Identity-Aware AI Security

An Enterprise Workbook to Govern AI Through the Lens of Identity

Version 5.0.2 | April 2026 | startmakingsense.org

1 Introduction

1.1 How to Use This Workbook

This workbook is for leaders responsible for using AI in ways that touch core business systems and data. It is written for CISOs, CIOs, CDOs, Chief Risk and Compliance Officers, enterprise architects, and senior business leaders who own AI risk throughout the enterprise.

The focus is identity-aware AI security: designing AI systems so that every chatbot, copilot, and agent has clear, enforceable answers to three questions:

- What data and capabilities is this identity allowed to access?
- What is it allowed to do with that data — aggregate, correlate, anonymize, act?
- What level of detail is it allowed to surface, to whom, and in what context?

This workbook is for enterprises that recognize the need for identity-aware AI security and are willing either to architect it themselves or to impose the construct on their technology vendors and systems integrator partners.

The workbook is organized around five pillars:

- Pillar A: Identity-Aware Authorization Policy Management
- Pillar B: Identity-Aware Retrieval
- Pillar C: Identity-Aware Abstraction
- Pillar D: Post-AI Security Operations
- Pillar E: Enterprise AI Governance

For each pillar you will find: the core problem it addresses, implementation patterns, key enforcement contexts and the vendors that operate there, and workbook prompts to guide assessment and planning. Two appendices follow the main pillar content:

- Appendix A — Inter-Pillar Interfaces: codifies the interfaces between pillar functions, the direction and nature of what flows across each, and the implementation variants that enterprises typically use.
- Appendix B — Key Vendor Solution Pairs Across Inter-Pillar Interfaces: maps known vendor integrations to each interface, with notes on integration depth and interoperability considerations.

The identity-aware AI security architecture is "graphed" in Figure 1 below. The graph attempts to visually simplify the orientation of the pillars relative to themselves and other key elements of the enterprise estate.

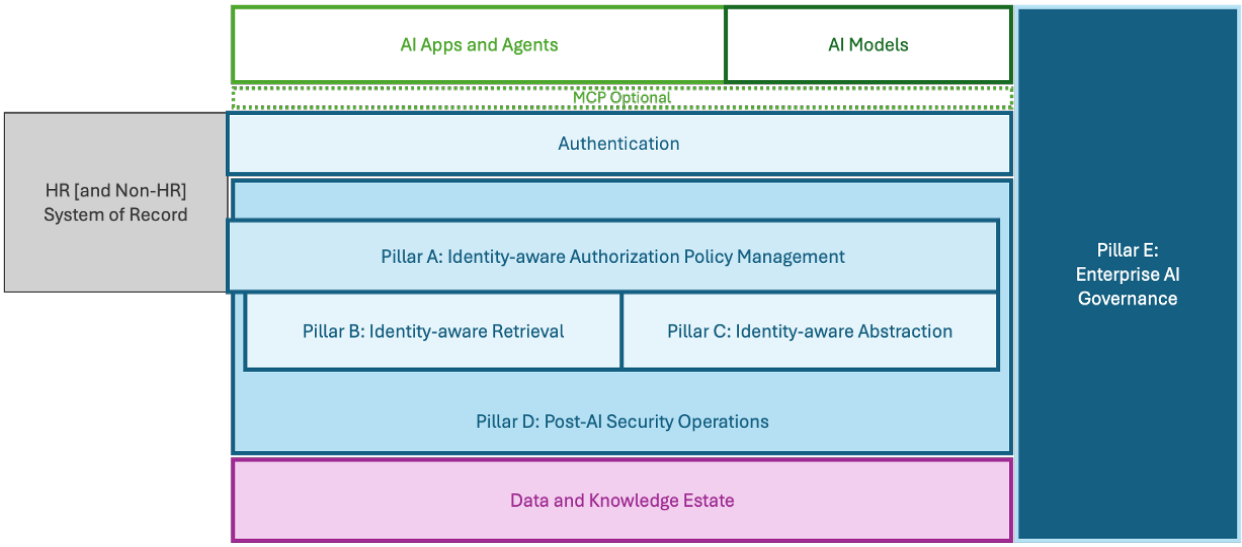

Figure 1: Identity-Aware AI Security Architecture — Five Pillars and Key Adjacent Systems

This workbook applies systems-level thinking to help enterprises iterate toward a more structured and more effective approach to AI security. The five pillars are conceptual categories — they are not a claim that every technology fits neatly into one and only one pillar. Most enterprise security and identity tools perform functions that span multiple pillars, and most enterprises will build coverage within each pillar from a portfolio of technologies rather than a single system of record. Within Pillar A alone, enterprises often rely on different tools for on-premises and cloud environments, for human and agent identity management, and for policy definition versus policy enforcement. The value of the pillar structure is the conceptual clarity it provides for architectural decision-making and gap analysis — not a vendor taxonomy that claims to be mutually exclusive or exhaustive.

This workbook is one in a series exploring how enterprises can align AI-enabled strategy, operations, and transformation. Further workbooks and practical guides are available at enttao.com.

1.2 Strategic Principle Hypotheses Served by This Workbook

This workbook supports three strategic principle hypotheses published at startmakingsense.org. Strategic principle hypotheses are explicit, testable recommendations about how some part of post-AI society should work, expressed with the best available logic and open to constructive critical dialogue.

Hypothesis	One-sentence summary
Identity-Aware AI Security	Enterprises should adopt identity-aware AI security as the backbone for how AI is securely deployed across their estate, implemented via proactive, real-time enforcement policies that govern what each human and AI agent identity may read, transform, and reveal.
Post-AI Security Operations	Enterprises should deliberately evolve — and where necessary redesign — their security operations to account for AI as both a new source of risk and a new security capability, adapting DLP, logging, SIEM/SOAR, and SOC workflows accordingly.
Enterprise AI Governance	Enterprises should establish Enterprise AI Governance as a supervisory oversight function — typically a cross-functional process or board — explicitly embedded in a governance stack alongside data governance, security governance, and business transformation governance.

1.3 Why Identity Is the Right Control Surface for AI

In the pre-AI enterprise, most access mistakes were local. A mis-scoped role on one application might leak one dataset or one workflow — painful, sometimes costly, but bounded.

AI changes the geometry of risk. Once you let copilots, chatbots, RAG systems, and agents read across mailboxes, document stores, SaaS applications, line-of-business systems, and observability platforms, you have effectively created a new access and insight layer on top of your estate. This layer can read faster and more widely than any human, connect previously disparate silos, and surface synthesized answers in natural language.

That is the point — and the risk. Any mis-scoped permission that would previously expose one dataset now potentially exposes any synthesized insight that dataset could contribute to. And if AI agents can act — invoke APIs, trigger workflows, write records — then a mis-scoped agent identity is a mis-scoped actor, not just a mis-scoped reader.

Traditional security controls — DLP rules, network inspection, anomaly detection — were designed for a world of human-generated, bounded transactions. They are valuable safety nets but they are inherently reactive and approximate: they catch patterns, not intent, and they accumulate false positives and false negatives at scale. Identity-aware controls are fundamentally more precise: they enforce what a specific, authenticated identity is permitted to do before the action occurs, eliminating entire categories of risk rather than detecting violations after the fact.

If you cannot point to a specific place in the architecture where identity-aware policy is enforced, you have likely introduced an AI access path that is less governed than your pre-AI systems.

1.4 Architecture at a Glance: The Five Pillars

The five pillars form a layered architecture grounded in a shared policy authority. The key design principle is that identity-aware controls in Pillars B and C are more precise and more preventive than the pattern-based safety nets in Pillar D — and that the precision of Pillars B and C depends entirely on the quality and currency of the shared policy authority in Pillar A.

Pillar A is the shared policy authority: it is not a single monolithic system but a capability — potentially composed of multiple technologies covering different environments and identity types — that provides a consistent, authoritative, and callable source of policy truth for all other pillars. Pillars B and C consume Pillar A policy at their respective enforcement points; they do not maintain their own independent policy definitions. Pillar D monitors for what the upstream architecture missed and feeds findings back into Pillar A for correction. Pillar E provides the human oversight and strategic direction that gives the entire architecture institutional legitimacy and ensures it remains aligned with enterprise values, obligations, and risk appetite.

Pillar	Role and Key Distinction
Pillar A: Identity-Aware Authorization Policy Management	The shared policy authority defines and maintains the authoritative set of identity-aware policies — who may access what, under what conditions, and with what permitted operations — across all AI-touching systems. Pillar A is a shared capability, not necessarily a single system; it serves multiple enforcement contexts in Pillars B, C, and beyond.
Pillar B: Identity-Aware Retrieval	Applies Pillar A policy at the point where AI reads data. Governs what data each identity may retrieve from knowledge stores, data platforms, and document repositories. The enforcement mechanism calls Pillar A for the authoritative policy decision at retrieval time.
Pillar C: Identity-Aware Abstraction	Applies Pillar A policy at the point where AI formulates its output. Governs what level of insight and detail may be revealed to each identity — independent of what data was retrieved. The enforcement mechanism calls Pillar A for the recipient identity's authorized disclosure depth.

Pillar	Role and Key Distinction
Pillar D: Post-AI Security Operations	The safety net layer monitors AI inputs (prompts) and outputs for policy violations, anomalies, and data exposure that the upstream identity-aware architecture missed. Feeds findings back to Pillar A as entitlement and policy corrections. Encompasses detection tools, DLP, DSPM, and SOC workflows. Distinguished from Pillars A-C by being compensating and detective rather than preventive and precise.
Pillar E: Enterprise AI Governance	The human oversight layer is the cross-functional function that translates leadership intent and risk appetite into Pillar A policy, ensures the governance architecture remains aligned with regulatory obligations and enterprise values, and provides board-level accountability for AI risk.

1.5 Key Architectural Concepts

Concept	Definition
Human identity	A specific, authenticated individual — an employee, contractor, or partner — with a defined set of entitlements reflecting their organizational role, responsibilities, and need-to-know. Human identities are provisioned and governed through IGA processes and are subject to access certification by their manager or role owner.
Personal agent identity	A non-human identity representing an AI agent that acts on behalf of a specific individual human — for example, a personal AI assistant or copilot operating in the context of one user's work. Personal agent identities should be granted entitlements derived from, but not exceeding, the sponsoring human's own entitlements. Access reviews for personal agents are typically performed by the manager of the sponsoring human, analogously to other entitlements granted to that person.

Concept	Definition
Enterprise agent identity	A non-human identity representing an AI agent that acts on behalf of an organizational function or process — for example, an enterprise awareness agent, an automated workflow agent, or a background intelligence service. Enterprise agent identities are scoped to their defined function and governed by a designated team or governance body that owns the risk of the agent's intended purpose. They are not derived from any individual human's entitlements.
Entitlement	A specific right granted to an identity: to read, write, invoke, or see data at a defined level of detail, under defined conditions.
Shared policy authority (Pillar A)	The capability — potentially composed of multiple technologies — that provides a consistent, authoritative, and callable source of identity-aware policy truth for all other pillars. May include IGA platforms, policy engines, identity providers, and CIEM tools, each covering different environments or identity types.
Identity-aware retrieval	A pattern where every data retrieval call is evaluated against the calling identity's entitlements as defined by the Pillar A shared policy authority before data is returned.
Identity-aware abstraction	A pattern where an AI agent or system shapes the detail and disclosure of its output based on the recipient identity's authorized disclosure depth, as defined by the Pillar A shared policy authority.
Read / Transform / Reveal	The three operations any AI performs on data, each independently governable: Read (what data it may access), Transform (what operations it may perform — aggregate, correlate, anonymize, predict), Reveal (what detail level it may surface to which recipient identity).
Clearance tier	A classification of a recipient identity's authorized disclosure depth: e.g., operational detail, functional summary, or enterprise aggregate. Governed by Pillar A; enforced by Pillar C.
DSPM	Data Security Posture Management: continuous discovery, classification, and monitoring of data across environments, together with assessment of the corresponding security posture — the degree to which data is appropriately protected, correctly classified, and accessible only to entitled identities. DSPM findings are a key input to Pillar A policy and Pillar B scope decisions.

