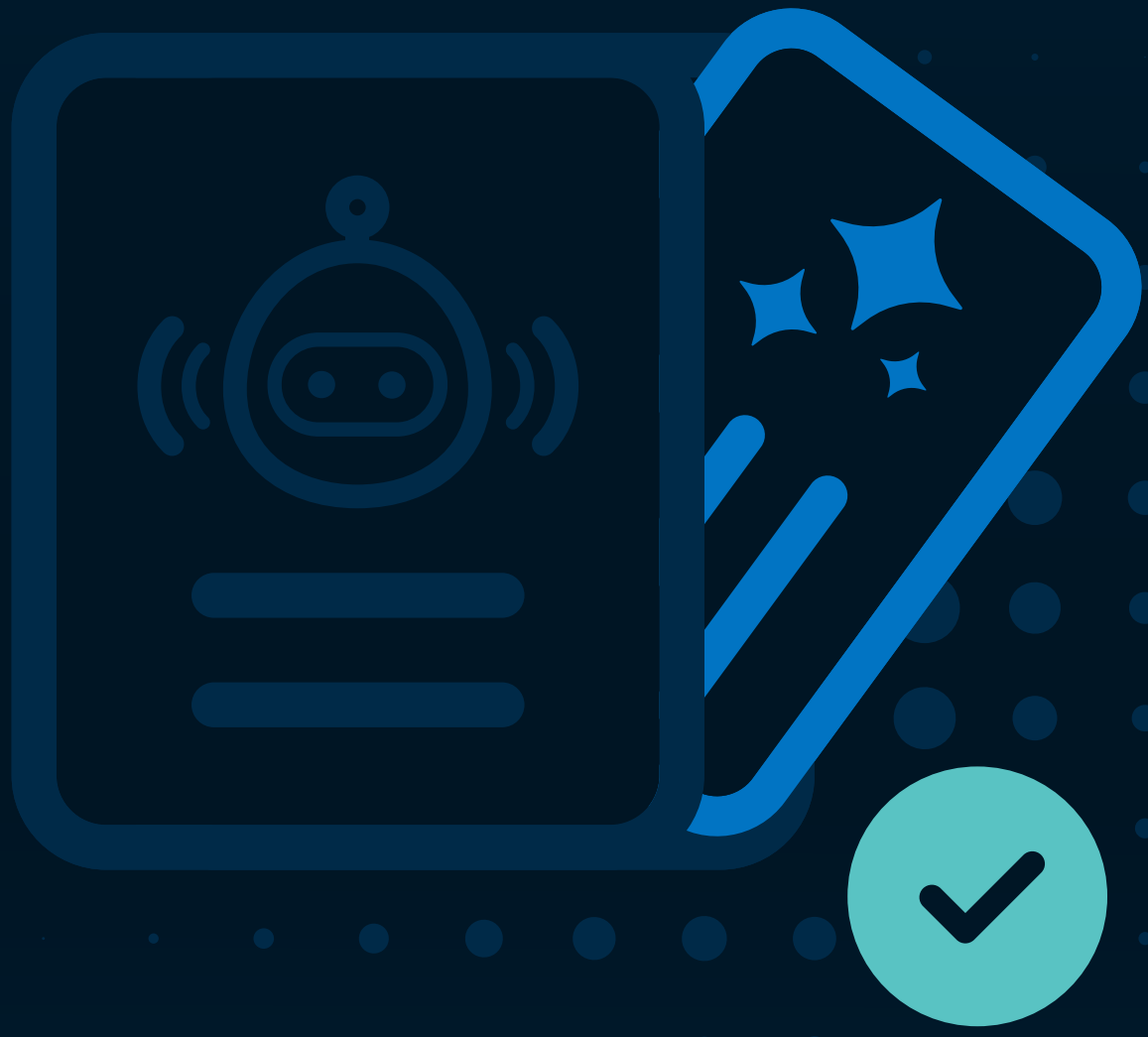
digicert®

How to trust AI Agents

A guide for achieving verifiable identity, authority, and control



Guide | 2026

Table of Contents

Introduction: When agent access grows faster than oversight 3

From answering to acting 5

The expanding trust problem 6

Why existing controls are not enough 8

What a trust layer must do 10

The DigiCert approach 11

Inside the DigiCert AI Passport 13

Trust across clouds and organizations 17

Act now: trust must precede authority 20

Introduction:

When Agent Access Grows Faster Than Oversight

In February, a regional insurer launched a customer-service agent with a deliberately narrow role: answer policy questions. It had read-only access to one policy database, no authority to change customer records, and an accountable owner in the service organization.

By August, its role had expanded. The agent could retrieve claims information, call three internal APIs, and issue refunds below a set threshold. Two other teams had cloned the agent for their own use, and all three versions operated through the same service account. Meanwhile, the original owner had moved to another role, and no one had reviewed the agent's overall authority since February.

No one had formally granted the agent authority to move money through the insurer's entitlement system. That capability emerged from a combination of changed instructions, access to a refund tool, and permissions already held by the shared service account. On paper, the agent still appeared to have a limited role. In practice, its reach was determined by everything its tools and credentials could access.

The problem surfaced when a monthly reconciliation identified several refunds that could not be matched to an approved customer-service case. The insurer tried to determine which agent had initiated them. It could not. At the enforcement point, all three copies appeared under the same identity. The insurer could revoke the shared credential, but doing so would stop all three agents. It could not disable one agent independently or reliably attribute each refund to the agent that initiated it.

There was no evidence that an attacker was involved. Yet the insurer was left with several seemingly simple but vexing questions: Which agent issued the refunds? What was it authorized to do at that moment? Where was that authority defined? Who was accountable for its actions? And could the insurer stop one agent without stopping all three?

Autonomy scales faster than the ability to account for it.

Agentic AI is not a faster version of the tools organizations already govern. They don't just generate content or answer questions. They can access data, invoke tools, make decisions, and take action on behalf of people and organizations. That changes the trust equation.

