automation is the question of whether governance can keep pace with agent deployment velocity. IBM survey data presented at Think 2026 projects that most large enterprises will have deployed digital workforces averaging sixteen hundred AI agents by the end of 2026. Gartner predicts that by 2028, an average global Fortune 500 enterprise will have over 150,000 agents in use, up from less than 15 in 2025. In that same survey, only 13% of organizations think they have the right AI agent governance in place. A key reason: quarterly attestation cycles and manual provisioning workflows are not just inadequate. These incumbent processes will never keep up with the deployment curve.

Conversely, successful automation in governing agentic AI first requires process change — the codification of human judgment into rules that execute at machine speed. The CISO does not review every agent. The CISO publishes the threshold tables that define when review is required, and the platform enforces those thresholds in real time. Business owners get autonomy within bounds; CISOs get oversight without bottleneck; auditors get a complete trail of how delegation evolved. This process model must be transparent, bidirectional with the business process mining and design tools (where the work is modeled) and facilitate overrides.

The legendary head of manufacturing at Harley-Davidson, Tom Gelb, once told me, “Technology amplifies processes — both good and bad. Automate a broken process, and you’ve simply made it fail faster.” Advice that transcends time.

6. Auditability

Auditability is the question of whether you can prove all of the above to a regulator, examiner, board, or court. It is the proof layer. Without it, governance is a claim. With it, governance is a verifiable fact.

Auditability for agentic AI cannot be retrofitted from log collection. Logs are mutable, scattered, and incomplete. The vendors that own evidence collection (e.g., OneTrust, ServiceNow GRC, Drata, Vanta) produce reports by gathering evidence from external systems and asserting that the assemblage represents the truth. Agentic AI requires an inverse model. Evidence must be generated as a byproduct of doing the governance, not collected after the fact. Every governance decision, credential issuance, runtime enforcement action, quarantine, and clearance is cryptographically hashed at the moment it occurs and chained to its predecessor in a tamper-evident hash chain (SHA-256 over canonical event data, Hyperledger-ready but not blockchain-dependent). Any alteration breaks the chain and is detectable. The compliance report is then a deterministic projection of an already-tamper-evident chain, not a manually assembled artifact whose integrity depends on the diligence of the team that assembled it.

This is the distinction that separates the six building blocks from a relabeling of governance, risk, and compliance tooling. GRC asks whether you can assemble proof, after the fact, that controls existed. Agentic Auditability asks whether the control executed at the moment the agent acted, and whether the proof generated itself. GRC is detective and assembled. Agentic Auditability is preventive and intrinsic. Prepending “agent” to a GRC platform does not produce the second; the second comes only from originating governance at the point of business intent and generating the evidence as the governance executes.

Regulators are moving in this direction. DORA Article 21 explicitly contemplates technical verification standards for autonomous systems. The EU AI Act’s Article 12 record-keeping obligations are scheduled to take effect on August 2, 2026, with evidentiary requirements that go well beyond “do you have policies.” NIST is developing an AI Agent Security Overlay. The Sonnenfeld framework places regulatory prescription as one of the four post-deployment governance variables. Auditability, done correctly, lets an enterprise satisfy every emerging regulatory regime with the same architectural investment, rather than rebuilding compliance for each new framework.

The right starting point is a foundation built on a recognized control catalog (NIST 800-53 Rev 5 is the natural anchor for federal, financial services, and healthcare), with a framework-agnostic auto-classification engine that extends to HIPAA, PCI DSS, EU AI Act, NIST AI RMF, and SOC 2 Type II as those regulatory overlays mature.

IV. What Recent Vendor Announcements Tell Us, and What They Do Not

None of these AI-related announcements delivers all six building blocks. Each addresses a slice.

IBM's HashiCorp Vault, with the native agentic capabilities shipped in mid-May, delivers Control at credential issuance and partial Auditability. They do not deliver business-intent Governance, multi-stakeholder Transparency, runtime behavioral Control, or the Automation of governance lifecycle. Vault remains a secrets platform retrofitting agent governance capabilities onto a model designed for persistent credentials. Its role in delivering ephemeral resource credentials to legacy systems is complementary, not competitive, with the governance layer.

IBM Verify, Vault 2.0, and Sovereign Core, unveiled at Think 2026, deliver strong Visibility, Control at issuance, Auditability, and Automation of identity orchestration. They do not deliver multi-stakeholder Governance origination. In my understanding, IBM's framing currently keeps governance authority in the hands of IT. The business owner, data owner, and compliance officer are not first-class actors in the Verify data model.

Cisco's planned acquisition of Astrix Security delivers excellent Visibility and partial Transparency of detected behavior. It does not deliver Governance origination, runtime Control with termination authority, or Auditability of business intent. Astrix retrofits governance onto agents that were already deployed; it does not govern the origination of new agents.

Strata Maverics, including the AI Identity Gateway shipped in late 2025, delivers runtime authentication and authorization for agents through delegated OAuth and SPIFFE/SVID, policy-driven access enforcement with optional human-in-the-loop for sensitive actions, and full-stack observability through OpenTelemetry. It does not deliver Governance origination, business process integration, multi-stakeholder Transparency, or tamper-evident Auditability of the governance lifecycle. Strata enforces policy at runtime; it does not originate or govern the policy it enforces.

CrowdStrike's SGNL (announced) acquisition delivers Control at the grant-revoke level and partial Visibility through behavioral signal. It does not deliver CISO review workflow, biometric human attestation, business process context, or full Auditability of the governance lifecycle.