Concept	Definition
Policy engine	A system (e.g., OPA, Cerbos, Cedar) that evaluates identity context and policy rules at runtime to produce an authorization decision. Policy engines are a key component of the Pillar A shared policy authority and are called by Pillar B and C enforcement contexts.
IGA (Identity Governance and Administration)	A platform managing the lifecycle of identities and entitlements: provisioning, access certification, role management, and joiner/mover/leaver workflows. A core component of the Pillar A shared policy authority.
SPIFFE / SPIRE	Secure Production Identity Framework for Everyone (SPIFFE) is a CNCF standard that assigns cryptographically verifiable identities — called SVIDs (SPIFFE Verifiable Identity Documents), typically X.509 certificates — to software workloads including AI agents. SPIRE is the production-grade runtime that issues and rotates SVIDs automatically. SPIFFE/SPIRE enables mutual TLS (mTLS) between services without long-lived secrets, providing the workload identity infrastructure that underpins Pillar A's non-human identity governance for AI agents in distributed and multi-cloud environments. Referenced by NIST NCCoE as a foundational standard for AI agent identity.
MCP (Model Context Protocol)	An emerging open standard for how AI applications and agents connect to external tools, data sources, and services. MCP defines a client-server protocol in which AI clients (apps and agents) invoke capabilities exposed by MCP servers. Enterprise-Managed Authorization — an official MCP extension — routes MCP server access through the enterprise identity provider (IdP), enabling centralized policy enforcement, SSO-anchored identity, and auditable authorization for AI agent tool access. MCP is shown as an optional integration layer in the architecture diagram; where present, it becomes a Pillar B enforcement context.
OAuth 2.0 Token Exchange (RFC 8693)	The IETF standard protocol that enables a service to exchange one security token for another with a different scope, audience, or identity context. In AI agent scenarios, RFC 8693 is used to propagate a human user's identity through a chain of agent calls — the resulting token carries both a subject claim (the originating human) and an actor claim (the delegating agent), creating a complete, auditable delegation chain. This is the formal standard underlying the OBO (On-Behalf-Of) pattern described in Pillar B Pattern B3.

Concept	Definition
SCIM (System for Cross-domain Identity Management)	The IETF standard protocol for provisioning and deprovisioning identities across enterprise systems. SCIM 2.0 is the wire protocol that enables IGA platforms to push identity lifecycle events (create, update, suspend, delete) to connected applications automatically. A new IETF draft extends SCIM to AI agents and agentic applications — introducing Agent and AgenticApplication resource types — enabling the same automated provisioning and lifecycle governance already used for human identities to be applied to AI agents.
NIST AI RMF	The NIST AI Risk Management Framework (AI 100-1, released January 2023) provides voluntary guidance for managing AI risks across the AI lifecycle. Its four core functions — Govern, Map, Measure, Manage — align closely with the role of Pillar E Enterprise AI Governance. The AI RMF's Govern function corresponds to E-AIG policy and oversight; Map and Measure correspond to AI use case triage and risk tiering; Manage corresponds to the lifecycle governance and feedback loops across Pillars A-D.
NIST SP 800-207 (Zero Trust Architecture)	The NIST Special Publication defining Zero Trust Architecture (ZTA) principles: verify explicitly at every access request, use least-privilege access, assume breach. The five-pillar architecture described in this workbook is an implementation of ZTA principles applied specifically to AI systems — with Pillar A serving as the policy engine, Pillars B and C as policy enforcement points, and Pillar D as the continuous monitoring and telemetry function that ZTA requires.

2 Pillar A: Identity-Aware Authorization Policy Management

The shared policy authority: defining, maintaining, and making callable the identity-aware policies that govern all AI activity.

2.1 The Problem

Every identity-aware control in Pillars B and C depends on one thing being true: there is an authoritative, consistently defined set of policies that specifies who — human or AI agent — may access what data, invoke what actions, and reveal what level of insight, under what conditions. Without a shared policy authority

in Pillar A, Pillars B and C are forced to define and manage policy locally — producing inconsistent, hard-to-audit, and unscalable governance.

Pillar A's role is not to perform all enforcement itself, but to be the canonical source of policy truth that enforcement points in Pillars B, C, and beyond can call. It is a shared capability rather than a single monolithic system: most enterprises will compose Pillar A from multiple technologies covering different environments (on-premises versus cloud), different identity types (human, personal agent, enterprise agent), and different policy domains (data access, action authorization, tool access, conditional access).

2.2 Core Principle

Pillar A provides a shared, authoritative, callable policy authority. Any system that needs to make an identity-aware authorization decision — whether at data retrieval (Pillar B), output disclosure (Pillar C), AI-to-AI delegation, AI tool access, or MCP server authorization — should resolve that decision by calling Pillar A rather than maintaining its own local policy definition.

2.3 What Pillar A Should Manage

What Should Be Managed	Description
Human identity lifecycle	Joiner/mover/leaver events trigger entitlement changes that propagate to all AI-touching systems, not just traditional applications. The IGA platform is the authoritative provisioning source; SCIM is the standard provisioning protocol for propagating these changes to connected systems.
Personal agent identity lifecycle	Personal AI agents are provisioned with entitlements derived from (but not exceeding) the sponsoring human's own entitlements. Access reviews are performed by the sponsoring human's manager, consistent with how other entitlements for that person are reviewed.
Enterprise agent identity lifecycle	Enterprise AI agents are provisioned with entitlements scoped to their defined function. A designated governance team or board proxy owns the risk of the agent's purpose and performs access reviews on a defined cadence.
Policy definitions and versioning	Authorization policies — for retrieval scope, disclosure depth, AI tool access, conditional access, and MCP server access — are managed as versioned, testable artifacts in a policy engine or IGA policy store, not hardcoded in application logic.

What Should Be Managed	Description
AI platform and tool access entitlements	Which identities are permitted to use which AI platforms, models, tools, and MCP servers is an entitlement question governed by Pillar A. The policy defines which AI tools are permitted, for which identity types, and under what conditions (e.g., managed device, specific data classification only).
Access certification and review	Human, personal agent, and enterprise agent entitlements are reviewed on a defined schedule. Uncertified access is automatically revoked. AI-specific entitlements are reviewed with the same rigor as traditional application access.
Secrets and credential governance	AI agents and services use short-lived credentials managed by a secrets vault. No long-lived shared secrets. SPIFFE/SPIRE provides cryptographic workload identity (SVIDs) for AI agents in distributed environments, replacing static credentials with automatically rotated X.509 certificates.
Cloud infrastructure entitlements (CIEM)	Cross-cloud visibility into what permissions AI workloads and service accounts actually hold — including permissions granted outside the IGA process. CIEM flags excessive permissions and configuration drift.
Authorization enforcement coordination	Pillar A serves as the shared policy authority for multiple enforcement contexts: identity-aware retrieval (Pillar B), identity-aware abstraction (Pillar C), AI-to-AI delegation, AI platform and tool access, MCP server authorization, and human-AI conditional access. For each context, the enforcement point calls Pillar A at runtime rather than maintaining independent policy.

2.4 Pillar A Enforcement Contexts

The following table describes the enforcement contexts that call on Pillar A as their shared policy authority. These are the points in the architecture where an identity-aware authorization decision must be made — and where the quality of Pillar A directly determines the precision of the control.

Enforcement Context	Where Enforcement Occurs	Representative Vendors / Standards
Identity-aware retrieval (Pillar B)	RAG pipeline pre-retrieval filter; vector store query; document repository access; graph database traversal	OPA, Cerbos, Cedar; SailPoint ISC, Entra ID Governance; Azure AI Search; Pinecone, Weaviate
Identity-aware output abstraction (Pillar C)	AI model output filter; structured output schema; enterprise agent output formatter	Guardrails AI, NeMo Guardrails; OPA, Cerbos; Microsoft Purview sensitivity labels
AI-to-AI delegation	Agent orchestration layer evaluates whether one agent may delegate to another and what scope may be passed; SPIFFE SVIDs authenticate agents to each other via mTLS	LangChain, AutoGen, CrewAI (with policy engine integration); SPIFFE/SPIRE for agent workload identity; emerging RFC 8693 token exchange patterns
AI platform and tool access	Identity provider (IdP) or SSO layer; zero-trust access policy; enterprise browser policy enforcement	Okta, Entra ID; Zscaler ZPA (as access policy enforcement point, calling Pillar A policy); Island Enterprise Browser
MCP server authorization	Enterprise-Managed Authorization extension routes MCP client access through enterprise IdP; IdP grants or denies MCP server access based on Pillar A policy; OAuth 2.1 tokens scoped to specific MCP methods	Okta (XAA/MCP extension); Entra ID; MCP Enterprise-Managed Authorization spec (modelcontextprotocol.io)
Human-AI conditional access	IdP / SSO enforces device state, location, step-up authentication conditions defined in Pillar A policy before permitting AI capability access	Entra ID Conditional Access, Okta Adaptive MFA; Zscaler ZPA (as enforcement point)

2.5 Implementation Patterns

2.5.1 Pattern A1: IGA as the Authoritative Entitlement Source

Your IGA platform should be the authoritative provisioning source for all AI system access — not just traditional applications. AI system access flows through IGA; entitlement changes propagate from IGA to AI systems via SCIM 2.0. Human, personal agent, and enterprise agent identities are all managed as first-

class IGA objects with their own lifecycle policies, certification schedules, and owner assignments. A new IETF draft extends the SCIM schema with Agent and AgenticApplication resource types, enabling automated provisioning and deprovisioning of AI agents through the same IGA workflows used for human users.

2.5.2 Pattern A2: Policy-as-Code

Authorization policies for AI — retrieval scope (Pillar B), disclosure rules (Pillar C), tool access permissions, MCP server access — are managed as versioned, testable code artifacts in a policy engine. Policies go through review and approval before production promotion. Policy tests run in CI/CD. Audit logs record which policy version was evaluated for each authorization decision.

2.5.3 Pattern A3: Differentiated Agent Identity Governance

Personal agent identities are provisioned with entitlements that are a constrained subset of the sponsoring human's own access — the agent should never be able to access data or capabilities that the human could not access directly. Access reviews for personal agents are performed by the sponsoring human's manager as part of the normal entitlement certification process for that person.

Enterprise agent identities are governed separately. Each enterprise agent has a named owning team or governance proxy that is accountable for the agent's defined purpose and that performs access reviews. Enterprise agent entitlements are scoped to the specific function the agent performs — they are not derived from any individual human identity and should not accumulate entitlements beyond what is needed for the agent's defined task. Credentials are short-lived, and the agent is deprovisionable when its function is retired.

