digicert®

The New Trust Architecture for AI

Securing agents, models, and content with cryptographic proof

White Paper | 2026

Executive summary

Artificial intelligence is reshaping the trust landscape. AI agents execute tasks across systems at machine speed with limited oversight and excessive privilege. AI models deliver powerful outcomes while handling sensitive data and valuable intellectual property, often without transparency or control. At the same time, synthetic media is blurring the line between authentic and manipulated content. Together, these shifts expose gaps in traditional security and governance models, creating new vulnerabilities, and eroding trust.

Today, trust cannot be assumed. It must be cryptographically verifiable across both AI systems and the content they generate. Across agents, models, and content, the core challenge is the same: organizations lack cryptographically verifiable control over what AI systems are, what they are authorized to do, and what they produce.

DigiCert addresses this challenge with a unified approach to AI trust that spans systems and content. Its three-layer defense architecture establishes control from the network edge to the execution core through DNS-based enforcement, standards-based agent identity with the AI Agent Passport, and hardware-rooted model protection via confidential computing. DigiCert also enables verifiable content provenance, allowing organizations to prove origin, integrity, and authenticity through standards-based, independently verifiable mechanisms.

Together, these capabilities define intelligent trust, extending the core principles of digital trust comprising identity, integrity, and secure communication to AI systems and their outputs.

From digital trust to intelligent trust

Historically, digital trust has focused on users, devices, software, and servers. Public key infrastructure (PKI), certificates, and encryption have been used to secure communication and authenticate endpoints, enabling trusted digital interactions across the internet.

AI requires the evolution of digital trust. Modern organizations must establish ways to verify:

- The identity and authorization of autonomous agents
- The integrity and provenance of AI models
- The authenticity of digital content

Meeting these requirements demands a more comprehensive approach: intelligent trust, a framework for establishing cryptographically verifiable identity, authorization, and integrity across AI agents, models, and the content they produce.

DigiCert's approach builds on its experience in digital trust to support this shift, applying established cryptographic practices to help organizations secure and scale AI across environments and use cases.

The AI governance crisis

Most organizations are already running AI. The real question is whether leaders know what AI is being used, where it's running, and on whose authority. This lack of oversight is amplified by the rise of shadow AI: the unapproved and untracked use of AI tools and agents outside formal governance.

Traditional identity and access management systems were built for human users and deterministic software. They are not equipped to manage probabilistic agents that can spawn sub-agents, interface with external tools, and operate across trust boundaries without direct human oversight.

98%

of orgs have employees using unsanctioned AI

Netskope

$4.2 M

average cost of a shadow AI data breach

SQ Magazine, 2026

40%+

agentic AI projects at risk of cancellation by 2027

PwC, 2025

5 questions CISOs struggle to answer

The shadow AI problem crystallizes around five unanswered questions that expose the governance gap in every enterprise:

1. What AI agents are employees using?
2. What regulated data, such as PII, PHI, or intellectual property, is flowing to those agents?
3. Which agents hold credentials, and whose?
4. Can a compromised or misbehaving agent be stopped immediately?
5. If an incident occurs, can the organization reconstruct a tamper-evident trail for regulators?

These gaps point to a broader issue: Traditional governance, security, and risk controls weren't designed for autonomous, probabilistic systems operating at machine scale. As a result, many agentic AI initiatives stall or fail.

DigiCert's thesis is that PKI, DNS, attestation, and cryptographic identity—disciplines proven at internet scale over three decades—are the natural foundation for securing AI.

> ***PKI, DNS, and cryptographic identity are the natural foundation for securing AI.***

The 3-layer defense architecture:

A unified control plane for AI trust

DigiCert ONE unifies three layers of defense, spanning from the network edge to the execution core, under a single control plane and policy engine with a unified kill switch that can terminate activity at any layer and propagate across the entire system.

Layer 3 | Network Edge: DNS-based agent trust

Agent trust begins at the network edge. Because every AI agent action starts with a DNS query, it becomes the only universal control point where policy can be enforced consistently, regardless of how or where the agent is deployed.

Whether resolving an API endpoint, connecting to an MCP server, or downloading a package, DNS is the universal gateway through which all network communication must pass. This allows for the creation of a network-edge enforcement layer that operates before any connection is established. This makes DNS a natural choke point to prevent unauthorized actions before they occur.

> ***"Every agent must make a DNS call. If you're able to tell me that an agent made a call to a domain that it should never have done, you immediately tell the MCP to kill that action."***
>
> **— CISO, $8B industry-leading company**

Extending domain-based trust to AI

DigiCert's approach follows a pattern established by DMARC for email authentication: domain-anchored identity verification, policy lookup, gateway enforcement, and automated response.

Similarly, we tie agent identity and authority to domain ownership. An organization publishes an agent policy record declaring its agent identities, the certificate authority (CA) that issued their credentials, and permitted scopes. When an agent presents credentials, the gateway queries DNS, verifies the credential, and either admits or blocks the agent.

