Arthur Whitepaper

The Enterprise Guide to Agent Discovery & Governance (ADG)

Contents

01 The Agent Explosion is Here

02 Why Traditional Oversight Falls Short

03 The Shadow Agent Crisis

04 Building Your Agent Discovery Strategy

05 Three Pillars of A Rock-Solid Agent Governance Strategy

06 The Governance Stack in Detail

07 How Arthur AI Powers ADG at Scale

01

The Agent Explosion is here

As enterprises race to deploy AI agents across every business function, a dangerous visibility gap has emerged. Agents are proliferating faster than organizations can track them, creating security blind spots, compliance risks, and operational chaos. This whitepaper provides a comprehensive framework for building a solid Agent Discovery & Governance (ADG) strategy that transforms agentic sprawl into a managed, high-performing AI portfolio.

The enterprise AI landscape has shifted dramatically. What began as experimental chatbots and copilots has rapidly evolved into autonomous agents that reason, plan, and act on behalf of employees and organizations. These agentic AI systems combine large language models with orchestration, tool access, and memory to automate complex workflows: customer support agents that resolve issues autonomously, internal automation that routes requests and updates systems without human handoffs, and product-facing agents that synthesize complex documentation into actionable decisions. These use cases are not futuristic; they are achievable today with current AI tooling, and the business cases are compelling.

Yet with all their promise, agents introduce a new category of enterprise risk. Unlike traditional software, agents are probabilistic, adaptive, and increasingly autonomous. They call external tools, access sensitive data, and make decisions that can have real business consequences. As deployment accelerates, the gap between innovation and oversight is widening at an alarming rate.

> *Security and risk are the #1 barrier to scaling AI agents*
>
> — McKinsey, 2026

This whitepaper introduces Agent Discovery & Governance (ADG) as the essential strategy to resolve that tension. ADG provides the framework enterprises need to gain full visibility into their agentic ecosystem, enforce consistent governance policies, and continuously improve agent reliability — all while enabling teams to innovate with confidence.

02

Why Traditional Oversight Falls Short

Enterprise IT has decades of experience governing traditional software systems. Configuration management databases (CMDBs), security operations control (SOC) monitoring and audit logs have served organizations well in the deterministic software era; however, agentic AI fundamentally breaks the assumptions these systems were built on.

Agents Are Probabilistic, Not Deterministic

Traditional software produces the same output given the same input. Agents do not. The same query can produce different reasoning chains, tool selections, and outputs depending on context, model state, and retrieval results. This means that traditional testing approaches, where you verify expected outputs against known inputs, are insufficient. You need continuous behavioral evaluation that accounts for the probabilistic nature of these systems.

Agents Need Broad Access Permissions to Function Effectively

Unlike traditional applications that execute predetermined logic, agents reason about which tools to call, what data to access, and how to sequence actions. An agent with overly broad permissions can access sensitive data it was never intended to reach, update systems it should not touch, or take actions that have irreversible business consequences. Traditional role-based access control was not designed for autonomous reasoning systems that dynamically select their own execution paths.

Agents Proliferate Faster Than Any Prior Technology

The barrier to creating an agent is remarkably low. Developers can easily spin up an agent in minutes using frameworks like LangChain, Crew AI, or cloud-native tools from AWS Bedrock, Google Vertex AI, or Microsoft Agent Foundry. Meanwhile, existing enterprise software vendors are embedding agents into products that have been deployed for years. The result is exponential growth in agent count with no centralized inventory or oversight.

The bottom line: Enterprises need a fundamentally new approach to discovering, governing, and improving AI agents. That approach is Agent Discovery & Governance (ADG).

The Shadow Agent Crisis

In 2026, a challenge more consequential than shadow AI has emerged: shadow agents. Unlike shadow AI, which involved individuals using unsanctioned external LLM tools, shadow agents are autonomous systems operating inside the enterprise, often without being cataloged, monitored or governed.

Think of the iceberg analogy: the registered agents your organization knows about are just the tip, while a much larger mass of unregistered agents operates below the surface.

Shadow agents arrive through three primary vectors.

VECTOR 1 Internal Application Development

Engineering teams across the enterprise are incorporating AI agents into nearly every new software project. Different teams use different frameworks, different cloud providers, and different orchestration patterns. Without a centralized registry, each new agent is an unknown risk surface.

VECTOR 2 New Third-Party Solutions

A wave of AI-native startups and technology vendors are building agent-powered solutions for legal, finance, customer service, HR, and every other enterprise function. Each new procurement brings agents into the environment that the central AI team may not know about.