Before an agent is permitted to act, an organization must be able to answer a few basic questions: what is this agent? Who is responsible for it? What is it allowed to do? And is that authority still valid?

These questions get harder as agents become more connected. An agent rarely operates by itself. It may call one or more models, and reach the outside world through Model Context Protocol (MCP) servers, which are the surface through which tools and data are exposed. An agent with a tightly scoped identity that talks to an ungoverned MCP server has a tightly scoped identity but an unbounded reach.

Put simply: trust must extend across the agent, the models it uses, and the systems that give it access to the outside world.

Existing security platforms detect threats and unsafe behavior. Enterprise identity platforms govern access within an identity environment. Both remain essential. However, agentic AI adds another requirement: portable, cryptographically verifiable trust that can be validated independently, wherever the agent happens to operate.

DigiCert AI Trust Manager provides a practical approach to governing AI Agents built around three capabilities:

Inventory

to identify AI agents and establish ownership

Passport

to provide cryptographically verifiable identity and authority

Control

to govern, revoke and audit trusted interactions

Together, these capabilities help organizations establish trust in agents operating across applications, clouds, and organizational boundaries.

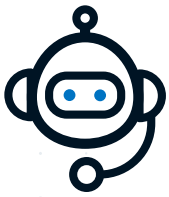
From answering to acting

The first wave of generative AI helped people create content, summarize information, and answer questions. The output was usually something a person reviewed before deciding what happened next.

Agentic AI changes that model. An agent can pursue a goal, decide which steps to take, call applications and tools, exchange information with other agents, and complete work with limited human intervention. In some environments, agents create or coordinate other agents.

This changes what an organization is governing. It is no longer governing only users and applications. It is delegating authority to digital actors that operate at machine speed.

Consider a few examples:



A customer-service agent retrieves account information and initiates a refund (like in the example given earlier)

A development agent writes code and submits it to a software pipeline

A procurement agent compares suppliers, negotiates terms, and places an order

The appeal is obvious. Agents can remove manual steps and get work done faster.

However, the same capability that makes them useful also gives them the power to adversely affect systems, data, money, and customers. That is the core challenge. The goal is not to slow adoption down, but to make sure their authority can be identified, verified, and revoked when necessary.

AI incidents cost more and most enterprises aren't prepared

Teams are deploying agents, models, and MCP servers faster than they can easily identify, govern, and control them. An AI-related breach costs materially more than ordinary ones because AI systems combine access to sensitive data with the ability to act across connected systems. When those systems operate outside established identity, security and monitoring controls, breaches can spread further, take longer to investigate, and require more extensive remediation.

$5.33M	+$630K	78%	40%
average cost of an AI-related breach	more than the average breach	lack an AI identity policy	will face an AI incident by 2030
IBM	IBM	CSA	Gartner

The expanding trust problem

AI agents will enter the enterprise from every direction. Built in-house, embedded in software, adopted without approval, or operated by partners. The result is a constantly changing mix of agents that are either home grown, vendor provided or externally operated that no single team can fully see or manage.

A new shape of shadow AI

Some agents are deployed without a security review. Others begin as sanctioned experiments and accumulate access as their responsibilities grow, which is the pattern in the insurer example. Today, an agent may use shared API keys, long-lived credentials, or the identity of the human who started it. Even when every connection looks legitimate, the enterprise may have no complete view of the agent, its owner, its dependencies, or the authority it has acquired.

Three absences tend to appear together, and each one makes the other two worse. No owner, so there is nobody to ask. No visibility, so there is nothing to review. No kill switch, so there is no way to stop it. Shadow AI in the agentic era is not primarily an employee-behavior problem. It is a visibility and control problem, and it is frequently created by the technical teams themselves.

Autonomy multiplies the impact of a compromise

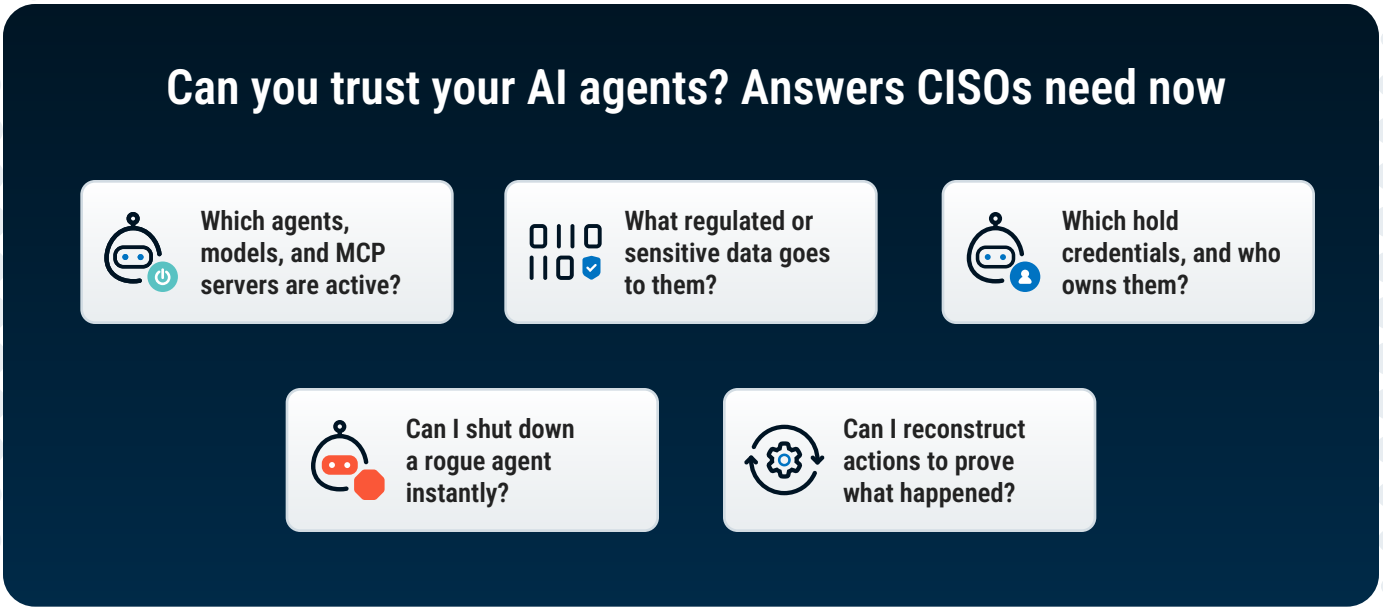
A compromised employee account can cause significant damage. A compromised or unauthorized agent can cause similar damage faster and on a much larger scale because it operates autonomously. It can make repeated decisions, access multiple systems, use tools, and pass work to other agents without waiting for human direction. It may also continue operating after its authority should have ended. Without a verified identity and reliable audit trail, the organization may not know what the agent did, which systems it affected or who was responsible. The result can be greater data exposure, more unauthorized transactions, wider operational disruption, and a longer, more expensive investigation.