From email trust to agent trust

DMARC (Email today)

- Authorized mail servers published via DNS TXT record (SPF)
- Email signed with keys referenced in DNS (DKIM)
- Policy record instructs receivers to reject messages that fail verification

Domain-based trust for AI agents

- Authorized agent identities and issuing CAs published in DNS record
- Permitted scopes defined in policy records
- Gateways enforce policy by allowing or terminating sessions that fail credential or scope verification

```
$ dig TXT _agentpolicy.ibm.com

"v=AGENTDMARC; ca=digicert.com; scopes=read:*,write:jira; p=reject"
```

Example: Agent policy TXT record

Enforcement capabilities

This DNS-based enforcement layer enables:

- Allowlist and blocklist control for agent communications
- Real-time visibility into outbound DNS queries from agents and MCP toolchains
- Pre-connection blocking of unauthorized domains
- Automated termination (kill switch) of sessions when anomalous patterns are detected

Because enforcement occurs at this universal control point, organizations gain immediate, system-wide leverage over agent behavior without requiring changes to endpoints or application code.

Real-world scenario: Claude Code installs npm packages

Scenario A: Legitimate package

A developer's agent runs npm install lodash. The agent resolves registry.npmjs.org, which is allowlisted. The DNS query is permitted, and the MCP gateway allows the action to proceed.

Scenario B: Typosquatted package

The agent is tricked (via prompt injection or hallucination) to run npm install lodash in a malicious registry. The domain isn't on allowlist. DNS blocks the query before a connection is established, and the kill switch terminates the session. Alert is logged.

Layer 2 | AI Agent Identity: Passports & policy

AI agents are increasingly embedded in enterprise workflows. They interact with APIs, retrieve data, trigger processes, and make decisions autonomously. As adoption accelerates, agents are reshaping both how work is performed and the composition of the workforce.

Yet even as this new class of "digital workers" operates at machine speed and across distributed systems, many still rely on shared API keys, long-lived credentials, and overprivileged access. Accountability is often unclear, lineage tracking is limited, agent discovery is murky, and shadow AI proliferates.

Organizations must extend trust to AI agents without introducing new visibility and security gaps.

Agent identity fundamentals

Many organizations lack a clear inventory of the AI agents operating across their environment. Before agents can be secured or governed, they must first be discovered and assigned identity. This is best achieved by applying existing workload identity principles.

Why workload identity rather than traditional identity and access management (IAM)? Although agents differ from more traditional software workloads with their non-deterministic nature and high degree of autonomy, they are not human users, and the identity concepts that were built for human users such as passwords and multifactor authentication (MFA) do not apply.

Agents need runtime attestation and ephemeral credentials, not login screens and passwords. The consensus is growing among industry bodies such as NIST and IETF that workload identity principles offer the best approach when it comes to effectively identifying and authorizing agents.

> ***AI agents require runtime attestation and ephemeral credentials, not the login screens and passwords used to secure human users.***

SPIFFE and SPIRE: Foundations of workload identity

Secure Production Identity Framework for Everyone (SPIFFE) is an open standard that defines how workloads securely identify themselves across distributed systems. Each workload is assigned a unique SPIFFE ID, a URI-based identifier (e.g., spiffe://example.com/ai-agent), which enables secure authentication and authorization between services.

SPIFFE Runtime Environment (SPIRE) is the reference implementation of the SPIFFE standard, operating as a certificate authority within a trust domain to issue identities to workloads. It continuously attests and verifies workloads and issues short-lived credentials called SPIFFE Verifiable Identity Documents (SVIDs), typically in the form of certificates for secure authentication and authorization.

Governing both built and third-party agents

Organizations operate in heterogeneous environments. They build agents on platforms such as Vertex AI, Amazon Bedrock, and Kubernetes, while also adopting third-party agents such as Microsoft Copilot, Salesforce Agentforce, or ServiceNow. The challenge is applying consistent identity, policy, and audit controls across both.