2.5.4 Pattern A4: Zero-Trust Identity and Application Layer Authorization

Every AI service call is authenticated and authorized at the identity and application layers, not just the network layer. This means that service-to-service calls carry verifiable identity claims (JWT tokens, mTLS certificates with identity metadata) and that authorization decisions are made at the application boundary — consulting the Pillar A shared policy authority — rather than relying solely on network-layer controls such as firewall rules or IP allowlists. This pattern aligns with the zero trust architecture principles of NIST SP 800-207 and is particularly important for agentic AI that makes lateral calls across internal services.

2.5.5 Pattern A5: SPIFFE/SPIRE for AI Agent Workload Identity

SPIFFE (Secure Production Identity Framework for Everyone) provides a CNCF-standard mechanism for issuing cryptographic identities — SVIDs —

to AI agents as software workloads. A SPIRE server acts as the trust anchor: it attests agent workloads (via platform-provided signals such as Kubernetes service accounts) and issues short-lived X.509 SVIDs that the agent uses to authenticate via mTLS to other services and agents. This eliminates long-lived secrets for AI agents in distributed and multi-cloud environments — a particularly important control for enterprise agent identities that span multiple infrastructure boundaries. HashiCorp Vault Enterprise (v1.21+) now natively issues SVIDs to authenticated NHI workloads, integrating SPIFFE into existing secrets infrastructure.

SPIFFE SVIDs encode the agent's identity in the certificate SAN (e.g., spiffe://enterprise.com/ns/finance/sa/budget-agent), making agent identity verifiable and traceable at every service boundary without a shared secret. Combined with policy engines (OPA, Cedar), the SPIFFE ID becomes the identity context for runtime authorization decisions — enabling consistent Pillar A policy enforcement even for ephemeral, dynamically deployed AI agent workloads.

2.6 Key Vendors and Solutions

Solution / Vendor	Role in Pillar A	Notes
SailPoint Identity Security Cloud Enterprise	IGA for human, personal agent, and enterprise agent identity lifecycle; access certification; AI entitlement governance; API for runtime entitlement queries; SCIM-based provisioning to AI systems.	Industry leader for large, heterogeneous estates; strong non-human identity capabilities; on-premises and cloud coverage.
Microsoft Entra ID Governance	IGA for Microsoft-centric environments; native integration with M365 Copilot and Azure AI; access reviews; entitlement management; lifecycle workflows.	Recommended for Microsoft-first estates; deep integration with Entra Conditional Access for human-AI access policy.
Saviynt	Cloud-native IGA and CIEM; strong non-human identity and PAM capabilities; cloud infrastructure entitlement visibility.	Strong for cloud-native AI workloads; CIEM module adds cross-cloud entitlement governance layer.

Solution / Vendor	Role in Pillar A	Notes
OPA (Open Policy Agent)	Policy-as-code engine for runtime authorization decisions across AI pipelines, microservices, and API gateways; a core component of the Pillar A shared policy authority.	De facto standard for policy-as-code; integrates with Kubernetes, API gateways, and custom AI systems.
Cerbos	Lightweight, Git-native policy engine for application authorization; called at runtime by Pillar B and C enforcement contexts.	Low-latency; well-suited for RAG pipelines and AI orchestrators needing fast policy decisions.
Cedar (AWS)	Policy language and engine for fine-grained authorization; Verified Permissions managed service on AWS.	Strong for AWS-native AI workloads; formally verified policy evaluation.
HashiCorp Vault / Azure Key Vault	Secrets management for AI agent credentials; short-lived tokens and automatic rotation; Vault Enterprise v1.21+ natively issues SPIFFE SVIDs to NHI workloads.	Essential for non-human identity credential governance; SPIFFE integration extends coverage to distributed AI agent workloads.
SPIFFE/SPIRE (CNCF)	Cryptographic workload identity for AI agents; SVID-based mutual TLS eliminates long-lived agent secrets; SPIRE server attests agent workloads across Kubernetes and multi-cloud environments.	Open standard; referenced by NIST NCCoE for AI agent identity; foundational for distributed agentic architectures.
Wiz / Orca / Tenable (CIEM)	Cloud infrastructure entitlement management — visibility into what AI workloads and service accounts actually hold across cloud environments.	Complements IGA for cloud-native AI deployments; flags drift between intended and actual permissions.
Okta / Entra ID (IdP)	Identity provider and SSO; issues verifiable identity tokens consumed by Pillar B and C enforcement contexts; enforces conditional access policies for AI tool and MCP server access; Okta Cross-App Access (XAA) extends enterprise identity governance to MCP servers.	Core identity infrastructure; the token issuer that other pillars depend on for verified identity claims.

2.7 Workbook Prompts

- A1. Is your IGA platform the authoritative provisioning source for AI system access, or are AI tools provisioned separately outside IGA workflows? What would it take to bring AI system provisioning under IGA?
- A2. Do you have a registry of all AI agent identities — both personal agents and enterprise agents? For each, can you identify: the owning team or person, the entitlements granted, the credential type and expiry, and the last certification date?
- A3. Are your AI authorization policies (retrieval scope, disclosure rules, tool access, MCP server access) documented, versioned, and testable? Or are they implicit in application code and configuration?
- A4. For each enforcement context listed in the Pillar A Enforcement Contexts table above — retrieval, abstraction, AI-to-AI delegation, tool access, MCP authorization, conditional access — does your architecture have a defined enforcement point that calls a shared policy authority? Which contexts have no enforcement point today?
- A5. For AI agents deployed across distributed or multi-cloud infrastructure: how are inter-agent and agent-to-service identities authenticated? Are long-lived secrets or static API keys still in use? What would it take to move to cryptographic workload identity (e.g., SPIFFE/SPIRE)?

3 Pillar B: Identity-Aware Retrieval

Applying Pillar A policy at the point where AI reads data from the enterprise knowledge and data estate.

3.1 The Problem

Most AI copilots and RAG systems are implemented with a shared service identity — a single, highly privileged account that retrieves documents on behalf of all users. The model can therefore read anything the service account can access, regardless of what the human user is allowed to see. The AI becomes the most privileged reader in the enterprise, often bypassing identity-specific controls already established for humans.

Identity-aware retrieval corrects this problem by inserting an explicit policy evaluation step — calling the Pillar A shared policy authority — before data is returned to the model. The model receives only data the calling identity is entitled to see.

A related problem arises where AI systems access enterprise knowledge through MCP servers: if those servers are not governed through the enterprise identity provider, they may be authorized with static API keys or personal access tokens, creating ungoverned, unauditable AI access paths that bypass the identity controls established for other enterprise systems. Astrix Security research

(2025) found that 53% of MCP server deployments relied on long-lived static credentials, with only 8.5% using OAuth.

3.2 Core Principle

For every AI capability that retrieves data, there must be an enforceable, auditable answer to: Which identity is making this request? What is that identity allowed to read? Only data within those entitlements may be returned. The answer comes from Pillar A, not from locally defined rules.

3.3 Pillar B Enforcement Contexts

The table below describes where Pillar B retrieval decisions are enforced and how those controls map to identity-aware policy inputs from Pillar A.

Enforcement Context	What Is Enforced (with Representative Vendors)
RAG pipeline pre-retrieval filter	Policy engine (OPA, Cerbos, Cedar) called with identity context before vector or keyword search; query scope limited to entitled corpora; entitlement source is IGA platform (SailPoint, Entra ID Governance).
Vector store query filter	Identity attributes applied as mandatory metadata filter before vector search returns results (Azure AI Search, Pinecone, Weaviate); metadata must reflect current IGA entitlement state.
API gateway / data mesh interface	Token-scoped entitlements evaluated at the API boundary before data domains are exposed to the AI layer; JWT claims from IdP (Okta, Entra ID) carry identity context; service mesh (Istio, Linkerd) enforces mTLS service identity; SPIFFE SVIDs authenticate agent workloads at service boundaries.
Document repository (SharePoint, Box, Confluence)	Native ACL or IGA-managed permission set evaluated at retrieval time; AI connector must propagate user identity token via OBO pattern — not substitute a shared service account (Microsoft Graph OBO; SharePoint Online permissions governed by Entra ID Governance).
Semantic layer / BI query engine	Row-level and column-level security evaluated against calling identity before aggregated data is exposed to AI (Databricks Unity Catalog, Snowflake RBAC, Looker user attributes).
Graph database traversal	Relationship-based access control (ReBAC) evaluated against calling identity before traversal returns nodes or edges (Zanzibar-style engines, OPA with graph context).

Enforcement Context	What Is Enforced (with Representative Vendors)
MCP server (tool and resource access)	Enterprise-Managed Authorization extension enforces IdP-managed access policy before AI client accesses MCP server tools or resources; OAuth 2.1 token scoped to permitted MCP methods; user authenticated via corporate SSO — not static API key (MCP spec Enterprise-Managed Authorization; Okta XAA; Entra ID).

3.4 Implementation Patterns

3.4.1 Pattern B1: Authorization-First RAG

Policy evaluation occurs before retrieval. The query is scoped to the corpora and document segments the calling identity is entitled to see. The model receives only pre-filtered results. Propagate the user's identity token from the UX layer through the retrieval pipeline — never substitute a shared service account.

3.4.2 Pattern B2: Permission-Scoped Vector Search

Entitlement attributes are stored as metadata at index time and evaluated as mandatory filters at query time. Semantic similarity search never returns results the caller cannot see. Metadata must be updated when IGA entitlements change. IGA platforms (SailPoint, Entra ID Governance) provide the authoritative entitlement set as input to the filter at session establishment or via real-time API call.

3.4.3 Pattern B3: Token Propagation via RFC 8693 (OBO)

The OAuth 2.0 Token Exchange standard (RFC 8693) enables a token to be exchanged at each service boundary such that the downstream data service sees the human's identity — not the AI service's identity. The resulting token carries both a subject claim (the originating human) and an actor claim (the delegating AI service), creating a complete, auditable delegation chain. This is the pattern Microsoft uses for Copilot in M365 (implemented as On-Behalf-Of, or OBO). For multi-hop agentic scenarios, nested RFC 8693 exchanges preserve the full delegation ancestry — enabling an audit log that shows "agent B called this API while acting on behalf of agent A, which was acting on behalf of user C."

3.4.4 Pattern B4: Minimum-Privilege Agent Retrieval Scope

Every AI agent that retrieves data is provisioned in Pillar A with a minimum-privilege retrieval scope: a precise definition of which data sources, corpora, classification tiers, and record types the agent may access. The retrieval filter enforces this scope at query time. Agent retrieval scope does not expand when

the agent is called on behalf of a human whose own entitlements are broader — the agent's own scope is the binding constraint.

3.4.5 Pattern B5: Zscaler as a Pillar A-Calling Enforcement Context

Zscaler's Zero Trust Exchange can function as an enforcement context in the Pillar B sense when it is configured to evaluate access policy fetched from the Pillar A shared policy authority — for example, determining whether a given identity is permitted to access a specific AI platform or internal AI service endpoint. In this mode, Zscaler is acting as a Pillar A-governed enforcement point, not as a detection or inspection tool. This is distinct from Zscaler's DLP capabilities, which belong in Pillar D.