Geordie, Reva, and the broader agent behavioral observability tier deliver Visibility and Transparency of behavior. They do not deliver Control with termination authority tied to a

governance baseline, or Auditability of business intent. Observation without enforcement is the layer that lets you watch the deviation. The control to stop it must come from elsewhere.

IBM has built a comprehensive portfolio and warrants particular attention:

- HashiCorp Vault's native agentic features for Control at credential issuance
- Verify and Vault 2.0 for Visibility and identity orchestration
- Sovereign Core for runtime policy enforcement and regulated-industry sovereign deployment
- Concert for operational observability
- Bob (IBM's AI coding agent) as agentic development partner
- watsonx Orchestrate as multi-agent control plane
- The Z mainframe and Spyre accelerator data center inference layer

For what IBM has accomplished very quickly, substantive architectural gaps remain.

- None of the announced products has a single living governance profile object on which the business owner, the CISO, the data owner, the compliance officer, and the operations team converge in real time, with full attribution and bidirectional propagation.
- None is wired bidirectionally into the business process design and mining tools where the agent's intent actually originates — the same gap that defines every other vendor in this wave.
- The portfolio is integrated by workflow and API, not by a shared trust root. Each product carries its own data model and policy language. Credential issuance, runtime enforcement, and audit each carry the imprint of separate product architectures. The customer remains the integrator, and the deployment velocity that agentic AI demands does not tolerate quarter-length integration projects.
- The agent owner is not the inherent Trust Owner.

In summary, these vendors are correctly addressing real problems in the current wave of innovation and M&A. Every one of them is solving various parts of this six-part trust model. None of them is delivering the whole. Addressing this required integration is also where the moat lives.

The strategic implications of this gap are the subject of the next section.

V. The Strategic Choice You Are Already Making

Every enterprise deploying agentic AI is already making a choice about how it will govern those agents. The question is whether the choice is being made explicitly, by the right people, with full awareness of the trade-offs. Since late 2025, I have observed many enterprises are assembling guardrails without fully understanding what they are committing to.

Five choices are available.

Option A — Deploy and Accept Risk. Deploy agents using extended human IAM, with any incremental governance tooling the IT team can configure on top. Accept that you cannot prove to a regulator which specific agent made a given decision. Accept that you cannot terminate a deviating agent mid-operation with cryptographic certainty. Accept that you cannot produce tamper-evident compliance evidence, only after-the-fact log archaeology. Hope that aggregate vendor solutions arrive before consequences do. This is the default posture for organizations not making an explicit choice.

Option B — Wait for Vendor Convergence. Defer significant agentic AI deployment until a major vendor ships an integrated platform covering all six building blocks. The vendors that shipped in May 2026 are not yet there. Industry analysts project integrated platforms in 2027 to 2028. The cost of waiting is two to three years of competitive disadvantage during what is, in my assessment, the most consequential platform transition since the arrival of the internet. For many large enterprises, the risks of waiting are deemed greater than the risks of proceeding with insufficient guardrails. Each entity must make this risk management tradeoff.

Option C — Evaluate the Convergence Architecture Now in Controlled Pilots. Evaluate business-intent governance architecture, against your three highest-risk agent use cases, before committing to production deployment. This architecture entails an abstraction layer over your strategic incumbent vendors, integration with new entrants, and in-house capabilities you build yourself. Leverage Agentic AI to accelerate this buildout. Objectively assess the extensibility and limitations of current tools without compromising the architectural non-negotiables laid out in this paper. In the process, fast-moving, yet *introspective* enterprises have a unique opportunity to change the rules of the game to create sustained differentiation.

Option D — Do Everything in Your Power to Resist the Steamroller. Leverage legislative, industry and regulatory resources to delay, constrain, or prevent adoption of AI in your sector.

Option E — Live off the Grid

There is, in truth, a sixth choice, and it is the worst: to make no deliberate choice at all. “If you choose not to decide, you still have made a choice.”

VI. What Comes Next

A companion paper to this one, Part Two, describes the architecture that delivers all six building blocks as an integrated trust model, with the multi-stakeholder governance convergence model at its center. This paper, Part One, was written to be read by everyone with a stake in agentic AI, from the board and C-suite to the technical leaders who will execute. Part Two is for the implementers: the CISO, CTO, and leaders of architecture, development, and operations. They need to see how the six blocks fit together, how the governance origination layer drives downstream enforcement, and how the architecture coexists with the rest of the agentic AI ecosystem.

I submit these Six Building Blocks of Agentic AI Trust capture the structural reality of what every enterprise will be expected to deliver, by every emerging regulatory framework, in every audit, in every board oversight cycle, for the next decade. The organizations that build to them deliberately, starting now, will define the standard.

> *Choose explicitly. Build deliberately. Govern at the speed of the agents, not at the speed of the quarterly review cycle.*

About the Author

Casey Plunkett is Co-Founder and CEO of Secure AI LLC. At IBM, he served as Chief of Staff to the General Manager of Tivoli, then as Director of Global Sales for IBM Security, leading 1,300 specialists serving 15,000 customers across 160 countries. In that role, he integrated three segments into the company's first unified IAM suite and launched the Federated Identity Management product, growing it from zero to fifty million dollars in revenue in under a year. He also led due diligence and integration for six IBM acquisitions. At Oracle, as Senior Practice Director of North America Security Consulting, he created the IAM and Database Security Practice and led the Oracle Tech Surge that stabilized Healthcare.gov in 2013. He is the author of The Agentic AI Steamroller and has led more than six hundred global engagements in digital identity, privacy, and cybersecurity over two decades.

Casey Plunkett, Co-Founder and CEO, Secure AI LLC · June 2026

Part One of a Two-Part Position Paper.

Part Two: "The Convergence Architecture" available separately.