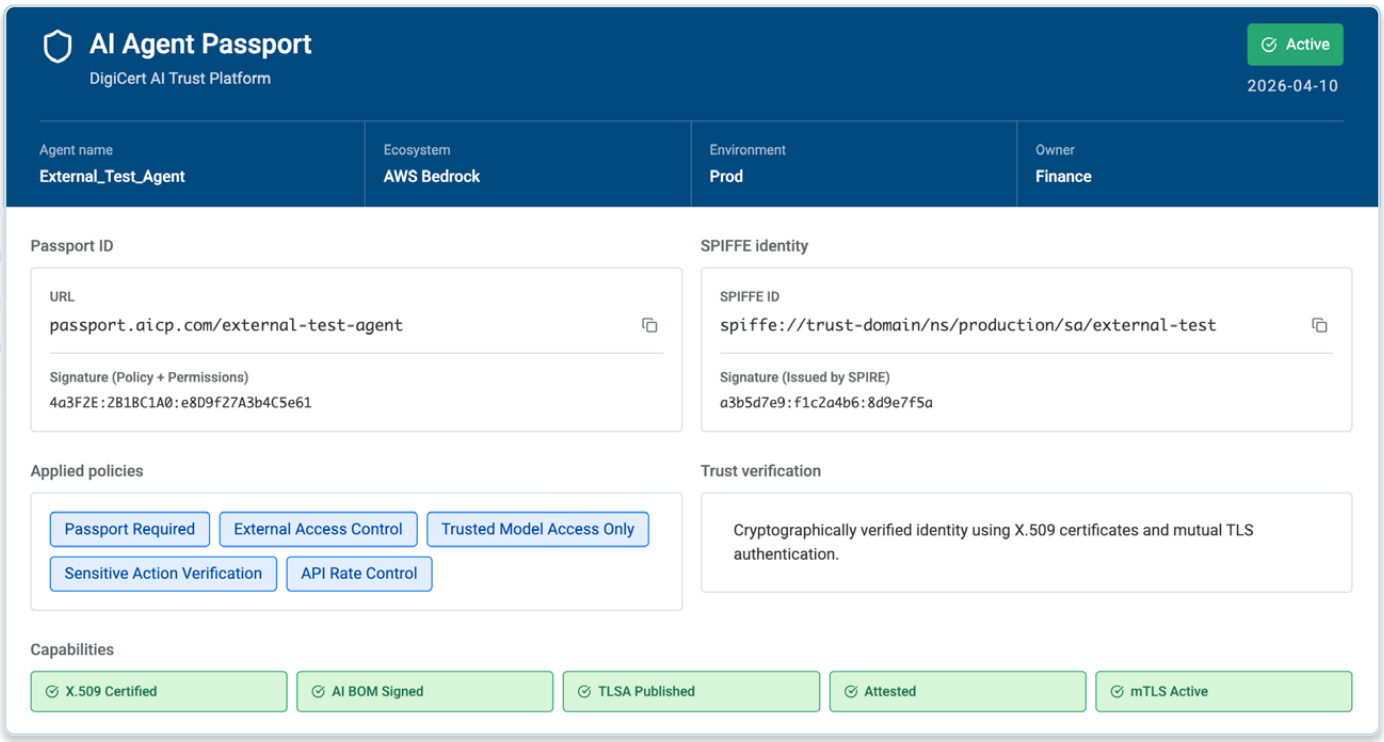
For agents you build

DigiCert issues Internally developed agents with workload identities based on SPIFFE and SPIRE. These identities use short-lived, with automatically rotated credentials, and are continuously validated through cryptographic attestation. This enables strong, portable identity without reliance on shared secrets or network boundaries.

For 3rd party agents

DigiCert embeds a Model Context Protocol (MCP) gateway within a third-party agent's toolchain to control access to tools, APIs, and internal agents. The gateway enforces Open Policy Agent (OPA) policies and issues short-lived credentials for each MCP-mediated connection. The external agent's identity is verified using a domain-anchored passport, validated through DNS TXT records.

Both lanes converge on a shared control plane: Federated identity is established through a single SPIRE server, anchored by a DigiCert CA, with policy enforced centrally using an OPA policy engine. A unified kill switch operates across both agent categories, triggered automatically or from a central dashboard.