3.4.6 Pattern B6: MCP Enterprise Authorization

Where AI applications and agents use the Model Context Protocol (MCP) to access enterprise tools and data, the MCP Enterprise-Managed Authorization extension replaces per-user static credential authorization with IdP-governed policy. The enterprise IdP (Okta, Entra ID) acts as the authorization authority: access to each MCP server is governed by Pillar A policy, enforced at the MCP authorization server. Employees authenticate via corporate SSO; OAuth 2.1 access tokens are scoped to permitted MCP methods. Revocation and lifecycle are managed centrally. This converts MCP from a potential shadow-IT access layer into a governed Pillar B enforcement context.

The Okta Cross-App Access (XAA) extension, now incorporated into the MCP specification as an authorization extension, enables enterprises to apply the same identity governance used for all other enterprise applications to AI agent tool access — making IdPs the control plane for AI agent connectivity.

3.5 Key Vendors and Solutions

Solution / Vendor	Role in Pillar B	Notes
Microsoft Entra ID Governance + M365 Copilot	Authoritative entitlement source for SharePoint, Teams, and Exchange; Copilot inherits Graph API permissions via OBO (RFC 8693); access reviews govern Copilot retrieval scope.	Deep native integration; access reviews are the primary governance mechanism for Copilot retrieval scope.

Solution / Vendor	Role in Pillar B	Notes
SailPoint Identity Security Cloud	Enterprise-wide entitlement management; runtime entitlement API for retrieval filter construction; non-human identity entitlements for agent retrieval scope.	Strong for heterogeneous estates; AI-specific entitlement policies; on-premises and cloud coverage.
Cerbos	Policy engine called as pre-retrieval authorization step in RAG pipelines; Git-versioned policy; low-latency API.	Ideal for LlamaIndex, LangChain RAG pipelines; clear separation of policy from application logic.
OPA (Open Policy Agent)	Policy engine for retrieval authorization decisions; called by API gateways, service meshes, and custom AI pipelines.	De facto standard; integrates with Kubernetes, API gateways, and custom systems.
Azure AI Search + Entra ID	Permission-scoped vector search using document-level ACL propagation; native M365 permission inheritance for SharePoint-indexed content.	Native Azure integration; entitlements governed by Entra ID Governance.
Pinecone / Weaviate	Vector stores with metadata filter support for permission-scoped search.	Require consistent metadata governance at ingest time; filter must reflect current IGA entitlement state.
Databricks Unity Catalog / Snowflake	Row-level and column-level security for AI agents querying governed data views; enterprise awareness agents see only abstraction-appropriate views.	Strong for data-platform-centric AI workloads; governed views prevent raw record access.
Zscaler ZPA (as enforcement context)	Zero-trust access policy enforcement for AI platform and service access, calling Pillar A policy — not inspection or detection.	Pillar B role is strictly as access enforcement point; DLP and inspection capabilities belong in Pillar D.
Okta (MCP/XAA)	Cross-App Access (XAA) extension, now incorporated in MCP spec; governs AI agent MCP server access through enterprise IdP; OAuth 2.1 token-scoped authorization replacing static API keys.	Converts MCP from ungoverned tool access to IdP-governed Pillar B enforcement; MCP Enterprise-Managed Authorization standard.

Solution / Vendor	Role in Pillar B	Notes
SPIFFE/SPIRE	Workload identity for AI agents at service mesh boundaries; SVID-authenticated mTLS ensures only attested agent workloads can reach retrieval endpoints.	Eliminates static credentials for service-to-service agent retrieval; consistent with NIST SP 800-207 zero trust principles.

3.6 Workbook Prompts

- B1. Inventory all AI capabilities that retrieve data. For each, record: What identity is used — the user's identity, a shared service account, or a governed agent identity?
- B2. For each AI capability, is there an explicit, auditable policy evaluation step — calling your Pillar A shared policy authority — before data is returned? If not, what would it take to introduce one?
- B3. Are your document stores, vector indexes, and data APIs consistently tagged with classification and ownership metadata that a policy engine can use as a filter? Which corpora are most urgent to classify?
- B4. Do any AI applications or agents in your environment use MCP to access enterprise tools or data? If so, how is MCP server access authorized — through the enterprise IdP with OAuth, or through static API keys or personal access tokens outside IGA governance?

4 Pillar C: Identity-Aware Abstraction

Applying Pillar A policy at the point where AI formulates its output.

4.1 The Problem

Identity-aware retrieval (Pillar B) governs what an AI model reads. But it does not fully govern what it can reveal. A model that retrieves only entitled data can still synthesize, summarize, and surface insights in ways that exceed the disclosure intent of the underlying entitlements.

This matters most for enterprise-wide AI: agents that synthesize insights across departments, management layers, or business units. Such agents can effectively create a cross-enterprise view — seeing patterns, anomalies, and performance signals that no individual would normally hold in one place. Even if each piece of retrieved data was individually entitled, the synthesized output may reveal more than any authorized disclosure policy intended.

Identity-aware abstraction adds a second governance layer at the output point: calling Pillar A to determine what level of insight and detail may be revealed to the recipient identity, independent of what was retrieved.

4.2 The Read / Transform / Reveal Framework

Operation	What It Means and Where Policy Applies
Read	What data the AI may retrieve — governed by Pillar B, calling Pillar A for the authorization decision.
Transform	What the AI may do with retrieved data: aggregate, correlate, anonymize, compare, score, predict — each transformation type can be independently governed as part of Pillar A policy.
Reveal	What the AI may surface in its output: raw records, summarized trends, aggregated metrics, or abstracted insights only — disclosure depth governed by the recipient identity's clearance tier, as defined in Pillar A and enforced at the output generation step.

4.3 Clearance Tiers for Enterprise AI

Tier	What the AI May Reveal at This Level
Tier 1: Operational detail	Specific records, transactions, or data points — subject to full retrieval-level entitlements (Pillar B).
Tier 2: Functional summary	Aggregated metrics, trends, and patterns for a defined scope (team, product, region) — suitable for managers and functional leads.
Tier 3: Enterprise aggregate	Cross-functional patterns, risk signals, performance comparisons — suitable for senior executives and board-level reporting.
Seal-break conditions	Defined circumstances under which a higher-tier identity may request operational detail: personal safety risk, confirmed fraud, regulatory investigation. Requires audit trail, justification, and approval workflow.

4.4 Pillar C Enforcement Contexts

The table below describes where Pillar C abstraction and disclosure controls are enforced for recipient identities.

Enforcement Context	What Is Enforced (with Representative Vendors)
AI model output filter / guardrail	Post-generation output reviewed against recipient identity's clearance tier; detail exceeding tier is abstracted or withheld (Guardrails AI, NeMo Guardrails, LangChain output parsers).
Enterprise insight agent identity	Agent operates under a distinct enterprise agent identity with its own read/transform/reveal entitlements in Pillar A; not inherited from any individual human (SailPoint ISC, Entra ID Governance — non-human identity management).
Structured output schema	AI output constrained to a governed schema encoding disclosure depth; detail fields populated only for entitled recipient identities (OPA, Cerbos evaluated at output formatting step).
Disclosure audit log	Every output involving synthesized or aggregated insight is logged with recipient identity, clearance tier, data sources referenced, and detail level surfaced.
Seal-break workflow	When a recipient requests detail beyond their normal tier, a structured workflow captures: identity, justification, approver, and specific data disclosed — then records in audit log.
Secure browser output controls	Enterprise browsers (e.g., Island) can enforce downstream use constraints on AI-generated content in the browser: what the recipient may copy, download, or share after delivery — a last-mile complement to server-side abstraction controls.

4.5 Implementation Patterns

4.5.1 Pattern C1: Read / Transform / Reveal Policy

Define explicit policy statements for each enterprise AI capability specifying: (1) what data it may retrieve (read — governed by Pillar B), (2) what operations it may perform on retrieved data (transform), and (3) what it may include in its output to a given recipient clearance tier (reveal). Encode these policies in the Pillar A shared policy authority alongside retrieval policies.

4.5.2 Pattern C2: Enterprise Awareness Agents

Enterprise awareness agents synthesize cross-functional insight — performance signals, risk patterns, strategic anomalies — from the full data estate. They operate under enterprise agent identities with defined read/transform/reveal entitlements in Pillar A. Their outputs are governed by recipient clearance tier: the same agent produces different disclosure depths for different audiences from

the same underlying analysis. Enterprise awareness agents do not hold raw data — they retain only the abstracted insight needed to serve their synthesis function.

4.5.3 Pattern C3: Data Abstraction Layer

For enterprise awareness agents operating over sensitive financial, HR, or operational data, a data abstraction layer mediates between the raw data estate and the agent's retrieval. The layer returns pre-aggregated views appropriate to the agent's entitlement tier. Treat abstraction latency as an engineering problem — pre-computed views, caching, materialized aggregates — not a reason to abandon identity-aware abstraction.

4.5.4 Pattern C4: Clearance-Tier-Aware Output Formatting

Structure AI outputs so detail levels are modular: an output schema contains fields for operational detail, functional summary, and enterprise aggregate. The Pillar A policy evaluation step populates only the fields the recipient's clearance tier permits. A single AI response pipeline serves multiple audiences without separate model calls.

4.6 Key Vendors and Solutions

Solution / Vendor	Role in Pillar C	Notes
Guardrails AI / NeMo Guardrails	Output-level guardrails that enforce disclosure rules on AI-generated text; configurable topic boundaries, output filtering, structured output constraints.	Purpose-built for LLM output governance; integrates with LangChain, LlamaIndex, custom pipelines.
LangChain / LlamaIndex	Orchestration frameworks where read/transform/reveal rules are implemented as pipeline steps; policy engine called at each stage.	Common in enterprise RAG; natural integration point for Pillar A policy calls at output formulation.
Microsoft Purview (sensitivity labels)	Data classification and sensitivity labels that feed into Pillar C disclosure rules; sensitivity labels can gate Copilot output disclosure depth.	Deep M365 integration; labels govern what Copilot surfaces in responses for labeled content.

Solution / Vendor	Role in Pillar C	Notes
SailPoint ISC (enterprise agent identity)	Lifecycle management for enterprise agent identities and their read/transform/reveal entitlements; access certification for enterprise awareness agents.	AI agent identity as a first-class governed object; certification by owning governance team.
Databricks Unity Catalog / Snowflake	Governed data views and row/column-level security for data abstraction; enterprise awareness agents query governed views, not raw tables.	Strong for data-platform-centric workloads; pre-computed abstraction views reduce latency.
OPA / Cerbos	Policy engine called at output formatting step to determine which response fields are populated for the calling identity's clearance tier.	Enables clearance-tier-aware output formatting with policy-as-code; consistent with Pillar A policy store.

4.7 Workbook Prompts

- C1. For each enterprise AI capability, have you defined explicit read / transform / reveal rules? Which AI can synthesize insights across the most sensitive data domains without a defined disclosure policy?
- C2. Do any AI systems produce outputs that aggregate or synthesize data across organizational boundaries? What governs the level of detail disclosed to which audience?
- C3. Have you defined clearance tiers for your enterprise AI outputs? Who owns the process for defining and maintaining these tiers, and where are they stored — in your Pillar A shared policy authority or locally in each AI system?
- C4. Once AI-generated content reaches a recipient, what controls exist on downstream use? Can a user copy a sensitive AI-generated report and share it externally?