Legislation adds to the pressure

Agent identity, authority and control have largely been treated as internal risk-management questions. They are now beginning to appear explicitly in legislative proposals.

The bipartisan Stop Rogue AI Act, introduced in September 2026, would direct NIST to develop standards and best practices for securely deploying AI agents. These would address continuously maintained agent inventories, verification of agent actions, security and reliability assessments, and tamper-resistant activity records. The AI AGENT Act of 2026, introduced as S. 5051 in July 2026, addresses custodial agents acting on behalf of individuals. It proposes verifiable, scope-limited and revocable delegation, together with agent identity verification, real-time revocation and auditable records.

Neither proposal is currently law, and their scope and requirements differ. But the direction is becoming clear: organizations should expect increasing pressure to demonstrate that they can identify an agent, define what it is authorized to do, revoke that authority and produce reliable evidence of its actions. Each of these requires a technical capability, not simply a policy document.



Most organizations can partially answer the questions on which agents, but struggle with the rest. An inventory can show that an agent exists and a policy can describe what it should do. However, trust requires a mechanism that connects the agent to a verifiable identity, binds that identity to defined authority, and lets systems enforce the decision at the moment of interaction.

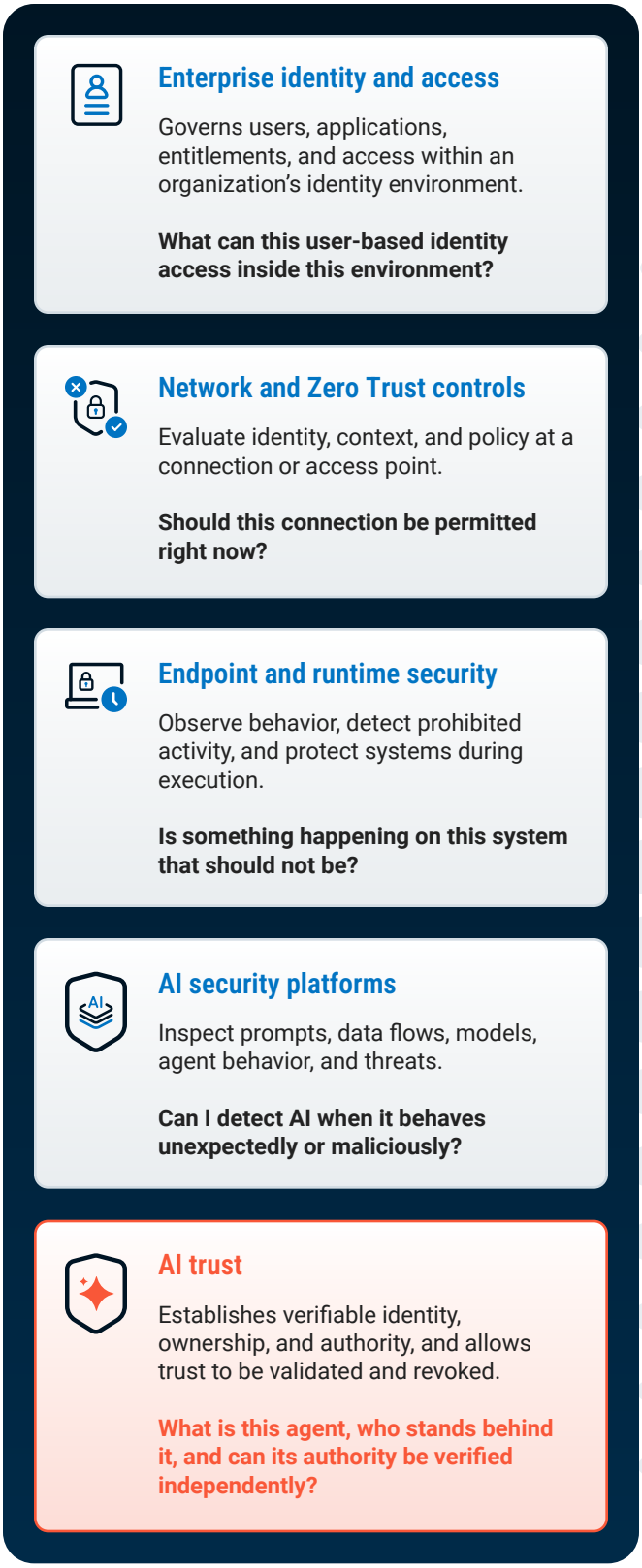
Why existing controls are not enough

Most organizations already have strong identity and security controls in place. Those tools remain essential.

Agentic AI introduces a different challenge. Agents can act across applications, tools, models, and organizational boundaries, which means trust must travel with them wherever they operate.

The figure shows how a layered security approach can build on existing enterprise controls and add the identity, authority, and verification needed for AI agents.

These are not substitutes for one another. Threat detection cannot establish an agent's identity and identity alone cannot determine whether an agent's behavior is safe. The layers are additive, and the AI trust layer is the one most organizations do not have.