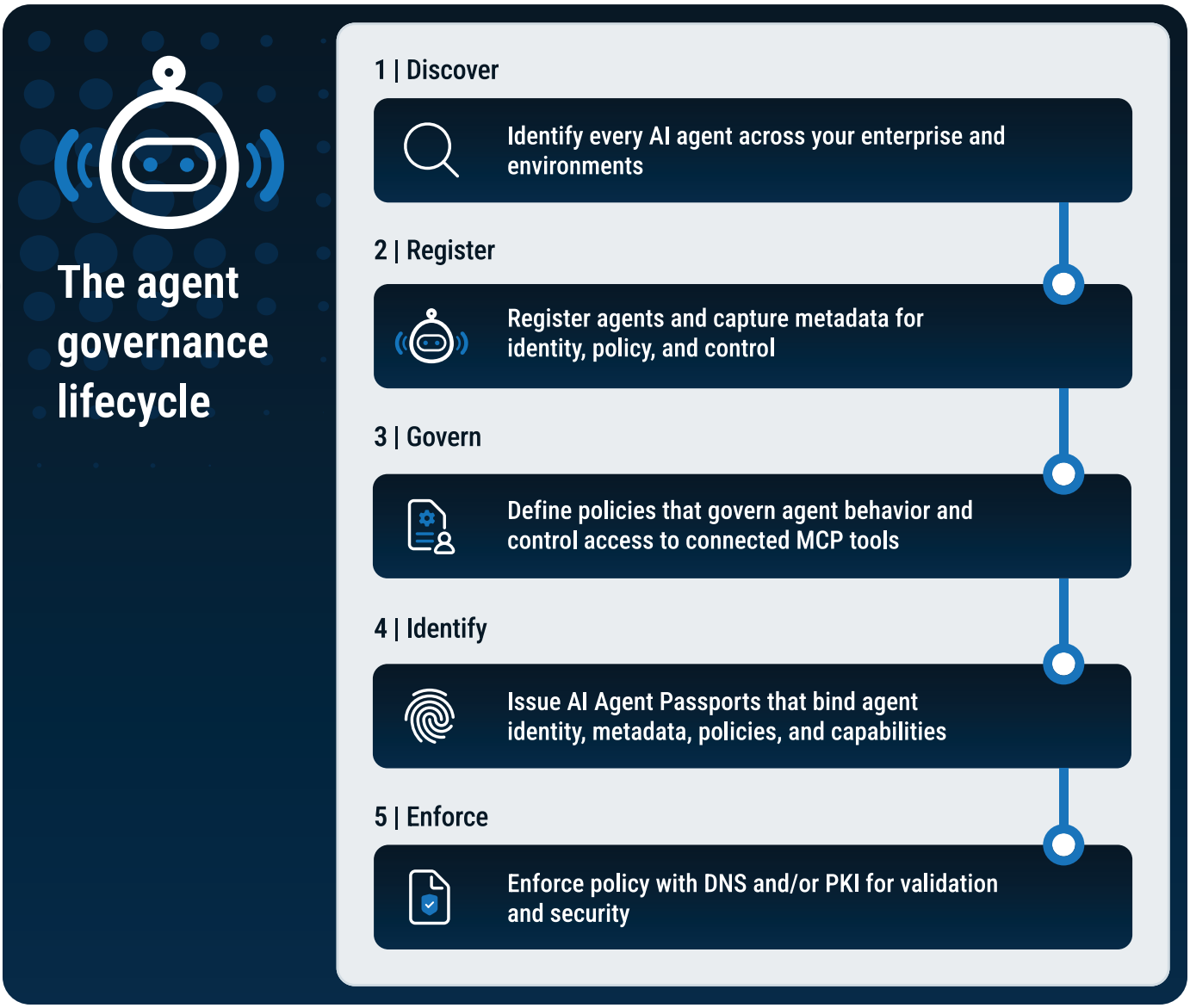
The AI Agent Passport for authorization

Identity alone is not sufficient; agents also require scoped authorization. Organizations must define where agents can operate, what actions they can perform, what data they can access, and who is accountable for their behavior.

To meet these requirements, DigiCert has introduced the AI Agent Passport, a tamper-evident artifact that cryptographically binds to an agent's workload identity and encodes key information such as:

- Approved systems and connections
- Permitted operations
- Authorized environments
- Data sensitivity classifications
- Expiration and renewal states
- Accountable human ownership
- Custom classifications based on internal requirements

Much like a passport that proves identity during travel, an AI Agent Passport serves as both proof of identity and authorization to operate across digital environments. Similar to visas and stamps, each action is recorded along with identity, permissions, and outcomes, creating a secure and tamper-evident audit trail.



Lifecycle trust and control

Establishing identity at deployment is only the beginning. Trust must persist across the full lifecycle, including issuance, renewal, expiration, and revocation.

Short-lived credentials and time-bound permissions reduce the risk of standing access. Renewal processes ensure that authorization is continuously revalidated. If an agent is compromised or no longer authorized, its identity can be revoked, immediately removing access across connected systems.

Layer 1 | Execution Core: Secure & confidential execution

AI models as critical infrastructure

AI models have become strategic assets. They power diagnostics, automate compliance, detect fraud, and influence high-impact decisions. For model creators, they represent valuable intellectual property. For users, they introduce operational risk. Both face uncertainty and evolving threats even as these systems drive innovation.

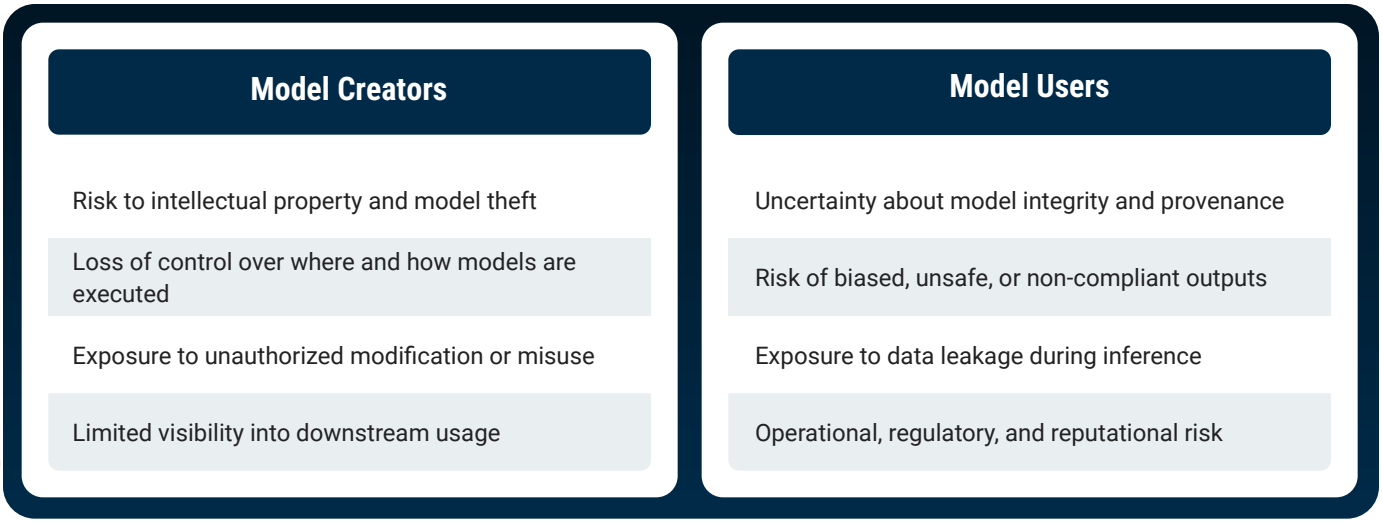
The model risk landscape

AI models introduce a new class of systemic risk by combining software supply chains, sensitive data processing, and high-impact decision-making into a single artifact. Unlike traditional applications, models are dynamic, data-dependent, and often distributed across organizational and cloud boundaries. This creates multiple points of exposure throughout the lifecycle:

- **Model tampering or unauthorized modification:** Malicious actors may alter weights, inject backdoors, or bias outputs.
- **Intellectual property theft:** Proprietary models may be extracted, reverse-engineered, or redistributed outside authorized channels.
- **Supply chain compromise:** Dependencies such as open-source libraries, pre-trained weights, or datasets may introduce hidden vulnerabilities.
- **Runtime manipulation:** Execution in untrusted environments may expose models to side-channel or integrity attacks.
- **Data leakage during inference:** Sensitive input data may be exposed in insecure execution environments.
- **Model drift and integrity degradation:** Models may diverge from approved baselines without continuous validation.