5 Pillar D: Post-AI Security Operations

The safety net: monitoring AI inputs and outputs, detecting what the upstream architecture missed, and closing the governance feedback loop.

The strategic principle hypothesis on Post-AI Security Operations holds that enterprises should deliberately evolve — and where necessary redesign — their security operations to account for AI as both a new source of risk and a new security capability, adapting DLP, logging, SIEM/SOAR, and SOC workflows accordingly.

5.1 The Problem

Pillars A through C establish the identity-aware governance architecture. Their controls are preventive and precise — they stop unauthorized access before it occurs, by enforcing policy against specific identities. But no preventive architecture is complete. Policies have gaps. Configurations drift. New AI capabilities are deployed before governance catches up. Adversaries probe edge cases. Employees find workarounds.

Pillar D provides the compensating and detective controls that catch what Pillars A through C missed: it monitors AI inputs (prompts and code submitted by users and agents to AI systems) and outputs (model responses, retrieved context, agent actions) for policy violations, anomalies, and data exposure. It also adapts traditional security operations — DLP, SIEM, SOAR, SOC workflows — to handle AI-specific behavior.

Critically, Pillar D's findings feed back to Pillar A as the primary mechanism for improving the upstream governance architecture. A DLP violation at the output layer is not just an alert — it is a signal that a Pillar A policy or Pillar B retrieval scope may need tightening. This feedback loop is what keeps identity-aware AI security effective over time.

It is equally important to distinguish Pillar D's technical feedback loop from Pillar E's governance feedback: Pillar D routes technical findings — specific policy violations, entitlement anomalies, DLP events — to Pillar A owners for remediation. Pillar E routes broader risk and governance insights — patterns in AI incidents, regulatory developments, changes in risk appetite — back into the governance function that owns Pillar A policy direction.

5.2 DLP for AI: Inputs and Outputs

Traditional DLP was designed for data at rest and in motion — files, emails, network traffic. AI security requires extending DLP to two new surfaces. Together, input and output DLP create a closed inspection envelope around every AI interaction.

DLP Surface	What It Governs and Why
AI prompt/code inputs	Prompts and code submitted by users and agents to AI systems are inspected before reaching the model. Policy can block, redact, alert, or log based on: sensitive data patterns in the prompt (PII, financial data, credentials, IP); whether the submitting identity is permitted to send that data type to the target AI system; whether the target AI platform is on the approved list for that identity and data classification; and prompt patterns indicating potential abuse (data exfiltration attempts, prompt injection strings, adversarial instruction overrides). This applies to internal AI systems and external AI services alike.
AI prompt/code outputs	Responses, retrieved context, and generated content are inspected after generation. Policy can block, redact, watermark, or log based on: sensitive data patterns in the output; whether the output exceeds the recipient's authorized disclosure tier (Pillar C); and whether the output is being routed to an unapproved destination. Traditional DLP rules apply alongside Pillar C disclosure tier rules.

5.3 What Pillar D Must Provide

Capability	Description
AI prompt/code input monitoring and DLP	Every prompt and code submission to an AI system is logged and inspected; policy enforced on what data types and prompt patterns are permitted based on identity and target AI platform.
AI prompt/code output monitoring and DLP	Every AI response is logged and inspected; policy enforced on disclosure tier compliance, sensitive data patterns, and permitted output destinations.
AI activity logging	Every AI interaction generates a structured log event including identity, timestamp, data accessed, policy version applied, and result.
Behavioral anomaly detection	AI-specific detection rules in SIEM/SOAR identify unusual patterns: unusual retrieval scope, prompt injection attempts, agent action anomalies, unusual AI platform access.
DSPM	Continuous discovery, classification, and security posture assessment of data that AI systems can reach; flags unclassified or over-permissioned data in AI-accessible corpora.

Capability	Description
Red-teaming and AI-specific testing	Regular adversarial testing for prompt injection, data exfiltration via prompt, jailbreaks, and entitlement bypass.
Technical feedback loop (Pillar D to Pillar A)	Technical findings — DLP violations, anomaly detections, red-team results, DSPM discoveries — routed to Pillar A owners as specific entitlement corrections, policy updates, and retrieval scope adjustments. This is the technical remediation channel.
Governance feedback loop (Pillar E to Pillars A-D)	Broader risk and governance insights — AI incident patterns, regulatory changes, shifts in enterprise risk appetite — routed from Pillar E to the Pillar A policy owners and technical pillar leads. This is the strategic governance channel, distinct from the technical Pillar D feedback loop.

5.4 Pillar D Enforcement Contexts

The table below describes the key Pillar D monitoring and inspection contexts used to detect and route AI security findings.

Enforcement Context	What Is Enforced / Observed (with Representative Vendors)
AI prompt/code input DLP	Prompts inspected before reaching models; sensitive data patterns, identity, and target AI platform evaluated; block, redact, alert, or log (Zscaler CASB/ZIA DLP; Microsoft Purview DLP; Symantec DLP; dedicated AI input inspection: Lakera, Prompt Security).
AI prompt/code output DLP	Model outputs inspected after generation; sensitive data and disclosure tier checked; block, redact, watermark, or log (Microsoft Purview AI Hub; Zscaler DLP; Guardrails AI in detection mode; HiddenLayer).
SIEM / SOAR with AI-specific detection rules	AI activity log events analyzed for anomalous access patterns, prompt abuse, policy violations; automated response playbooks (Microsoft Sentinel, Splunk, IBM QRadar).
DSPM	Continuous visibility into data classification and access posture of AI-reachable data stores; proactive identification of over-permissioned or unclassified data (Securiti AI, Varonis, Microsoft Purview DSPM).

Enforcement Context	What Is Enforced / Observed (with Representative Vendors)
AI output monitoring / guardrails	Runtime monitoring of model outputs for hallucination, toxicity, policy violation, or disclosure exceeding authorized tiers (Guardrails AI, NeMo Guardrails, Lakera, Prompt Security).
Red-team / penetration testing (AI-specific)	Structured adversarial testing of AI systems; findings drive Pillar A policy and entitlement corrections (Protect AI, Garak, HiddenLayer, specialist red-team services).
Prompt audit log	Every prompt logged with user identity, session context, target AI system, and what the prompt caused to be retrieved or generated.
Agent action log	Every API invocation or workflow step by an AI agent logged with agent identity, parent human identity if applicable, action type, policy version, and outcome.
Secure browser AI interaction audit	Enterprise browsers (e.g., Island) provide a complete audit trail of browser-based AI interactions — what was input, output, and done with the output — covering AI tools accessed via browser where server-side logging is unavailable.

5.5 Implementation Patterns

5.5.1 Pattern D1: AI Prompt Input Inspection Pipeline

A prompt inspection layer sits between the user or agent interface and the AI model. It identifies the submitting identity, classifies the content of the prompt for sensitive data patterns, evaluates the target AI platform against the approved list for that identity and data classification, applies policy (block, redact, log, alert), and passes approved prompts through with a logged audit record. This pattern applies to internal AI platforms and to external AI tools accessed via browser or API.

5.5.2 Pattern D2: AI Output Inspection and Disclosure Enforcement

AI outputs are inspected after generation against two rulesets: standard DLP rules for sensitive data patterns, and Pillar C disclosure tier rules for the receiving identity. Output that fails either check is blocked, redacted, or watermarked. All outputs are logged with the full context needed for audit.

5.5.3 Pattern D3: AI Activity Logging Pipeline

A structured logging standard for AI interactions is consistent across all AI capabilities. Log events include: identity (human and agent), timestamp, AI system, action type (prompt, retrieval, generation, agent action), data sources accessed, policy version evaluated, and output classification. These logs feed the SIEM alongside traditional security event data.

5.5.4 Pattern D4: AI-Specific SIEM/SOAR Detection Rules

AI-specific detection rules cover: scope anomalies (AI accessing data far outside the calling identity's normal work pattern), volume anomalies (retrieval or generation volume far exceeding baseline), prompt injection indicators, output anomalies (sensitive data patterns inconsistent with the use case), and agent action anomalies (agent invoking APIs or writing records at unexpected times or with unusual parameters).

5.5.5 Pattern D5: Red-Teaming AI Systems

AI-specific adversarial tests include: prompt injection (crafting prompts that cause the model to ignore system instructions or reveal withheld data), data exfiltration via prompt (prompts that surface sensitive data despite retrieval filters), jailbreaks, and entitlement bypass (inducing an agent to use its entitlements beyond its intended purpose). Red-team findings drive Pillar A entitlement and policy corrections — not only model-level patches.

5.5.6 Pattern D6: The Technical Governance Feedback Loop

A formal process routes Pillar D technical findings back to Pillar A owners: DLP input violations identify identities attempting to submit unauthorized data types, driving entitlement review. DLP output violations signal Pillar C disclosure rules need tightening. Anomaly detections trigger Pillar A entitlement scope review. DSPM discoveries of unclassified or over-permissioned data in AI-accessible stores trigger Pillar A and Pillar B remediation. This feedback loop is maintained by security operations and distinct from Pillar E's governance oversight channel.

5.6 SOC Adaptation for AI Security Operations

Enterprises should deliberately evolve — and where necessary redesign — their security operations to account for AI as both a new source of risk and a new security capability.

SOC Adaptation	Description
AI-specific runbooks	Dedicated incident response procedures for AI scenarios: prompt injection confirmed, sensitive data exfiltration via AI input or output, rogue agent action, AI input DLP violation, AI system compromise.
AI-assisted SOC	Use AI (under appropriate governance from Pillars A-C) to accelerate alert triage, pattern discovery, and incident summarization — the same identity-aware principles apply to AI used by the SOC itself.
AI activity as first-class telemetry	AI interaction logs treated as security-relevant telemetry on par with endpoint and network logs — fed to SIEM with equal retention and analysis priority.
DSPM integration	DSPM findings trigger automatic protective actions: reclassification notification, access restriction review, AI scope assessment before expansion into newly discovered data stores.

5.7 Key Vendors and Solutions

Solution / Vendor	Role in Pillar D	Notes
Zscaler (CASB / ZIA DLP)	AI prompt input DLP and network-level controls for AI traffic; inspects prompts before they reach AI services; governs which AI platforms can receive corporate data.	Pillar D role is DLP and traffic inspection; access policy enforcement is a separate Pillar B function via ZPA.
Microsoft Purview (AI Hub + DLP)	AI prompt and output DLP for M365/Copilot; native Copilot activity logs; AI interaction audit trail for Sentinel ingestion.	Deep M365 integration; covers Copilot prompt and response governance end-to-end.
Lakera / Prompt Security	Purpose-built AI input inspection and prompt injection detection; integrates with AI orchestration frameworks.	Real-time inspection at the AI service boundary; designed for the AI input surface that traditional DLP misses.
Securiti AI / Varonis (DSPM)	Continuous data discovery, classification, and AI access posture management; identifies sensitive data in AI-reachable stores.	Proactive identification of data exposure before AI incidents occur.