VECTOR 3 Existing Software Updates

Perhaps the sneakiest vector: established enterprise software vendors are embedding agents into products that have been deployed for a decade or more. A routine software update to a CRM, financial ledger, or project management tool can introduce agentic capabilities into systems that were previously fully deterministic. Even if an organization purchases no new software, its agent count can grow significantly through vendor updates alone.

The compound effect of these three vectors is that enterprises are going from dozens of agents to thousands and tens of thousands — with no systematic way to track what exists, where it runs, what data it accesses, or what actions it can take. Without a comprehensive discovery and governance strategy, every new agent is a potential security incident, compliance violation, or operational failure waiting to happen.

Building Your Agent Discovery Strategy

You cannot govern what you cannot see. The first pillar of any ADG strategy is establishing comprehensive, automated discovery of every agent operating in your environment.

Manual self-reporting processes cannot keep pace with the rate of agent proliferation. Organizations need automated, multilayered detection that catches agents regardless of how they enter the enterprise.

Technique 1: Telemetry-Based Discovery

The industry is coalescing around OpenTelemetry (OTEL) as the standard for agent telemetry. By implementing listeners on OTEL streams across your environment, you can detect new agents, new tools, configuration changes, and behavioral patterns.
Forward-leaning enterprises establish an enterprise-wide standard for agent telemetry collection as a foundational capability. Scanners in your OTEL-supported cloud loggers can discover agent framework signatures automatically, providing continuous visibility into what's running and how it's behaving.

Technique 2: MCP Server Monitoring

The Model Context Protocol (MCP) is becoming the standard interface through which agents expose capabilities to other agents and tools — think of it as the agent equivalent of APIs, but designed for unstructured reasoning. By monitoring for new MCP servers appearing in your environment and tracking changes to existing ones, you can flag new agents as they come online and detect capability changes in real time.

Technique 3: Network Layer Analysis

By analyzing network traffic for LLM API call signatures — whether through a dedicated LLM proxy or general network monitoring — you can detect AI usage that may not be instrumented through telemetry or MCP. Monitoring HTTP bodies of network requests for LLM-specific patterns catches agents built with non-standard frameworks or those that bypass standard instrumentation.

Technique 4: API-Driven Discovery

Cloud AI platforms like AWS Bedrock, Google Cloud Vertex AI, and others are beginning to offer API endpoints that advertise running agents and their configurations. While this technique is still maturing, it provides valuable coverage for agents built on managed cloud services. Enterprises should push their platform vendors to improve these discovery APIs, as this vector will become increasingly important.

Discovery alone is not enough. Once an unregistered agent is detected, your ADG platform should make it easy to triage and onboard that agent into your governance framework. This means assigning it to an application, identifying an accountable owner, classifying its risk level, and applying the appropriate governance policies.

The goal is to make this process as frictionless as possible so that it scales to the thousands of agents running across a large enterprise. When a new unregistered agent is flagged, you should be able to quickly organize it into an application that can then be governed — turning unknown risk into managed operations.

BEST PRACTICE

Establish automated discovery as a continuous process, not a one-time audit. New agents appear every single day through development, procurement, and software updates. Your discovery infrastructure must run continuously to maintain real-time visibility.

Three Pillars of A Rock-Solid Agent Governance Strategy

Once you have visibility into the agents running across your enterprise, the next challenge is establishing effective governance. This is the question enterprise leaders are asking more than any other right now: how do I effectively govern these systems at scale?

Three Pillars of A Rock-Solid Agent Governance Strategy

PILLAR 1

A Unified Policy Framework

The most common governance failure is fragmentation. When individual application teams implement their own guardrails and monitoring in isolation, the result is inconsistent standards, gaps in coverage, and no centralized reporting.

A unified policy framework establishes organization-wide standards for agent governance while allowing the flexibility needed for diverse use cases. This framework should cover guardrails (input and output controls), continuous evaluations, access management policies, monitoring and alerting rules, and audit trail requirements.

One of the most interesting shifts happening right now is the growth of first-line governance — application teams themselves demanding governance capabilities because they don't feel comfortable pushing agents to production without the right controls. This is distinct from traditional second-line (compliance) and third-line (audit) governance functions, though those remain important too.

PILLAR 2

Agnostic Governance

Agents in a typical enterprise are built on multiple frameworks, deployed across multiple cloud providers, and powered by multiple LLM providers. Your governance approach must be able to monitor and manage agents across all of these environments and stacks.

Agnostic governance means implementing a single, central AI control plane that works across AWS, Google Cloud, Azure, and any other environment where agents operate. It must integrate seamlessly with agents built on any framework — whether Vertex AI, Bedrock, Agent Foundry, LangChain, Crew AI, or custom implementations. This ensures that governance standards are consistent regardless of technology choices made by individual teams.

