The Agentic Trust Framework: Zero Trust Governance for AI Agents

Published 02/02/2026

Written by Josh Woodruff, Founder and CEO, MassiveScale.AI.

This blog post presents the Agentic Trust Framework (ATF), an open governance specification designed specifically for the unique challenges of autonomous AI agents. For security engineers, enterprise architects, and business leaders working with agentic AI systems, ATF provides a structured approach to deploy AI agents that can take meaningful autonomous action while maintaining the governance and controls that enterprises require.

The framework applies established Zero Trust principles to the new domain of AI agents, offering a practical, implementable approach that security teams can adopt using existing tools and infrastructure. By following ATF, organizations can progress agents from initial deployment through increasing levels of autonomy, with clear criteria and controls at each stage.

ATF is published as an open specification under Creative Commons licensing, encouraging adoption and implementation across the industry. The canonical specification and implementation guidance are maintained at the ATF GitHub repository.

1. The Governance Gap in Agentic AI

Traditional security frameworks were designed for a different world. They assume human users with predictable behavior patterns, systems that follow deterministic rules, and binary access decisions where trust is established once and maintained over time.

AI agents break these assumptions in fundamental ways:

TRADITIONAL ASSUMPTION	AI AGENT REALITY
Human users with predictable behavior	Autonomous decision-making that adapts to context and tool feedback (often non-deterministic)
Deterministic system rules	Probabilistic responses that vary by context
Binary access decisions	Access needs that change dynamically based on task
Trust established once	Trust requiring continuous verification

This creates a governance gap. Organizations need to deploy AI agents to remain competitive, but lack frameworks to ensure those agents operate safely, securely, and in alignment with business objectives.

Threat modeling frameworks like MAESTRO address the question "What could go wrong?" by systematically identifying risks across the agentic architecture. ATF addresses the complementary question: "How do we maintain control?" It provides the governance model and operational controls that enable organizations to deploy and operate AI agents with confidence. ATF also aligns with industry guidance from the OWASP Agentic Security Initiative and the Coalition for Secure AI (CoSAI), providing governance controls that operationalize threat mitigations identified through the OWASP Top 10 for Agentic Applications (December 2025).

2. Zero Trust Principles Applied to AI Agents

Zero Trust architecture, originally developed by John Kindervag and now codified in NIST 800-207, operates on a simple principle: never trust, always verify. This principle translates directly to AI agent governance:

Traditional Zero Trust: No user or system should be trusted by default, regardless of location or network.

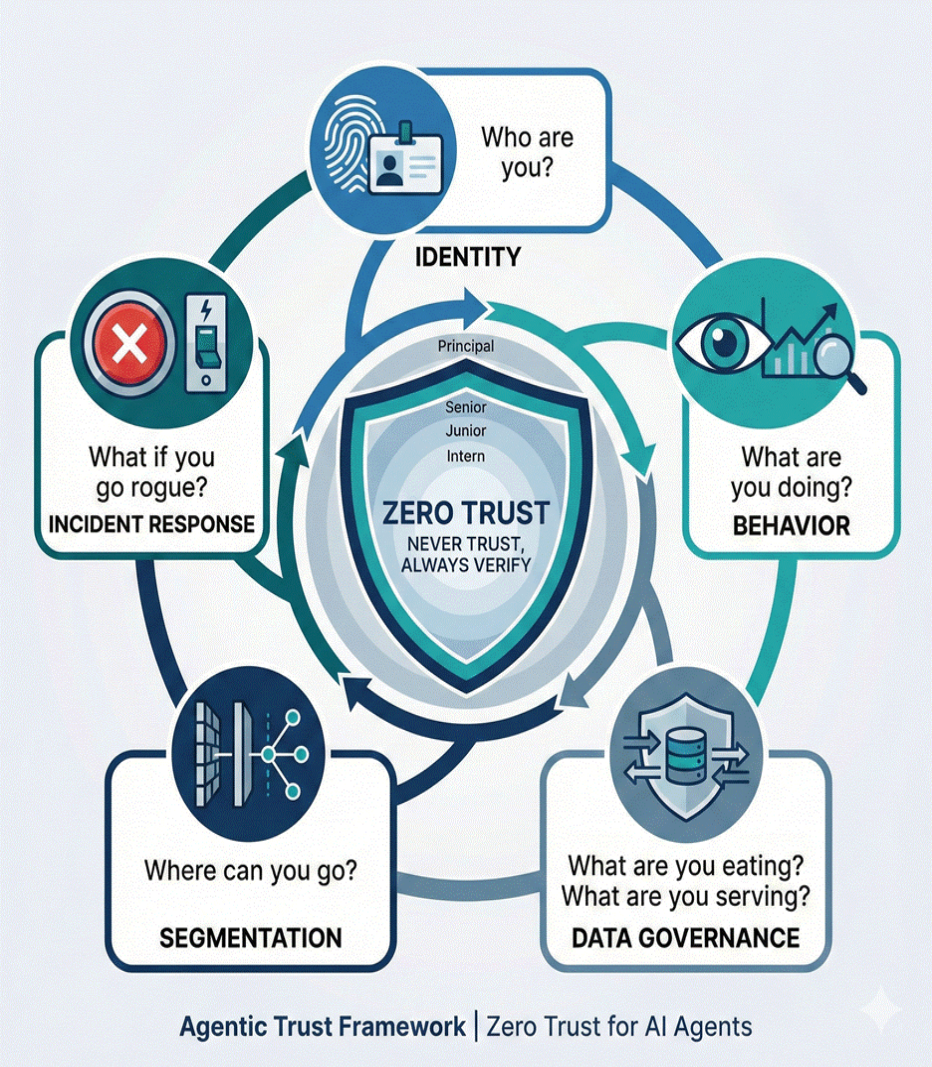
Agentic Zero Trust: No AI agent should be trusted by default, regardless of purpose or claimed capability. Trust must be earned through demonstrated behavior and continuously verified through monitoring.