Solution / Vendor	Role in Pillar D	Notes
Microsoft Sentinel + Purview	SIEM with native M365 Copilot activity integration; AI activity logs from Purview AI Hub; AI-specific analytics rules.	Best for Microsoft-centric environments with Copilot deployments.
Splunk	SIEM for AI activity log ingestion and AI-specific detection rules; broad ecosystem for custom integrations.	Flexible; AI activity logs added as a new source; growing library of AI security content.
Protect AI / Garak	Red-teaming and vulnerability scanning for ML models and AI pipelines; adversarial testing tooling.	Addresses model-layer and pipeline-layer risks that traditional security tools miss.
HiddenLayer	AI/ML security: model scanning, adversarial attack detection, AI supply chain security.	Focused on model-layer risks; complements access-layer governance in Pillars A-C.
Island Enterprise Browser (AI Protect)	Complete audit trail for browser-based AI interactions; DLP for both prompt inputs and outputs in the browser; observability over AI tool access where server-side logging is unavailable.	Closes the browser observability gap; Pillar D role only — governance of AI platform access is a Pillar A/B function.
CrowdStrike / SentinelOne	Endpoint and identity detection extended to AI agent behavior on endpoints and in cloud environments.	Existing EDR/XDR platforms with growing AI-aware detection capabilities.

5.8 Workbook Prompts

- D1. Do you have a prompt input inspection layer for any AI systems? Which AI tools do users access where prompt inputs are not inspected before reaching the model? Which carry the highest risk of sensitive data exposure via prompt?
- D2. Do you have visibility into what employees are inputting into browser-based AI tools (ChatGPT, Copilot web, other external AI services)?
- D3. Do your SIEM detection rules include AI-specific scenarios (prompt injection, unusual retrieval scope, agent action anomalies)? What AI-specific runbooks does your SOC have?
- D4. Have you red-teamed any AI systems? What were the most significant findings, and were they resolved at the Pillar A policy and entitlement level or only at the model level?

- D5. Is there a formal process for routing Pillar D technical findings — DLP violations, anomaly detections, red-team results — back to the Pillar A owners responsible for entitlement and policy corrections?

6 Pillar E: Enterprise AI Governance

The human oversight layer: translating leadership intent and risk appetite into policy, ensuring regulatory alignment, and providing board-level accountability for AI risk.

The strategic principle hypothesis on Enterprise AI Governance holds that enterprises should establish Enterprise AI Governance as a supervisory oversight function — typically a cross-functional process or board — explicitly embedded in a governance stack alongside data governance, security governance, and business transformation governance, because without such a peer function there is no durable mechanism to translate leadership intent and AI risk appetite into the concrete policies, controls, and change decisions those boards make.

Most critical where AI systems materially influence financial risk, regulatory exposure, safety, workforce outcomes, or customer decisions and trust.

6.1 The Problem

Pillars A through D are technical security mechanisms. They can control access according to pre-defined policies. But they cannot, on their own, answer the questions that ultimately determine whether an enterprise's AI is aligned with its values, obligations, and risk appetite:

- Is this AI initiative consistent with our board's articulated AI risk appetite?
- Have we assessed and disclosed the regulatory exposure this AI capability creates?
- Are there workforce, fairness, or reputational risks that no technical control can fully mitigate?
- Who is accountable when this AI system produces a consequential, wrong, or harmful output?
- Are our existing AI controls adequate to maintain our acceptable level of AI risk?

These questions require cross-functional human judgment, recurring oversight, and clear decision rights — not just better access policies or more sophisticated detection rules. Enterprise AI Governance (E-AIG) is the function that provides this judgment and accountability layer, sitting above and informing the technical pillars.

6.2 The Governance Stack

E-AIG is designed to be a peer governance function alongside existing governance boards that own adjacent domains:

Governance Board	Domain and Relationship to E-AIG
Data Governance Board	Owns data quality, semantics, and stewardship. E-AIG depends on data governance for classification standards that feed Pillar A policy and Pillar B retrieval filters; E-AIG escalates AI-specific data classification questions to data governance.
Security Governance Board	Owns security risk, controls, and compliance. E-AIG depends on security governance for the Pillar A-D technical control framework; security governance implements technical controls that E-AIG policy decisions require.
Transformation Governance Board	Owns operating model change, culture, and change management. E-AIG depends on transformation governance to manage workforce impacts of AI initiatives.
Enterprise Risk Management / Board	Owns enterprise risk appetite and board accountability. E-AIG routes significant AI risks to ERM; feeds board-level AI risk reporting; ensures AI risk is represented in the enterprise risk register.
Enterprise AI Governance (E-AIG)	Owns AI system behavior end-to-end: use case approval, lifecycle governance, cross-cutting AI risk, regulatory alignment, and translation of risk appetite into Pillar A-D technical controls. The function that sees the whole picture.

6.3 What E-AIG Should Own

E-AIG Responsibility	Description
AI risk appetite definition	Translate board and leadership risk appetite into explicit, scope-defined policies for what AI can and cannot be used for, and at what risk tier additional oversight is required.
AI use case triage and approval	A structured intake process that assesses new AI initiatives against risk tiers; routes low-risk cases to fast-track; applies full governance to high-risk cases.

E-AIG Responsibility	Description
Lifecycle governance	Oversight from use case approval through deployment, monitoring, and retirement — including scheduled re-assessment of live AI systems as context, regulations, or risk profiles change.
Regulatory and compliance alignment	Maintain a current map of AI-relevant regulatory obligations — including the EU AI Act (high-risk AI system requirements, GPAI obligations, prohibition categories) and sector-specific regulations — and ensure the enterprise's AI portfolio and controls satisfy those obligations. The NIST AI RMF (AI 100-1) Govern and Map functions provide a practical operational framework for this responsibility.
Cross-domain risk identification	Identify AI risks that fall between existing governance boards: fairness and bias, customer trust, AI-generated misinformation, workforce impact.
AI incident oversight	Define the threshold at which an AI-related security or compliance event escalates from technical teams to E-AIG; own the enterprise response and learning for significant AI incidents.
Board-level AI reporting	Produce regular, structured AI risk and performance reporting for board and audit committee; normalize AI risk in enterprise risk reporting.
Governance feedback to technical pillars	Translate E-AIG policy decisions — what AI can do with what data, for whom, under what conditions — into specific inputs to Pillar A policy engine configurations, Pillar B retrieval scope, and Pillar D monitoring rules. This is the strategic governance channel, distinct from the technical Pillar D feedback loop.

6.4 E-AIG Structure and Operating Model

6.4.1 Membership

E-AIG should include representatives from: CISO, CDO, CIO or CTO, Chief Risk Officer, General Counsel or Chief Compliance Officer, Chief People Officer, Chief Data and AI Officer where the role exists, and senior business representatives for high-risk AI use case domains.

6.4.2 AI Risk Tiers

Risk Tier	Characteristics and E-AIG Engagement
Tier 1: Low risk	Narrow, bounded use case; no sensitive data; limited consequential decisions. Fast-track approval; standard Pillar A-D controls. E-AIG reviews quarterly.
Tier 2: Moderate risk	Access to sensitive data; some consequential outputs. Standard approval; Pillar A-D controls plus DLP for inputs and outputs. E-AIG reviews at launch and periodically.
Tier 3: High risk	Significant regulatory exposure; influences financial, safety, workforce, or customer outcomes; broad data access. Full E-AIG review and approval; lifecycle checkpoints; board-level disclosure.
Tier 4: Prohibited	Use cases incompatible with regulatory obligations, ethical commitments, or board-articulated risk appetite. E-AIG maintains and publishes the prohibited list; enforced via Pillar A policy.

6.5 E-AIG Maturity Path

Maturity Stage	Characteristics
Stage 1: Ad hoc oversight	AI initiatives reviewed case-by-case; no standing governance function; decisions made informally by security, legal, or technical teams.
Stage 2: Scoped E-AIG	Standing E-AIG function with defined membership and cadence; governance applied to Tier 3 use cases; basic AI risk register; defined interface with data and security governance boards.
Stage 3: Integrated governance stack	E-AIG fully integrated with data, security, transformation, and risk governance; lifecycle-wide oversight for Tier 2 and 3 use cases; board-level AI risk reporting normalized; policy decisions systematically translated into Pillar A-D technical controls.

6.6 Key Vendors and Solutions

E-AIG is primarily a human governance function, not a technology function. The following tools support it:

Solution / Vendor	Role in Pillar E	Notes
OneTrust AI Governance	AI use case inventory, risk assessment workflow, and regulatory compliance mapping.	Purpose-built AI governance platform with risk tier management and audit trail.
ServiceNow GRC	AI risk and policy management integrated with existing GRC workflows.	Leverages existing GRC investments; AI risk as extension of enterprise risk management.
Responsible AI toolkits (IBM OpenScale, Microsoft Responsible AI)	Fairness, bias, and explainability assessment for AI systems under governance.	Technical inputs to E-AIG risk assessments; particularly for Tier 3 use cases.
EU AI Act compliance tooling (Armilla, Holistic AI)	Regulatory classification and compliance evidence management for EU AI Act obligations.	Critical for organizations with EU market exposure or regulated sector use cases.
Microsoft Purview Compliance Manager	Compliance posture management including AI-relevant regulatory controls.	Useful for Microsoft-centric environments with Copilot and Azure AI deployments.

6.7 Workbook Prompts

- E1. Does your organization have a standing Enterprise AI Governance function or board? If not, where do high-risk AI use case decisions currently get made — and who bears accountability if they go wrong?
- E2. Does your E-AIG have a defined interface with your data governance, security governance, and transformation governance boards? Are there documented decision rights for cross-domain AI questions?
- E3. Has your organization defined an AI risk tier model? Are existing AI initiatives tiered and governed accordingly?
- E4. Do you have a current map of AI-relevant regulatory obligations (EU AI Act, sector regulations, NIST AI RMF alignment)? Is there a named owner responsible for keeping that map current as regulations evolve?
- E5. Does your board receive structured, regular AI risk reporting? If not, what would a useful minimum board-level AI risk disclosure look like for your organization?

7 Architecture Summary: Putting the Pillars Together

The five pillars form a coherent, layered, interdependent architecture. They all likely exist in your enterprise already, but the key is to evolve them together with intention to unlock the business value of AI within your enterprise's business risk tolerance.

Architectural Layer	Pillar / Description
Shared policy authority (Pillar A)	The foundation: defines the authoritative identity-aware policies — composed of IGA platforms, policy engines, IdPs, CIEM tools, and SPIFFE/SPIRE workload identity infrastructure covering different environments and identity types. All other pillars depend on Pillar A for consistent, current, callable policy decisions.
Identity-aware retrieval (Pillar B)	Calls Pillar A at the point where AI reads data. Ensures the model receives only data that the calling identity is entitled to. MCP servers, where present, become a governed enforcement context within Pillar B.
Identity-aware abstraction (Pillar C)	Calls Pillar A at the point where AI formulates its output. Ensures the detail and disclosure depth of AI responses is bounded by the recipient identity's authorized clearance tier.
Post-AI security operations (Pillar D)	The safety net: monitors AI prompt inputs and outputs for policy violations and anomalies; feeds technical findings back to Pillar A as entitlement and policy corrections. Encompasses DLP, DSPM, SIEM/SOAR, and red-teaming.
Enterprise AI Governance (Pillar E)	The oversight layer: translates leadership intent and risk appetite into Pillar A policy; ensures regulatory alignment (EU AI Act, NIST AI RMF); routes governance-level insights back to all technical pillars. Provides board-level AI risk accountability.

The architecture works as a governed closed loop: Pillar A defines policies. Pillar B applies them at retrieval. Pillar C applies them at disclosure. Pillar D monitors for gaps and routes technical findings back to Pillar A. Pillar E provides the human oversight that gives the loop institutional legitimacy and strategic direction — and routes governance-level signals to all technical pillars when strategy, regulations, or risk appetite changes.