PILLAR 3

Customizable, Use-Case-Specific Policies

One-size-fits-all governance is doomed to fail. A customer support agent for an airline has fundamentally different governance requirements than an inventory management agent for a warehouse or a patient intake agent for a hospital.

With traditional human-based governance, supervisors naturally adapt their oversight approach from one context to another — monitoring customer service reps for friendliness and brand compliance while monitoring back-office workers for accuracy and process adherence. As you automate governance, you need equally adaptable automation.

Your governance platform must support highly customizable policies that can be tailored to each agent's specific context, risk profile, and business requirements.

The Governance Stack in Detail

Real-Time Guardrails

Guardrails are the first line of defense — automated controls that evaluate every agent interaction in real time. Essential guardrails include:

- PII Detection and Handling
- Toxicity and Brand Safety Filtering
- Hallucination Detection
- Prompt Injection Defense

Continuous Evaluations

Beyond real-time guardrails, agents need ongoing behavioral evaluation — automated assessors that continuously monitor agent performance against defined quality standards. These evaluators act as the automated equivalent of a human supervisor, monitoring every interaction for compliance, accuracy, and quality. Examples include:

- Tone and brand guideline adherence
- Answer correctness and factual consistency
- Goal accuracy
- Context recall
- Topic adherence
- Domain-specific evaluators

Access Management Policies

Agents require the same least-privilege principles applied to human users, but implemented through automated enforcement:

- Explicitly defined tool access boundaries for each agent
- Database read and write permissions scoped to the minimum required
- API access controls limiting which external systems the agent can interact with
- Clear escalation paths for actions that exceed the agent's authorized scope

Monitoring & Alerting

Production agents require comprehensive observability:

- Full trace capture of every prompt, retrieved context, tool selection, parameter, result, and approval
- Configurable alerting when governance policies are violated or performance thresholds are crossed
- Centralized dashboards providing real-time visibility into agent health, cost, and compliance across the entire portfolio

How Arthur AI Powers ADG at Scale

Arthur AI built the industry's first comprehensive Agent Discovery & Governance platform, purpose-designed to address the challenges outlined in this whitepaper. Developed through our extensive work with Fortune 500 enterprises deploying production-grade agents from financial services, to customer operations, to data platforms, and more, the Arthur platform provides end-to-end ADG capabilities.

Automated Agent Discovery

Arthur automatically scans cloud environments to discover and catalog agents as they appear, using the multilayered detection techniques described in this whitepaper — telemetry monitoring, MCP server detection, network layer analysis, and API-driven discovery. Unregistered agents are flagged for triage and can be quickly onboarded into your governance framework.

Agnostic, Cross-Platform Governance

Arthur integrates seamlessly whether you are building on Google Cloud Vertex AI, AWS Bedrock, Microsoft Agent Foundry, or any other platform. This allows engineering teams to use the best tools for each job while maintaining a single governance standard across the enterprise. Arthur is available on the Google Cloud Marketplace or AWS Marketplace for streamlined procurement, with spend counting toward existing commit thresholds.

Customizable Policy Engine

Arthur's governance platform supports highly customizable guardrails, evaluators, and access management policies. Out-of-the-box capabilities include PII detection, toxicity filtering, hallucination detection, and prompt injection defense. Custom evaluators can be configured for any domain-specific requirement — from ensuring responses are grounded in clinical data in healthcare to SQL semantic equivalence in data analytics to brand guideline adherence in customer-facing agents.

Business-Aligned Metrics

Arthur ties agent performance directly to relevant business KPIs, providing centralized dashboards that communicate agent health, cost, compliance status, and evaluation pass rates to diverse stakeholders — from application developers to compliance officers to the C-suite.

End-to-End ADLC Support

Arthur supports the complete Agent Development Lifecycle:

- Full agent observability across prompts, tool calls, decisions, and outcomes
- Automated evaluations that replace subjective "vibe checks" with measurable reliability signals
- Rapid iteration and experimentation — compare prompts, models, and agent logic without introducing regressions
- Production-grade monitoring with real-time failure detection and configurable alerts
- Policy enforcement ensuring outputs are on-brand, sensitive data is secure, and agent behavior aligns with organizational standards

Ready to build your ADG strategy?

Arthur AI's platform provides the industry's most comprehensive Agent Discovery & Governance capabilities — trusted by Fortune 500 enterprises to govern mission-critical agent deployments.

Book a demo → Talk to our AI experts to understand how Arthur can bring visibility and governance to your agentic ecosystem.