Model risk: two sides of the same coin

AI model risk affects both those who build models and those who deploy them, creating a shared trust gap across the ecosystem.



Without verifiable integrity and execution assurance, neither model makers nor model users can fully trust the system.

What's required: End-to-end model trust

Addressing these risks requires more than version control or internal policy. It requires end-to-end model trust: cryptographically verifiable integrity from packaging to deployment, hardware-backed assurance during execution, and lifecycle governance across the AI supply chain.

Building verifiable model integrity

This approach enables cryptographic proof that a model is untampered, traceable, and authorized for use. Verifiable model integrity ensures that each stage of the lifecycle is anchored in strong identity, tamper-evident controls, and independently provable provenance.

Ensuring model integrity requires:

- Model encryption and cryptographic signing (e.g., SHA-256 hashing with verifiable chain of custody)
- OCI-compliant model packaging with CLI-based signing tools (Sigstore and Cosign)
- A cryptographically verifiable Model Bill of Materials (MLBOM) documenting weights, datasets, dependencies, and packaging components
- Version hashing and integrity verification at each distribution point
- Controlled distribution workflows with scoped API access enforced at runtime

These practices extend software supply chain security principles to AI systems. Organizations gain transparent lineage, tamper detection, and stronger governance across the model lifecycle.

Runtime trust and confidential execution

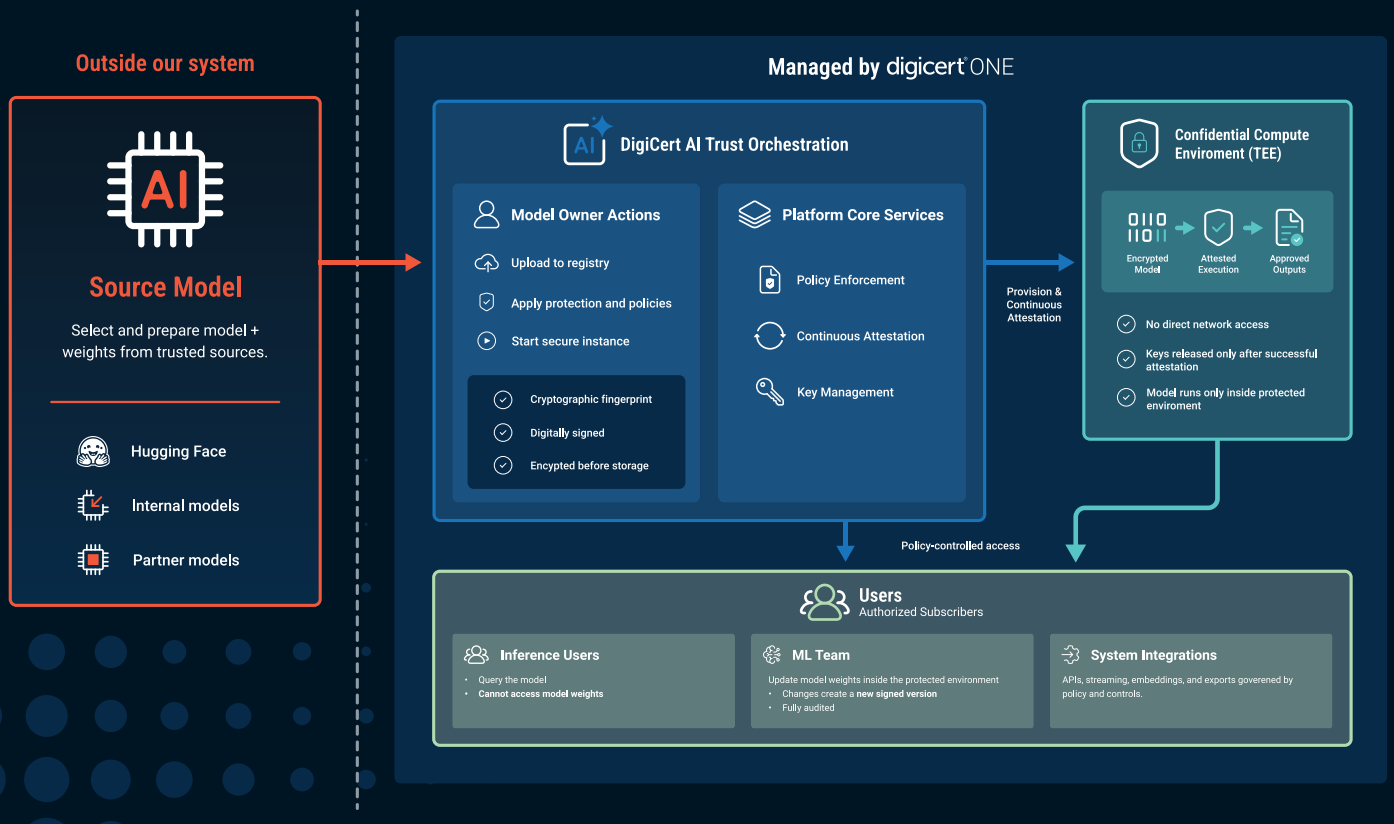
Models must be protected at runtime to ensure both integrity and confidentiality. Model creators often lack visibility and control once their models are deployed, while model users face risks such as data exposure, model drift, and unintended behavior, particularly in cloud environments where infrastructure is not fully trusted.