7.1 The Closing Challenge

For every AI use case in your portfolio, you should be able to answer these questions with confidence:

- Who or what is this AI acting for, and what is that identity entitled to read, transform, and reveal? (Pillar A)
- Is there a specific, enforced, auditable point in the architecture where that entitlement is checked before data is retrieved? (Pillar B)
- Is there a specific, enforced, auditable point where the output's disclosure depth is checked against the recipient's authorized tier? (Pillar C)
- Are both the inputs to this AI (prompts) and its outputs (responses, actions) monitored against security policy? (Pillar D)
- Does a cross-functional governance body have visibility, accountability, and decision authority over this AI's use and risk? (Pillar E)

If you cannot answer all five questions with confidence, you have governance gaps. This workbook is designed to help you close them systematically, with patterns you can reuse and evolve as AI adoption expands across your enterprise.

To discuss how these principles apply to your organization, contact the author at 1joegottlieb@gmail.com or visit www.enttao.com.

8 Appendix A: Inter-Pillar Interfaces

The five pillars are interdependent. Each pillar produces outputs that other pillars consume. This appendix codifies the interfaces between pillar functions — what flows across each interface, whether it is uni- or bi-directional, and the implementation variants that enterprises typically use.

Note: where an interface is described as bi-directional, it means that both pillars exchange information in both directions — but the nature and timing of what flows in each direction typically differ, as described in the interface detail tables below.

8.1 Interface Summary

Interface	Direction	What Flows Across It
A-B	Bi–directional	A to B: Entitlement definitions and runtime policy decisions consumed by retrieval enforcement points. B to A (feedback): Retrieval coverage gaps and ungoverned access paths identified during Pillar B operations surfaced back to Pillar A for policy correction.

Interface	Direction	What Flows Across It
A-C	Bi–directional	A to C: Policy definitions for read/transform/reveal rules and clearance tier assignments consumed by abstraction layer output governance. C to A (feedback): Abstraction tier compliance gaps and seal-break events surfaced back to Pillar A for policy refinement.
A-D	Bi–directional	A to D: Policy version metadata for audit log correlation, entitlement state for anomaly detection baselines, and credential or access change events for SOC alerting. D to A: DLP violations, anomaly detections, red-team findings, and DSPM discoveries consumed by Pillar A for entitlement corrections and policy updates — the primary technical governance feedback loop.
A-E	Bi–directional	A to E: Policy implementation status, entitlement coverage metrics, and access certification results consumed by E-AIG for governance reporting and risk assessment. E to A: E-AIG policy decisions translated into Pillar A policy engine configurations, IGA role definitions, and prohibited action lists.
B-D	Bi–directional	B to D: Retrieval audit logs (which identity retrieved what, from which corpus, under which policy version) consumed by SIEM for anomaly detection and DLP for scope analysis. D to B: Retrieval scope anomalies and prompt injection findings fed back into Pillar B pre-retrieval filter tuning.
B-E	Uni–directional	Retrieval coverage reports and ungoverned AI access paths surfaced to E-AIG for risk register and use case governance decisions.
C-D	Bi–directional	C to D: Output classification and clearance tier applied to each AI response consumed by DLP output inspection and disclosure audit logs. D to C: Output DLP violations and disclosure anomalies fed back into Pillar C clearance tier and reveal policy corrections.
C-E	Uni–directional	Abstraction tier compliance and seal-break event reports consumed by E-AIG for risk reporting and disclosure policy review.

Interface	Direction	What Flows Across It
D-E	Bi–directional	D to E: AI security incident summaries, DLP trend data, and red-team findings consumed by E-AIG for risk register updates and board-level AI risk reporting. E to D: AI risk tiers and incident escalation thresholds defined by E-AIG consumed by Pillar D as criteria for alert severity classification and escalation routing.

8.2 Interface Detail: A-B (Policy-Retrieval)

This is the most operationally critical interface. Every AI retrieval call depends on Pillar A delivering an accurate, current authorization decision.

Variant	How Policy Reaches Enforcement	Interoperability Considerations
IGA API call at request time	Enforcement point calls IGA platform (SailPoint, Entra) for the calling identity's current entitlements immediately before constructing the retrieval filter.	Highest accuracy; some latency; IGA must be available in the critical path; entitlements always current.
JWT / token claims	Identity token carries entitlement claims (groups, roles, sensitivity clearance) issued by IdP; enforcement point reads claims from token.	Low latency; claims may be stale if token is long-lived; requires token refresh strategy aligned with IGA lifecycle events.
Policy engine sidecar (OPA, Cerbos, Cedar)	Enforcement point calls a policy engine sidecar with identity context; engine evaluates against a locally cached, periodically synchronized policy bundle.	Sub-millisecond latency; bundle must be synchronized with IGA state; cache invalidation on entitlement change is critical.
Pre-computed metadata filter	Retrieval query includes pre-computed entitlement filter (e.g., department = Finance, classification <= Confidential) derived from IGA at session start.	Fastest; risk of stale filters if entitlements change mid-session; acceptable for low-risk use cases, insufficient for high-risk.

Variant	How Policy Reaches Enforcement	Interoperability Considerations
SPIFFE SVID + mTLS	AI agent authenticates to retrieval service using SPIFFE SVID (X.509); service mesh enforces mTLS; policy engine reads SPIFFE ID as identity context for authorization decision.	Suitable for distributed agent retrieval across multi-cloud; eliminates static credentials; requires SPIRE server deployment.
MCP Enterprise Authorization	AI client accesses data via MCP server; Enterprise-Managed Authorization routes authorization through enterprise IdP; OAuth 2.1 token scoped to permitted MCP resources; policy governed by Pillar A via IdP.	Converts MCP from uncontrolled access to governed Pillar B enforcement; requires MCP server to implement Enterprise-Managed Authorization spec.

8.3 Interface Detail: A-C (Policy-Abstraction)

Pillar C depends on Pillar A for the enterprise agent identity's entitlements (what it may read and transform) and the clearance tier assignment for each recipient identity (what it may reveal to this specific caller).

Variant	How Policy Reaches Enforcement	Interoperability Considerations
IGA-managed role / clearance attributes	Recipient clearance tier stored as an IGA-managed attribute; policy engine reads it at output generation time.	Consistent with IGA governance; requires IGA to carry clearance tier as a first-class attribute with lifecycle management.
Purview sensitivity labels	Recipient clearance determined by Microsoft Purview sensitivity label on user record or group membership.	Native to M365 environments; deep Copilot integration; limited to Microsoft AI ecosystem.
Policy engine rule with identity context	Policy engine evaluates clearance tier from identity context at output formulation time.	Most flexible; requires policy to encode tier logic explicitly; all tier variants must be modeled in the policy store.

Variant	How Policy Reaches Enforcement	Interoperability Considerations
Structured output schema with field-level access control	Output schema defines which fields are populated for which clearance levels; policy engine resolves field set at output time.	Clean for structured outputs; less suited for free-form narrative AI responses.

8.4 Interface Detail: A-D (Policy-Operations)

This feedback interface is the least formalized in most organizations and the most important for keeping governance effective over time. Without it, Pillar D findings accumulate as alerts without improving the upstream architecture.

Variant	How Findings Reach Policy	Interoperability Considerations
Manual escalation	SOC analyst or DLP team raises a ticket or email escalation to the IGA/policy team when a finding requires an entitlement or policy change.	Simple; slow; prone to deprioritization; no systematic tracking of resolution.
Automated SOAR playbook with IGA API call	SOAR playbook triggers an automated IGA access review or entitlement suspension when specific alert thresholds are crossed.	Fast; requires SOAR-IGA integration; scope of automation must be carefully defined to avoid false-positive revocations.
AI risk register owned by E-AIG	Pillar D findings classified and logged in AI risk register; E-AIG reviews on a defined cadence and routes policy/entitlement changes to Pillar A owners.	Systematic; governed; slower than automated but appropriate for policy-level changes requiring governance judgment.
SIEM-to-IGA event correlation	SIEM correlates AI activity anomalies with IGA access events; correlated findings surfaced simultaneously to SOC and IGA governance teams.	Most comprehensive; requires investment in SIEM-IGA integration; maximizes both detection speed and governance response quality.

8.5 Interface Detail: A-E (Policy-Governance)

This interface translates E-AIG policy decisions — expressed in governance terms — into specific technical artifacts in Pillar A. The quality of this translation determines whether governance intent is faithfully implemented.

Variant	How Governance Reaches Policy	Interoperability Considerations
Governance decision to manual policy authoring	E-AIG decision document reviewed by policy engineers who manually update OPA/Cedar/Cerbos rules or IGA configurations.	Human judgment at translation; risk of interpretation gaps; recommended for complex or novel policy decisions.
Governance decision to IGA role change	E-AIG defines new AI access roles or clearance tiers; IGA team implements as role definitions and assignments.	Natural for IGA-managed entitlements; works well for use case approvals that map to specific data access grants.
Governance decision to prohibited action list	E-AIG adds use cases or prompt patterns to a prohibited list; Pillar A policy engine enforces via Pillar D DLP input rules.	Effective for categorical prohibitions; requires clear specification of what constitutes the prohibited pattern.
GRC platform with policy translation workflow	E-AIG decisions logged in GRC platform (ServiceNow, OneTrust); workflow routes implementation tasks to technical pillar owners with tracking.	Most auditable; highest governance rigor; requires GRC-pillar integration investment.

9 Appendix B: Key Vendor Solution Pairs Across Inter-Pillar Interfaces

This appendix maps known vendor solution pairs to each inter-pillar interface. For each pair it notes the interface spanned, the nature of the integration, and key interoperability considerations. This is a living reference: verify current integration depth with each vendor before architectural commitment.

9.1 Interface A-B: Policy-Retrieval

Vendor Pair	Integration Description	Interoperability Notes
SailPoint ISC + Azure AI Search	SailPoint entitlements for SharePoint content synchronized to Azure AI Search metadata fields, enabling permission-scoped vector search that reflects IGA-managed access.	Requires consistent metadata schema between SailPoint and AI Search index; IGA change events must trigger index metadata updates.
Microsoft Entra ID Governance + M365 Copilot	Entra ID Governance manages SharePoint, Teams, and Exchange entitlements that Copilot inherits via Microsoft Graph OBO (RFC 8693); access reviews and lifecycle workflows in Entra directly govern Copilot retrieval scope.	Native, deep integration; Entra access reviews are the primary governance mechanism for Copilot retrieval scope.
Cerbos + LlamaIndex / LangChain	Cerbos policy engine called as pre-retrieval authorization step in LlamaIndex or LangChain RAG pipelines; policy bundle defines which identity attributes permit access to which document metadata.	Low-latency; Git-versioned policy; requires consistent attribute passing from identity token to Cerbos call.
OPA + Pinecone / Weaviate	OPA policy evaluates identity context and returns allowed metadata filter set; filter applied to Pinecone or Weaviate vector query before semantic search executes.	OPA policy bundle must be kept synchronized with IGA entitlement state via event-driven update.
Saviynt + AWS Bedrock (Knowledge Bases)	Saviynt manages AWS IAM entitlements and service account permissions; Bedrock Knowledge Bases use IAM for retrieval authorization; Saviynt governs IAM role assignments that determine Bedrock retrieval scope.	Native AWS IAM integration; Saviynt CIEM module adds cloud entitlement visibility.
Cedar (AWS Verified Permissions) + AWS Bedrock Agents	Cedar policy language used to define agent retrieval entitlements; Verified Permissions called by Bedrock agent orchestrator before each retrieval or tool invocation.	Native AWS integration; Cedar's formal verification provides high-confidence authorization.