Agents are not people. They need identities of their own.

Many organizations still rely on identity systems designed primarily for people, using directory accounts, passwords, multifactor authentication and login-based access controls. That model is not sufficient for autonomous software that can operate without a person present and act across multiple systems.

Modern IAM platforms, including Microsoft Entra and Okta, increasingly support applications and software workloads. They can issue short-lived credentials, apply conditional access and revoke access. These capabilities remain essential, but agentic systems introduce additional requirements:

- Identities that can be independently verified across systems and organizations
- Authority tied to an agent's purpose, task and delegation, and evaluated when it attempts to act
- Continuous verification that the agent's identity and authority remain valid
- Reliable evidence linking each action to the agent, its accountable owner and the authority it held at that time

Modern IAM can provide many of these controls for agents operating within a managed environment. When agents cross clouds, platforms or organizational boundaries, however, their identity, authority and accountable owner must remain independently verifiable wherever they act.

Most AI governance products fall short

Many AI governance tools concentrate on prompt inspection, anomaly detection, and post-execution monitoring. These capabilities are valuable and worth considering, but they observe behavior rather than establish who is acting and what they're allowed to do.

An organization that only observes can describe an incident after the fact. An organization that can establish identity and authority can prevent a class of incidents and prove what the rules were at the time.

Governance is the result, not the starting point

Governance is achieved when these layers work together. A written policy alone is not enough because the system must be able to identify the agent, determine which rules apply, allow or deny the requested action, and record the decision.

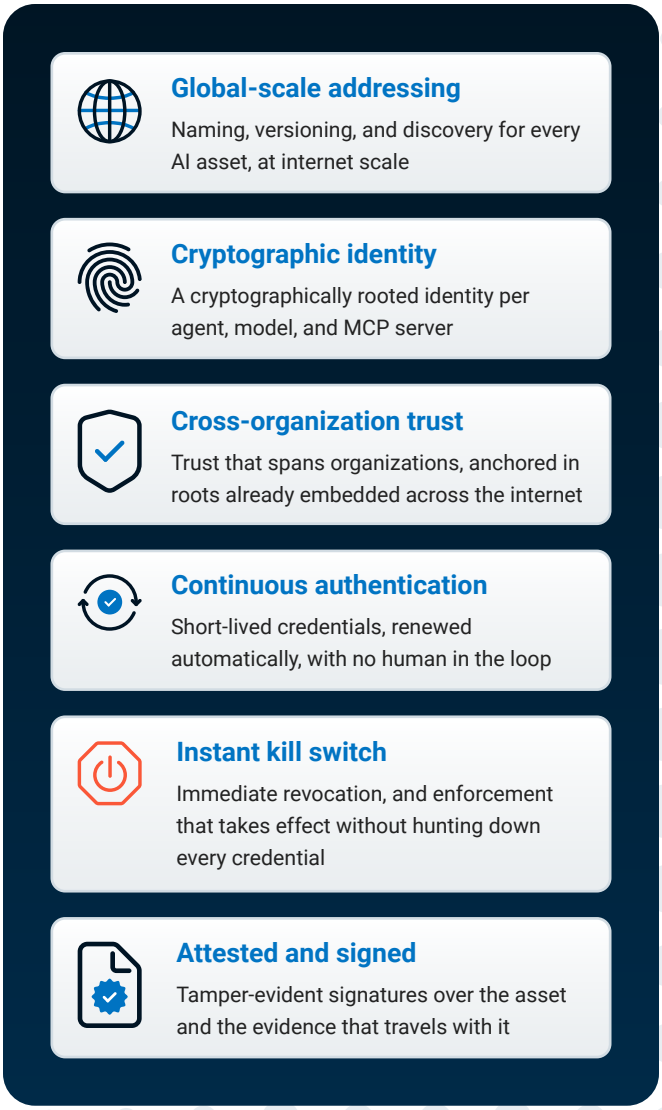
This is also important for compliance. Regulations such as HIPAA and GDPR require organizations to control access to sensitive or personal data and demonstrate that those controls are working. When an agent's verified identity and permissions are checked before access is granted, the system can enforce the policy and record what happened. This provides evidence that the organization's policies are being applied, rather than simply documented.

What a trust layer must do

A useful agent trust architecture must work for agents built internally, agents embedded in third-party services, and interactions that cross clouds, applications, and organizational boundaries. It must support decisions at machine speed without it requiring every participant belong to the same enterprise identity tenant. That last constraint rules out most of what is already deployed. That is why agent identity must be portable across organizational boundaries and anchored in a root of trust that each participant can independently verify.