Trusted execution environments and confidential computing address these challenges. When models run inside a Trusted Execution Environment (TEE), they remain encrypted in memory, are isolated from the host operating system, and are protected from external access, including the cloud provider.

This ensures:

- Data is protected in transit, at rest, and in use
- Execution occurs only in verified, trusted environments
- Sensitive information remains inaccessible to unauthorized parties
- Runtime integrity is continuously validated
- Access is tightly controlled

Deploying and Protecting AI Models



Adding continuous monitoring and lifecycle controls ensures that events such as registration, attestation, key release, fine-tuning, and retirement are recorded, creating a verifiable audit trail. This approach aligns with emerging frameworks such as IETF Supply Chain Integrity, Transparency, and Trust (SCITT).

Hardware-rooted trust and attestation

DigiCert supports TEEs across major platforms, including Intel TDX for cloud-based confidential VMs, AMD SEV-SNP for memory encryption with integrity guarantees, and TPM 2.0/vTPM for device-level and virtualized attestation.

These hardware roots of trust follow the IETF Remote ATtestation ProcedureS (RATS) architecture defined in RFC 9334, where:

- The execution environment acts as the Attester
- DigiCert's verification service acts as the Verifier
- Downstream systems act as the Relying Party

This model enables independent verification of execution integrity across environments.

DigiCert attestation signing service

DigiCert operates a confidential computing attestation service for major cloud providers. Crucially, attestation comes from a neutral third party rather than the cloud operator itself, ensuring independence that is essential for regulatory and contractual trust.

Collect evidence

Workload collects attestation evidence from TDX hardware via go-tpm-tools.

Submit

Evidence sent to DigiCert's verification service via REST API.

Verify

Certificate chain and measurements cross-referenced against firmware endorsements.

Sign

Signed attestation token returned for use or publication.

Tamper-evident audit trail

Every lifecycle event—registration, attestation, key release, fine-tuning, retirement—is recorded as an immutable, cryptographically signed entry in a transparency log (e.g., Sigstore Rekor). This enables independent verification of model integrity and execution history and aligns with emerging standards such as IETF SCITT.

Attestation directly addresses the FDA's 2023 cybersecurity guidance requirements for software integrity verification, SBOM enforcement, and coordinated vulnerability disclosure.

Customer scenario: Medical device manufacturer

Challenge

A medical device manufacturer deploys AI diagnostic models on MRI machines in hospitals worldwide. The FDA requires that the algorithm cleared through the 510(k) or PMA pathway is the exact algorithm running in clinical deployment. But today, there's no technical mechanism to enforce this—it's purely procedural. Meanwhile, hospitals need assurance that patient scan data (PHI under HIPAA) is not exfiltrated, and the AI model vendor needs protection of their proprietary model weights and architecture.

Solution

DigiCert's solution seals the AI model decryption key to the device's TPM state. The attestation agent verifies model integrity before every inference by comparing the running model's hash against the DigiCert-signed manifest of the FDA-cleared version. Signed attestation certificates provide cryptographically verifiable proof that the cleared model is the running model. If the model is tampered with or the execution environment is compromised, attestation fails, sealed keys cannot be released, and the model will not run.

Expected Outcome

- Cryptographically verifiable proof of model integrity for FDA audit trail
- PHI data isolation via hardware-backed enclaves
- Continuous compliance evidence for hospital HIPAA teams

Trust in content

AI is rapidly becoming a primary producer of digital content. Images, videos, text, and structured outputs are now generated at machine speed and global scale. Digital content is the primary means by which a business presents itself to the world. It underpins how a firm operates, communicates, and earns trust, and it ranks among the most vital assets any organization possesses.

Yet despite the acceleration in content creation, a fundamental problem remains unresolved: once content leaves the system that created it, there's no reliable way to prove where it came from, whether it's been altered, or how it has evolved over time. Traditional trust signals such as branding, platform reputation, or metadata are no longer sufficient in an environment where synthetic content can be generated, modified, and redistributed with ease.

This creates a new class of enterprise risk. Legitimate AI-generated content can be misattributed or manipulated, while malicious content can be made to appear authentic. The result is a breakdown in trust—not just in suspicious content, but in all digital interactions.

Impact of malicious content on organizations

Insurance

Fraud

Media

Misinformation

Healthcare

Tampered diagnostics

Government

Forged records

Retail

Fake products

Legal

Altered evidence

Why transparency is not enough

Most approaches to content authenticity focus on transparency. Metadata fields, visual indicators, and platform-level signals attempt to provide context about how content was created. But these approaches are inherently fragile. Metadata can be stripped or altered, platform signals don't persist across ecosystems, and self-asserted claims can't be independently verified.

As AI-generated content scales, trust can't rely on visibility alone. It must be grounded in cryptographic proof—proof that's bound to the content, persists across systems, and can be validated by any recipient without relying on a central authority.

> ***Visibility is not trust. Without cryptographic proof, authenticity remains an assertion, not a guarantee.***

The approach needed for trusting content

Establishing trust in AI-generated content requires a shift from implicit trust to verifiable authenticity. This means embedding proof directly into the content and ensuring that the proof travels with it.