Vendor Pair	Integration Description	Interoperability Notes
OPA + Kong / Apigee (API Gateway for Agent Retrieval)	OPA sidecar deployed with API gateway; evaluates agent retrieval entitlements before API calls are forwarded to data services.	Strong for agentic retrieval authorization; requires OPA policy to model agent retrieval entitlement schema.
SPIFFE/SPIRE + service mesh (Istio/Linkerd)	SPIRE issues SVIDs to AI agent workloads; Istio/Linkerd service mesh enforces mTLS using SPIFFE certificates; policy engine reads SPIFFE ID as identity context for retrieval authorization decisions.	Eliminates static service credentials; suitable for distributed AI agents across Kubernetes and multi-cloud; requires SPIRE server and mesh deployment.
Okta XAA + MCP servers	Okta Cross-App Access (XAA), incorporated in MCP spec, routes MCP client authorization through Okta as IdP; OAuth 2.1 tokens scoped to permitted MCP methods govern AI agent access to MCP-exposed data and tools.	Converts MCP from shadow-IT access pattern to governed Pillar B enforcement; requires MCP server implementation of Enterprise-Managed Authorization extension.

9.2 Interface A-C: Policy-Abstraction

Vendor Pair	Integration Description	Interoperability Notes
SailPoint ISC + Guardrails AI	SailPoint manages clearance tier attributes on user identities; at output generation, clearance tier passed to Guardrails AI as a topic/field restriction parameter.	Requires SailPoint to expose clearance tier via API or token claim; Guardrails AI configured with tier-aware topic rules.
Microsoft Entra ID + Purview Sensitivity Labels + Copilot	Entra group membership determines Purview sensitivity label policy scope; Purview labels gate what Copilot includes in responses for labeled content.	Deep native M365 integration; sensitivity labels as de facto clearance tier for Copilot; limited to M365 ecosystem.

Vendor Pair	Integration Description	Interoperability Notes
SailPoint ISC + Databricks Unity Catalog	SailPoint manages user and enterprise agent entitlements; Databricks Unity Catalog enforces row/column-level security; SailPoint entitlements map to Unity Catalog group memberships.	Requires mapping between SailPoint role model and Databricks group model; change events must propagate to Unity Catalog.
OPA + LangChain output filter	OPA policy encodes read/transform/reveal rules; LangChain calls OPA at output formatting step to determine which response fields are populated for the calling identity's clearance tier.	Flexible; requires output schema design supporting field-level tier enforcement; OPA call adds latency at output stage.

9.3 Interface A-D: Policy-Operations

This interface is bi-directional in the architecture, but the most mature vendor integrations today are concentrated in the operations-to-policy feedback path shown below.

Vendor Pair	Integration Description	Interoperability Notes
Microsoft Sentinel + SailPoint ISC (via SOAR)	Sentinel SOAR playbook triggered by AI anomaly detection calls SailPoint API to initiate an access review, suspend an entitlement, or flag an identity for review.	Requires SailPoint SOAR connector; scope of automated action must be carefully defined; recommended for high-confidence anomalies only.
Microsoft Sentinel + Microsoft Entra ID	Native Sentinel-Entra integration: playbooks trigger Entra conditional access policy changes or user risk elevation in response to AI anomaly detections.	Deep native integration; risk elevation in Entra triggers step-up authentication or access restriction automatically.
Splunk SOAR + SailPoint ISC	Splunk SOAR connectors for SailPoint allow playbooks to trigger IGA actions in response to AI security events in Splunk.	Well-suited for organizations using Splunk as primary SIEM; requires SOAR SailPoint connector configuration.

Vendor Pair	Integration Description	Interoperability Notes
Zscaler (DLP event) + SailPoint ISC	Zscaler DLP events for AI prompt input or output policy violations correlated with SailPoint entitlement data to identify systemic access gaps; manual or automated remediation workflow.	Currently requires API or SIEM correlation; Zscaler-SailPoint native integration depth varies by deployment.
Lacework / Wiz (CIEM alert) + SailPoint ISC	Cloud entitlement anomaly detected by CIEM tool triggers SailPoint access review for the affected AI workload or service account.	Closes the loop between cloud infrastructure entitlement drift and IGA-governed remediation; requires API integration.

9.4 Interface A-E: Policy-Governance

This interface is bi-directional: Pillar A reports policy coverage and operational health upward to governance, while Pillar E translates governance decisions back into enforceable policy changes.

Vendor Pair	Integration Description	Interoperability Notes
OneTrust AI Governance + SailPoint ISC	OneTrust AI use case approval decisions trigger SailPoint access request workflows; AI system access granted only after E-AIG approval recorded in OneTrust.	Requires webhook or manual integration between OneTrust approval events and SailPoint provisioning workflows; closes loop between governance approval and technical access grant.
ServiceNow GRC + OPA / Cerbos	ServiceNow GRC records E-AIG policy decisions; engineering teams consume these to author policy-as-code updates in OPA or Cerbos; change management workflow tracks from governance decision to policy deployment.	Standard GRC-to-engineering workflow; maturing with AI governance tooling.

Vendor Pair	Integration Description	Interoperability Notes
OneTrust AI Governance + Microsoft Purview	OneTrust documents AI data handling policies; Purview implements corresponding sensitivity label policies and DLP rules; approved use cases in OneTrust trigger Purview configuration changes.	Native integration emerging; requires mapping from OneTrust policy concepts to Purview label taxonomy.
ServiceNow GRC + SailPoint ISC (role provisioning)	E-AIG Tier 3 use case approval in ServiceNow triggers SailPoint role provisioning workflow for AI system access; E-AIG approval is a prerequisite condition in the SailPoint provisioning workflow.	Strong governance control; approval is a gate in the technical provisioning path, not a parallel track.
SailPoint ISC + ServiceNow GRC / OneTrust	SailPoint access certification reports, entitlement coverage metrics, and policy exception logs exported to GRC platform; E-AIG consumes in AI risk register and governance reporting.	Requires report export automation; SailPoint governance dashboards provide source data for E-AIG oversight of Pillar A coverage.
Microsoft Entra ID Governance + Purview Compliance Manager	Entra access review completion rates, lifecycle workflow coverage, and Copilot permission scope reports surfaced in Purview Compliance Manager; E-AIG reviews AI entitlement governance posture alongside broader compliance posture.	Native M365 integration; best suited for Microsoft-centric AI governance programs.
OPA / Cerbos policy metrics + SIEM dashboards	Policy evaluation coverage, policy decision distribution, and policy error rates surfaced from policy engine telemetry; E-AIG consumes as indicators of Pillar A operational health.	Requires investment in policy engine observability; increasingly important as policy-as-code matures in the enterprise.

Vendor Pair	Integration Description	Interoperability Notes
SPIRE + SIEM (operational health)	SPIRE certificate issuance rates, attestation failures, and SVID expiry events surfaced to SIEM; E-AIG uses as indicator of workload identity infrastructure health and coverage gaps across AI agent deployments.	Requires SPIRE telemetry export; increasingly relevant as SPIFFE-based agent identity becomes standard in distributed AI environments.

9.5 Interface B-D: Retrieval-Operations

Vendor Pair	Integration Description	Interoperability Notes
Azure AI Search + Microsoft Sentinel	Azure AI Search emits query logs including identity context; Sentinel ingests as custom data connector; AI-specific analytics rules detect unusual retrieval scope against identity baseline.	Native Azure integration; requires enabling diagnostic logging on AI Search; Sentinel workbook for AI retrieval anomaly detection.
LlamaIndex / LangChain + Splunk	RAG pipeline emits structured trace logs with identity, query, retrieved documents, and applied policy version; Splunk ingests via HTTP Event Collector; AI-specific detection rules built on indexed trace data.	Requires instrumentation of RAG pipeline to emit structured events; growing Splunk AI security content library.
Cerbos + Splunk / Elasticsearch	Cerbos emits structured authorization decision logs for every retrieval authorization call; SIEM ingests for anomaly detection and audit.	Clean audit trail for policy decisions; Cerbos decision log schema maps naturally to SIEM alert criteria.
OPA + Elastic SIEM	OPA decision logs (identity, policy version, decision, resource) exported to Elastic SIEM; AI retrieval anomaly detection rules built on OPA audit data.	Requires log forwarding from OPA to Elastic; well-suited for organizations using Elastic for security analytics.

Vendor Pair	Integration Description	Interoperability Notes
SPIFFE/SPIRE + SIEM (Splunk / Sentinel)	SPIRE audit logs record SVID issuance, renewal, and attestation events per agent workload; SIEM ingests for anomaly detection on agent identity lifecycle and unusual certificate request patterns.	Provides workload-level identity audit trail alongside application-level retrieval logs; useful for detecting impersonation or unauthorized agent deployments.

9.6 Interface C-D: Abstraction-Operations

Vendor Pair	Integration Description	Interoperability Notes
Microsoft Purview AI Hub + Microsoft Sentinel	Purview AI Hub logs Copilot prompt and response activity including sensitivity label context; Sentinel ingests for DLP violation detection and disclosure anomaly alerting.	Native M365 integration; covers Copilot AI interactions end-to-end; limited to Microsoft AI ecosystem.
Guardrails AI + SIEM (Splunk / Sentinel)	Guardrails AI emits policy violation events (topic blocked, output redacted, disclosure rule triggered) as structured logs; SIEM ingests for SOC alerting on disclosure violations.	Requires custom log forwarding integration; Guardrails event schema well-suited for SIEM rule authoring.
NeMo Guardrails + Datadog / Splunk	NeMo guardrail policy triggers logged with context; forwarded to observability platform; AI-specific dashboards for disclosure compliance tracking.	Good for LLM-heavy environments using NVIDIA stack; requires log forwarding configuration.
OPA (output filter) + Sentinel / Splunk	OPA decision logs for output-stage authorization calls (clearance tier evaluations) forwarded to SIEM; disclosure anomalies detected against baseline clearance tier usage.	Consistent with Pillar A policy-as-code approach; same OPA audit infrastructure covers both retrieval and output decisions.

9.7 Interface D-E: Operations-Governance

Vendor Pair	Integration Description	Interoperability Notes
Microsoft Sentinel + ServiceNow GRC	Sentinel AI security incidents escalated to ServiceNow GRC as AI risk register entries when they meet E-AIG escalation thresholds.	Requires threshold definition in Sentinel playbooks; ServiceNow GRC provides structured tracking of AI risk items.
Splunk + OneTrust AI Governance	Splunk AI security trend reports exported to OneTrust AI risk register; E-AIG reviews aggregated AI security posture in OneTrust alongside use case governance.	Currently requires manual or scheduled export; integration automation emerging.
Protect AI / Garak (red-team findings) + ServiceNow / OneTrust	Red-team and adversarial testing findings logged as structured risk items in GRC platform; E-AIG reviews and determines whether findings require policy-level response.	Indirect integration via ticketing; structured findings improve governance response quality even without direct API integration.