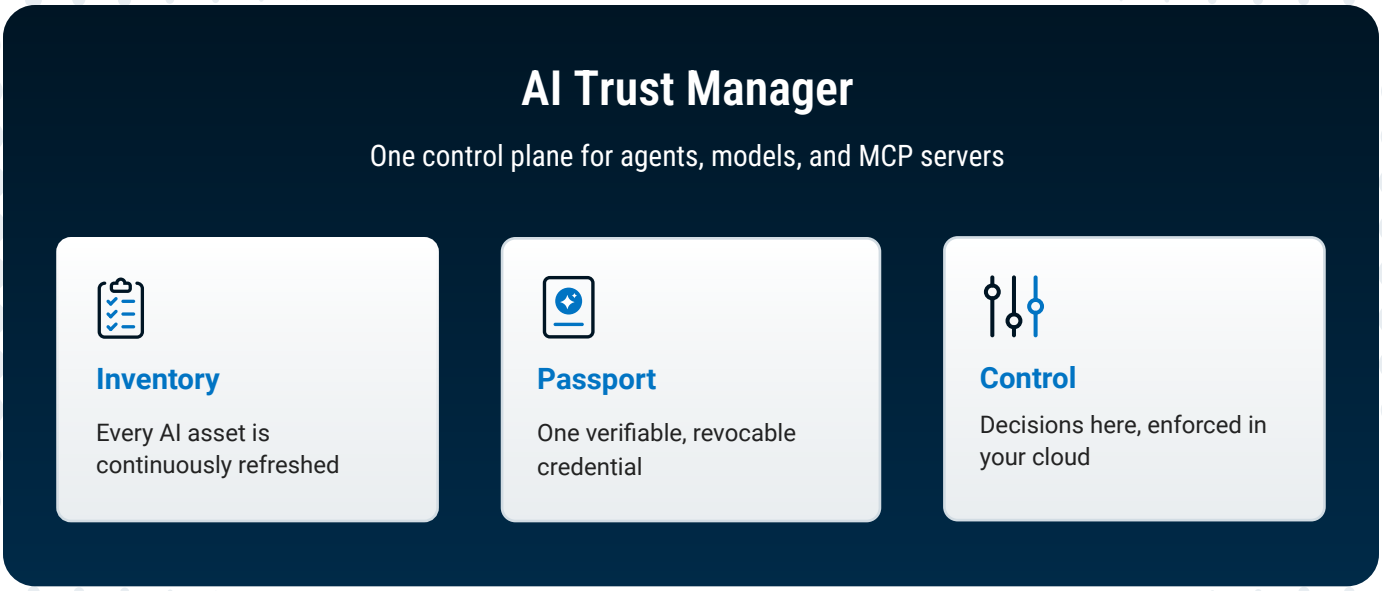
An effective AI trust approach rests on six foundational requirements. Use them to evaluate any solution, whether built in-house or provided by a vendor.

Get The New Trust Architecture for AI to learn more



The DigiCert Approach

DigiCert AI Trust Manager extends the trust infrastructure already used to identify and secure machines and workloads to the new class of software actors created by agentic AI.



What customers ask on their path to trust

Customers usually begin with a visibility question: What agents, models, and MCP servers are running, and who is responsible for them?

Once these assets are known, the next question is whether each one has a verifiable identity and clearly defined authority. Finally, that identity and authority must affect what happens when the agent acts. Policies must be enforced consistently, whether the organization built the agent or subscribes to it as a service. Authority must be revocable, and every decision must produce reliable evidence of what was allowed or denied.

DigiCert addresses these requirements through three connected capabilities: Inventory, Passport and Control.

Inventory: know which assets exist

Before anything can be trusted or governed, an organization must know it exists. Discovery should be continuous rather than periodic because agents are created, modified and retired more frequently than traditional enterprise applications.

AI Trust Manager uses integrations with cloud, CI/CD, IT service management and other systems, together with its built-in discovery capabilities, to identify supported agents, models, and MCP servers across environments the organization can observe and manage. It brings these assets into a centralized private registry. The architecture also supports a public registry through which organizations can make approved assets independently verifiable outside the enterprise.

Not all agentic activity appears as a standalone workload. An employee using agentic capabilities through a browser-based service may operate under the employee's authenticated identity and the permissions granted to that service. In these cases, the control point shifts to the boundary where the service attempts to reach enterprise tools or data. The organization can apply its own policies there, restrict what the service may access and record the resulting actions.

Passport: bind identity to authority

Once an asset is visible and owned, it needs an identity of its own, and that identity must be bound to what the asset is allowed to do. The mechanism is the Agent Passport, and because it is the distinguishing capability in this architecture, it gets its own chapter.

The approach to issuing identity differs by how the agent arrived:

	Agents you build	Agents you buy
Identity	Agents receive short-lived identities that are automatically replaced before they expire, allowing them to verify one another and communicate securely. This capability uses the Secure Production Identity Framework for Everyone (SPIFFE), with identities issued through the SPIFFE Runtime Environment (SPIRE).	A Model Context Protocol (MCP) gateway acts as a checkpoint between the agent and the tools it uses. It presents the agent's Passport and checks every request before allowing it to proceed.
Where it runs	Agents running in your environment receive trusted identities and secure connections. Before they can operate, their digital signatures and declared software components are verified.	Nothing to install. The gateway brokers access to tools, APIs, and internal agents.
What it replaces	Shared API keys and long-lived static secrets.	Unmediated tool and data access from a vendor-operated agent.

Control: enforce trust when the asset acts

Identity and policy create value only when they change a decision. At defined enforcement points, an agent presents its identity and Passport. The credential and trust chain are validated, policy is evaluated, and the request is allowed, quarantined, or refused. The decision and its outcome are recorded.

The default matters as much as the mechanism. Policy is fail-closed: no policy, no access. An asset that has not been brought under governance does not get a permissive default; it gets nothing. And when an agent is compromised, retired, or no longer authorized, revocation removes its authority centrally, so configured enforcement points reject it without anyone hunting down every credential it might have used. In the insurer example, revocation would have covered the two cloned copies as well, because they present the same identity.

Inside the DigiCert AI Passport

This chapter brings the guide's core ideas together. The DigiCert AI Passport (AI Passport) turns identity, authority, and control into a verifiable artifact that makes agent trust practical and extends governance beyond dashboards and policy documents.