A solution for content trust must provide four core capabilities: a verifiable link to the identity of the creator, tamper-evident integrity to detect unauthorized changes, a complete record of provenance that captures how content has evolved, and the ability for any third party to independently verify these claims.

Identity

Who created it

Integrity

Has it changed

Provenance

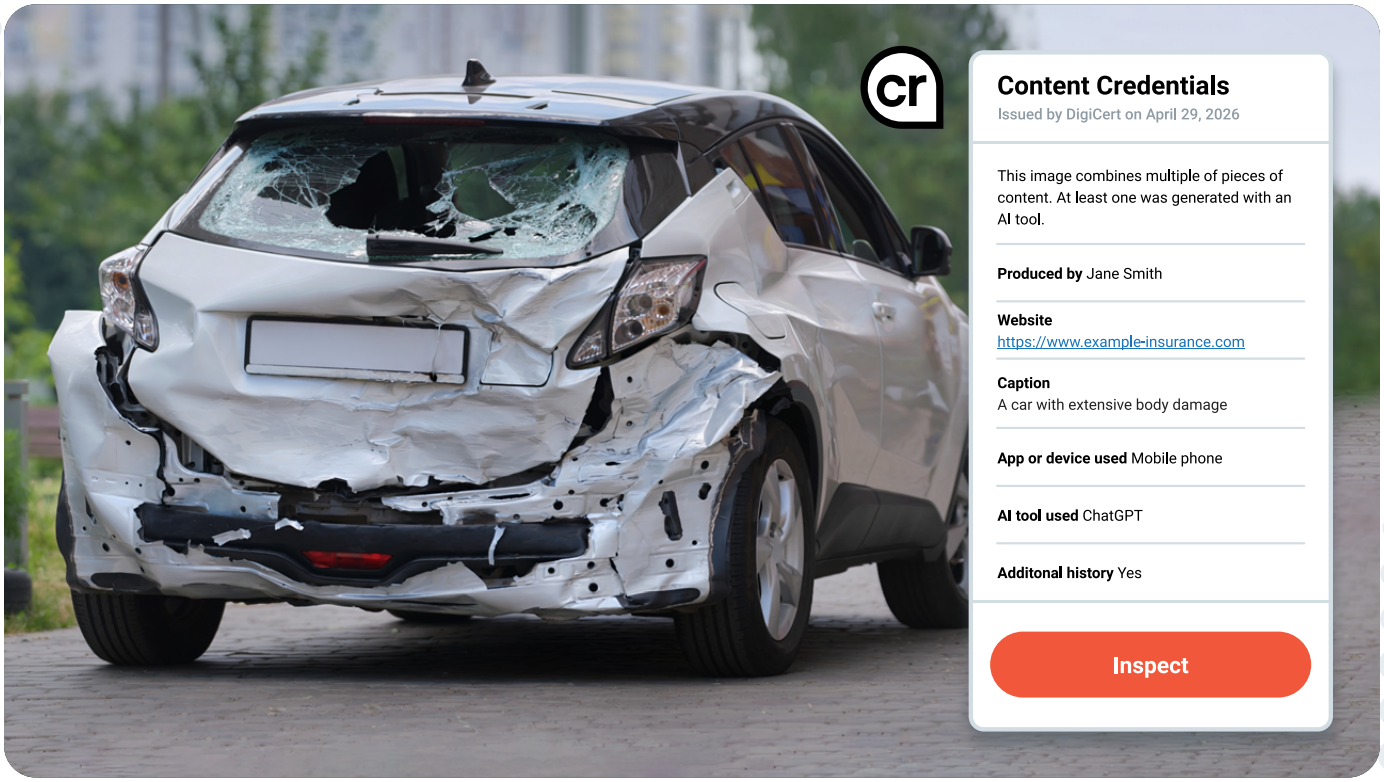
What happened

Verification

Can it be validated

Built on the C2PA standard

The Coalition for Content Provenance and Authenticity (C2PA) defines an open standard for attaching verifiable information to digital content. It uses a cryptographically signed “content credential” to record details like who created the content, what tools were used, when it was created, and any subsequent modifications.



While C2PA provides the structure for expressing provenance, it does not establish trust on its own. Trust depends on the identity behind the signature, the security of the signing process, and the ability to validate those claims through independent, trusted infrastructure.