Why a passport, and not a permission

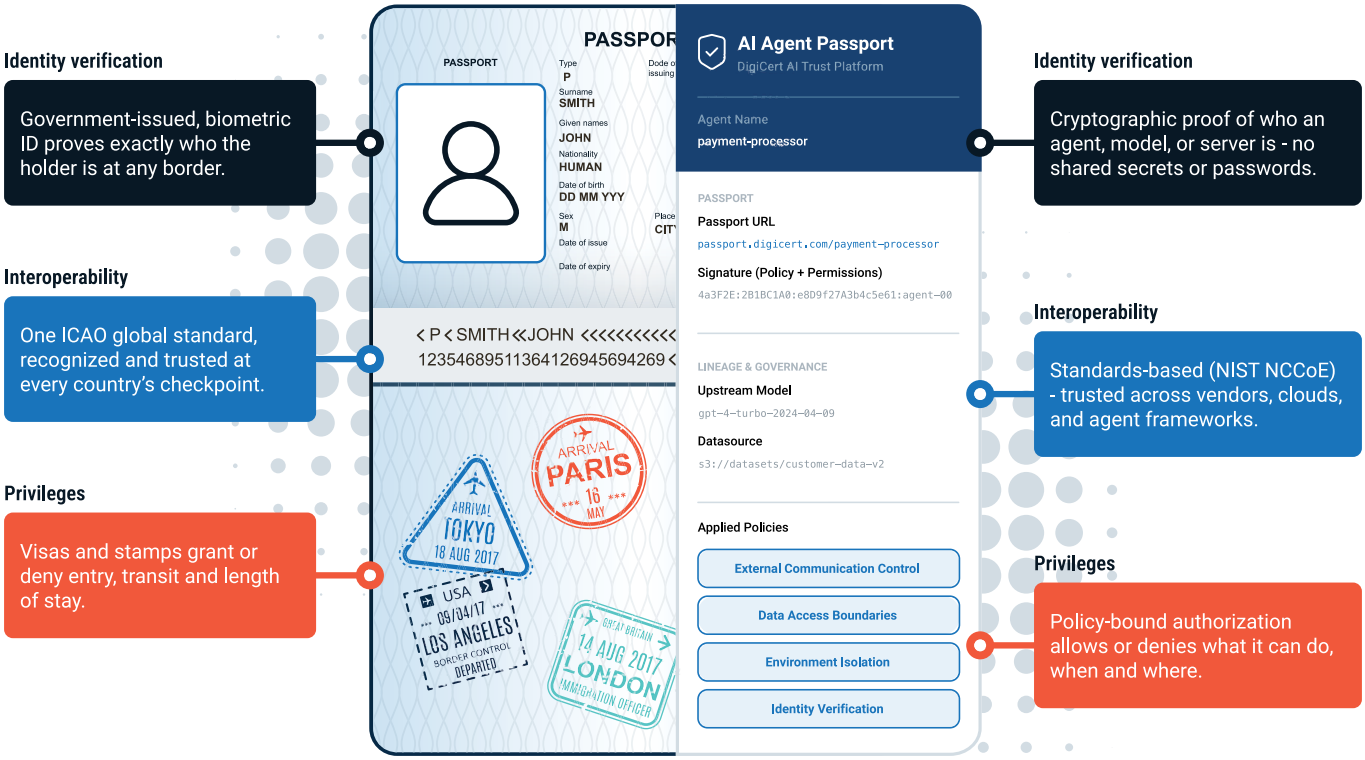
The physical passport is useful as an analogy because it separates proof of identity from the decision to grant access. A government-issued passport proves who a person is and identifies the authority that issued it. Visas and entry rules determine whether that person may enter a particular country, for a particular purpose and period. A border official verifies the information and makes the final decision. Holding a valid passport does not guarantee entry.

The AI Passport works in much the same way. It provides verifiable information about an agent's identity and authority, but it does not automatically approve the agent's request. The receiving system verifies the passport, applies its own policies and decides whether the requested action should be allowed.

Key features and benefits:

- Provides verifiable proof of identity.
- Recognizes and verifies across different systems and environments.
- Carries information about the agent's authority without taking control away from the receiving organization.

Interoperability is especially important. An identity that can only be verified inside the system that issued it, has limited value when agents interact across clouds, platforms and organizations. An AI Passport is designed so that another system can verify who issued it, confirm that its contents have not been changed and determine whether it is still valid.



Four questions a passport helps answer

The diagram explains how an AI Passport works. The information carried by the Passport helps answer four practical questions about the agent.

1. Identity: Who is this agent?

The passport identifies the agent and the trust domain in which its identity was issued. The identity is cryptographically protected rather than relying only on a name or directory entry. Where supported, hardware-based verification can provide additional assurance about the environment in which the agent is running.

2. Authority: What is it permitted to do?

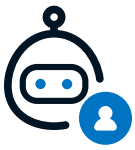
The Passport can describe the agent's approved capabilities and operating limits. These may include which systems or agents it may contact, which actions it may perform, what level of data it may access and whether it may delegate work to another agent.

This information does not grant access. It gives the receiving system verified information that it can evaluate against its own policies.

3. Lineage: Where did it come from?

The Passport can include information about the agent's origins and dependencies. This may include the model or software on which it is based, evidence about the integrity of its components and the chain of agents through which a task was delegated.

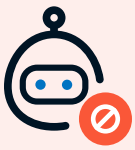
Lineage helps an organization understand how the agent was created, determine whether it was authorized to act on behalf of another agent and reconstruct events during an investigation.



4. Ownership: Who is responsible for it?

The Passport identifies the organization that owns or operates the agent and, where appropriate, the individual or business function accountable for it. It can also reference the governance and data-handling requirements that apply to the agent.

This makes accountability part of the information presented with the agent, rather than leaving it in a separate system that the receiving organization may not be able to access.



What the Passport does not do

The AI Passport does not make the final access decision.

It provides verified information about the agent's identity, authority, lineage and ownership. The receiving system evaluates that information against its own policies and decides whether to allow, limit or reject the requested action.

This separation allows organizations to rely on a shared method of establishing trust while retaining control over their own systems and decisions.

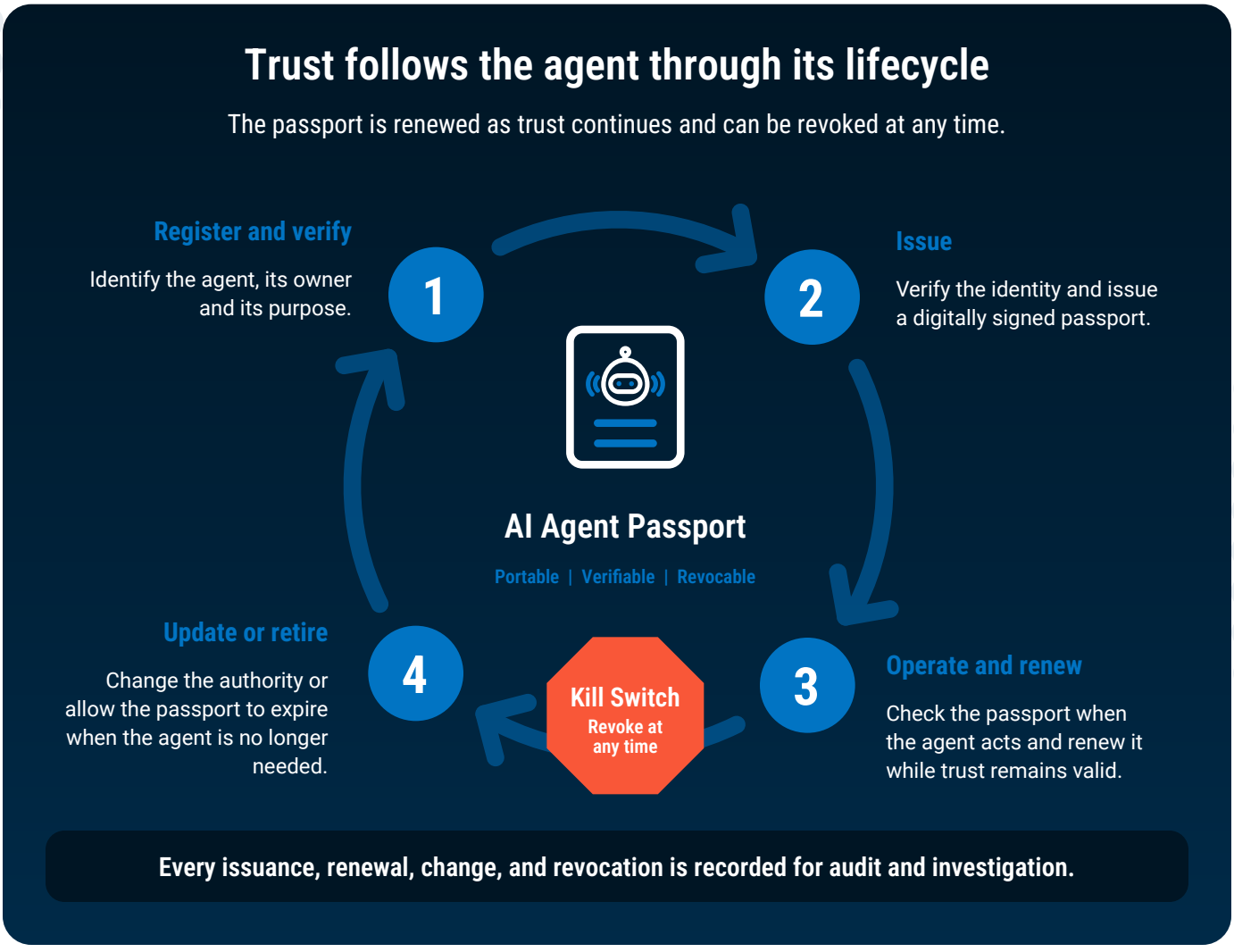
Managing trust throughout the agent lifecycle

Trust in an agent can change. An agent may be modified, given new responsibilities, compromised or retired. Its passport must change with it.

DigiCert's AI Passports can be issued when an agent is approved, renewed while it remains trusted, updated when its authority changes and revoked when it should no longer operate. Because agents act at machine speed, these processes must be automated.

Short-lived credentials reduce the time in which stolen credentials can be misused. Automated renewal allows approved agents to continue operating without relying on long-lived secrets. Revocation allows an organization to withdraw trust when an agent is compromised, violates policy or no longer has a legitimate purpose.

These events also create an audit record. Organizations can determine when a Passport was issued, what authority it represented, when it was used, and when its status changed. This provides the evidence needed to monitor agent activity, investigate incidents, and demonstrate that controls were applied.



Built on established standards

DigiCert's AI Passport builds on established approaches to public key infrastructure, digital signatures and workload identity. It can also incorporate standards-based information about hardware verification, software supply-chain integrity, model provenance, and agent naming and discovery.

Using established standards makes it possible for different systems and organizations to verify an agent without relying on the same vendor, cloud platform, or identity directory. This is important because agentic AI will operate across technology and organizational boundaries.

DigiCert is also contributing to industry work on agent identity and trust.

Trust across clouds and organizations

Enterprise identity and access management are essential for controlling access within an organization. Agentic AI, however, creates a broader challenge. An agent may need to interact with a partner's service, an external MCP server, or another agent running in a different cloud.

The two organizations may use different identity systems, cloud platforms, and security products. Before allowing the interaction, each organization still needs to establish:

What is the agent?

Which organization is responsible for it.