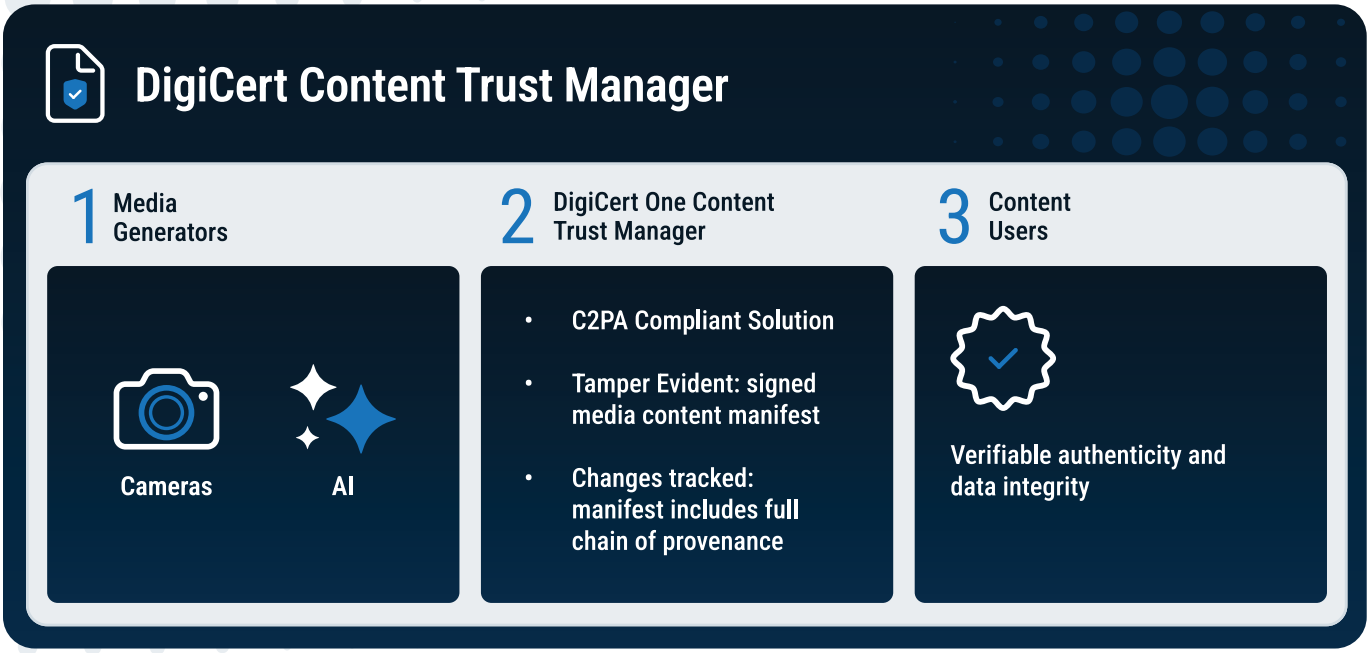
DigiCert ONE's trust infrastructure for authentic content

DigiCert complements this approach by providing the trust infrastructure required to make content authenticity verifiable at scale. With DigiCert ONE, organizations can embed cryptographic proof directly into AI-generated content as part of their creation and publishing workflows.

Each piece of content is signed using a certificate issued by DigiCert, binding it to a verified organizational identity. This ensures that authenticity isn't based on self-assertion but on independently validated identity. Time-stamping services provide an additional layer of assurance by recording when content was created or signed, supporting auditability and regulatory requirements.

Because DigiCert operates as a neutral, third-party certificate authority aligned with emerging standards such as C2PA, content signed through this infrastructure can be validated across platforms and ecosystems. This allows trust to persist beyond the boundaries of any single application or distribution channel.

In practice, this framework integrates directly into the lifecycle of AI-generated content. Content is generated by an AI system, relevant metadata is captured, and a content credential is created. DigiCert then signs the credential, embedding cryptographic proof into the asset. As the content is distributed, this proof travels with it, enabling recipients to independently verify its authenticity and integrity.



Content Trust in action with DigiCert ONE

Customer scenario: Global news organization

Challenge	Solution	Expected Outcome
A leading global media organization publishes breaking news, video, and images on a massive scale. AI-generated and manipulated content has made it harder to distinguish real reporting from fake. The company recognizes that bad actors can create synthetic content and attribute it to the organization, undermining trust. To reduce the growing risk, the company is actively exploring ways to protect its content without slowing down its newsroom.	DigiCert can enable the organization to embed verifiable proof of origin and attribution directly into content at creation. Each asset carries a secure, tamper-evident record of its source and any edits, allowing platforms and audiences to confirm authenticity and detect manipulation or misattribution.	The organization can prove content authenticity in real time and at scale, reducing misinformation and false attribution.

From uncertainty to verifiable trust

As AI-generated content becomes ubiquitous, the ability to establish authenticity will become a foundational requirement for digital trust. By embedding cryptographic proof directly into content and anchoring that proof in trusted, independent infrastructure, organizations can move from assumption to verification.

This extends the principles of intelligent trust beyond systems and models to the outputs they produce, ensuring that what AI creates can be trusted wherever it's consumed.

Conclusion:

The future of trust in the age of AI

As AI systems take on more decisions and actions, the stakes rise. Autonomous agents act, models decide, and content shapes perception at an unprecedented scale. In this environment, trust cannot rely on visibility or policy alone. It must be built into systems, continuously enforced, and cryptographically proven.

DigiCert delivers this foundation through a unified, three-layer architecture: DNS-based enforcement at the network edge, cryptographic identity and authorization for agents, and hardware-backed integrity for model execution. Together, these layers create a control plane for AI trust where actions are governed, identities are verified, and outcomes are auditable.

Trust must also extend to what AI produces. By embedding verifiable provenance into digital content, organizations can ensure authenticity, detect manipulation, and restore confidence in the information others rely on.

This is the shift from digital trust to intelligent trust, extending proven PKI principles to the systems and content that now create, decide, and influence on behalf of organizations. DigiCert is uniquely positioned to lead this shift. Operating a significant portion of the world's authoritative DNS infrastructure, alongside its global PKI platform and leadership in emerging standards, DigiCert brings the scale, neutrality, and expertise required to establish verifiable trust across the AI ecosystem.

> ***Organizations that succeed with AI will be those that can operate with verifiable trust at scale.***

Organizations that succeed with AI will not be those that move fastest, but those that can operate with verifiable control and trust at scale.