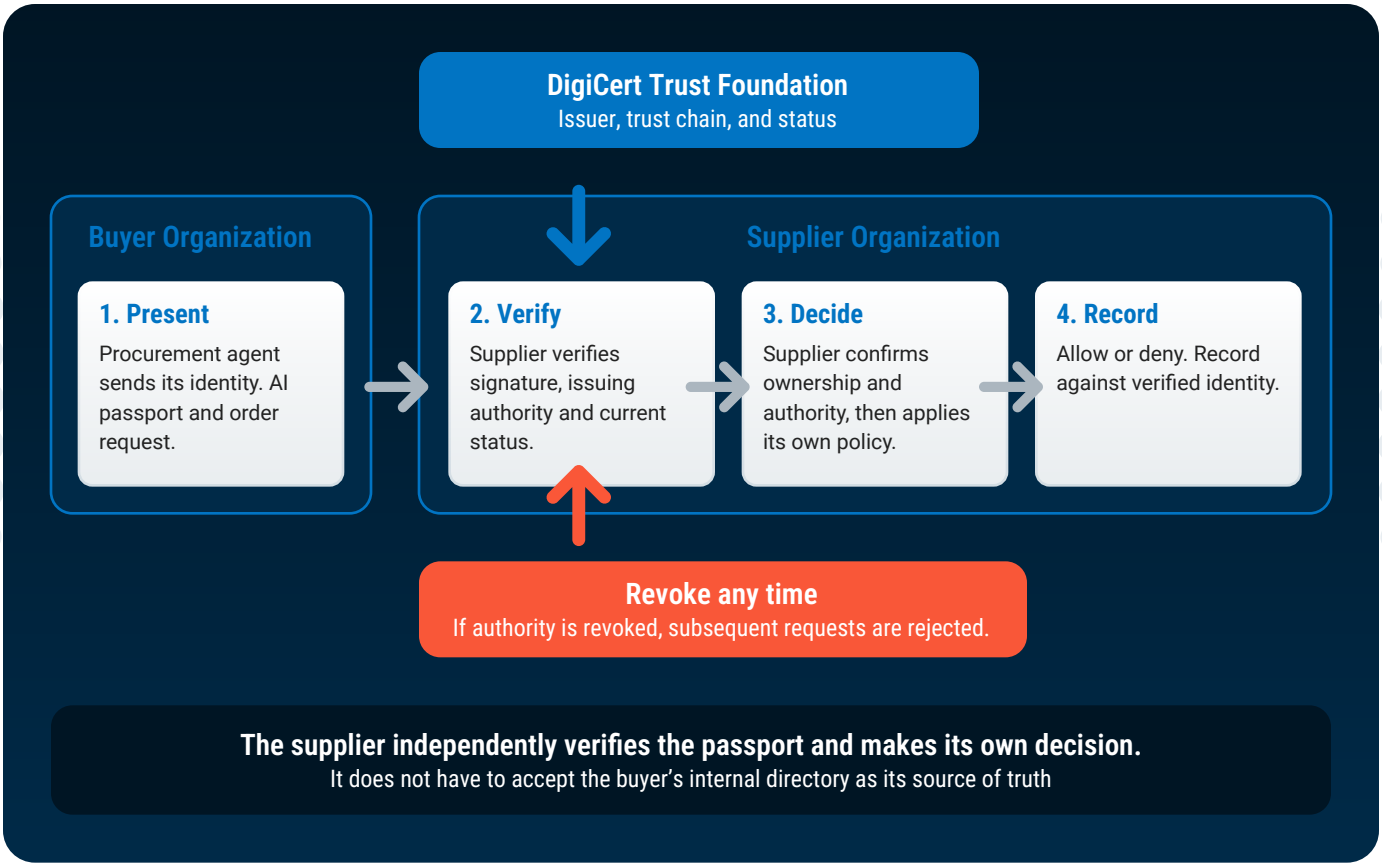
What it is authorized to do.

Whether its identity and authority remain valid.

This requires trust that can extend beyond a single enterprise identity environment.

A procurement agent in action

Consider a buyer agent placing an order through a supplier's agent. The two organizations use different cloud platforms and identity systems. However, the supplier does not have to treat the buyer's internal directory as its own source of truth. It can independently verify the agent's identity and authority, then make its own decision.



A shared basis for verification

Agents interacting across organizations may not share the same cloud, identity directory or security provider. They still need a common way to verify one another before access is granted.

The DigiCert AI Root of Trust provides a managed hierarchy for issuing and verifying the DigiCert AI Passport. A receiving system can confirm who issued the passport, whether it has been altered and whether it remains valid. The receiving organization then applies its own policies to allow or deny the request.

How the passport is presented depends on where the agent runs. An agent operating within an organization's environment can present a passport bound to its workload identity. For an approved SaaS agent, an MCP gateway can assert the passport when the agent attempts to access the organization's tools or data. This allows the organization to control the interaction without installing software inside the SaaS provider's environment.

DigiCert operates the certificate authority and key-management infrastructure, so organizations do not have to build and manage their own. The dedicated AI passport hierarchy builds on DigiCert trust anchors already distributed across operating systems, devices and cloud environments, reducing the need to distribute an entirely new trust anchor.

This does not make passport verification automatic. Each enforcement point must include a verifier that can interpret the passport, validate its trust chain and status, and apply local policy. This gives organizations a shared basis for verification while leaving each one in control of which issuers and agent requests it trusts.

How the security layers work together

The division of responsibility is straightforward:

- Security platforms assess whether an agent or interaction appears dangerous.
- Enterprise identity platforms govern access within a managed identity environment.
- DigiCert AI Trust Manager provides verifiable proof of what the agent is, who is responsible for the agent, and whether its authority remains valid.

Together, these controls allow organizations to interact with external agents without surrendering control over their own access decisions.

Act now: trust must precede authority

Agentic AI is useful because it can act.
That is also what makes trust so important.

As agents gain more responsibility, organizations need to know exactly what each agent is, who is accountable for it, what it is allowed to do, and whether that authority is still valid. The more authority an agent has, the less room there is for ambiguity.

The insurer from the opening pages is a good example. Nothing had been compromised. Every change to the agent had been approved. The problem was that no one could see the full picture anymore. Its access had expanded; its credentials had spread; ownership had changed, and the original boundaries no longer reflected what the agent could do.

That is how a governance problem can become a security problem.

Existing security and identity tools still matter. They detect threats, protect systems, and govern access. Agentic AI adds another requirement: a way to establish trust in an agent itself and carry that trust across the systems where it operates.

That means giving each agent a verifiable identity, clearly defining its authority, enforcing those boundaries, and revoking that authority when conditions change.

DigiCert AI Trust Manager is designed around that model. Through AI Trust Manager in DigiCert ONE, organizations can discover AI assets, issue verifiable passports, define and control authority, and revoke it when needed.

As AI agents become more capable, the question will not simply be whether they can act. It will be whether you can trust them to act within the authority you intended to give them.

Learn more about **AI Trust Manager** or **see it for yourself in action**.

Discover every agent. Give it a verifiable passport.
Control what it is authorized to do.