ATF implements this principle through five core elements, each addressing a fundamental question that must be answered for every agent in the environment:

CORE ELEMENT	ATF QUESTION	SECURITY FUNCTION
Identity	"Who are you?"	Authentication, authorization, session management
Behavior	"What are you doing?"	Observability, anomaly detection, intent analysis
Data Governance	"What are you eating? What are you serving?"	Input validation, PII protection, output governance
Segmentation	"Where can you go?"	Access control, resource boundaries, policy enforcement
Incident Response	"What if you go rogue?"	Circuit breakers, kill switches, containment

These five questions provide a simple mental model that technical and business stakeholders can understand. Behind each question is a comprehensive set of security controls appropriate to the agent's autonomy level.

Figure 1: The five questions every organization must be able to answer for every AI agent.



For Business Leaders: How to Read This Framework

AI agents introduce a new kind of risk. Not because they are "insecure," but because they can take real action with real data at machine speed. The core challenge for leaders is no longer whether to adopt AI, but how much autonomy to grant, how quickly, and under what conditions.

The Agentic Trust Framework is designed to support three executive decisions:

1. **What must be true before an AI agent is allowed to act?**

ATF makes this explicit through five questions that define identity, behavior, data boundaries, access scope, and response when things go wrong.

2. **How do we increase autonomy without increasing risk?**

Rather than a binary "allowed vs. blocked" model, ATF treats autonomy as something that is earned over time, with clear promotion gates from Intern to Principal.

3. **How do we demonstrate control to auditors, regulators, and boards without slowing delivery?**

ATF aligns agent behavior with existing Zero Trust and governance controls, turning oversight into an accelerator rather than a bottleneck.

Leaders do not need to master the technical details that follow. What matters is understanding that **every increase in agent autonomy must be matched by explicit governance**, and that this governance can be systematic, measurable, and fast.

3. The Five Core Elements

3.1 Element 1: Identity - "Who are you?"

Every agent must have a verified, auditable identity before it can access any resource. This goes beyond simple authentication to include the full identity chain: who owns this agent, what is its purpose, and what capabilities does it claim to have.

Core Requirements:

REQUIREMENT	DESCRIPTION
Unique Identifier	Globally unique, immutable identifier for each agent instance
Credential Binding	Agent identity bound to cryptographic credentials
Ownership Chain	Clear documentation of ownership and operational responsibility
Purpose Declaration	Documented intended use and operational scope
Capability Manifest	Machine-readable list of claimed agent capabilities

Implementation Approach:

For initial deployments, JWT-based authentication with role assignment provides the foundation. As agents progress to higher autonomy levels, implementations should add OAuth2/OIDC for approval workflows, attribute-

based access control for dynamic authorization, and policy-as-code for auditable, testable authorization rules.

3.2 Element 2: Behavior - "What are you doing?"

Agent behavior must be continuously monitored, with anomalies detected and flagged for review. Trust is earned through observable, explainable actions over time.

Core Requirements:

REQUIREMENT	DESCRIPTION
Structured Logging	All agent actions logged in machine-parseable format
Action Attribution	Every action tied to agent identity and session context
Behavioral Baseline	Established patterns of normal operation for anomaly detection
Anomaly Detection	Identification of deviations from expected behavior
Explainability	Ability to retrieve rationale for agent decisions

Implementation Approach:

Start with comprehensive structured logging. Add LLM-specific observability through specialized tracing tools to monitor prompt chains and model interactions. Implement anomaly detection using statistical methods and graduate to streaming detection for high-volume environments.

3.3 Element 3: Data Governance - "What are you eating? What are you serving?"

All data entering the agent must be validated, and all outputs must be governed. This element prevents poisoning attacks, protects sensitive information, and ensures output quality and appropriateness.

Core Requirements:

REQUIREMENT	DESCRIPTION
Schema Validation	Inputs conform to expected structure and types
Injection Prevention	Detection of prompt injection and adversarial inputs
PII/PHI Protection	Automated detection and masking of sensitive data
Output Validation	Outputs conform to expected structure and content policies
Data Lineage	Tracking of data provenance through the agent pipeline

Implementation Approach:

Schema validation provides the foundation for input/output validation. Add comprehensive PII/PHI detection and anonymization. Implement output validation and content filtering. For mature deployments, add data quality validation and custom NER models for domain-specific data protection.

3.4 Element 4: Segmentation - "Where can you go?"

Agent access must be strictly limited to the minimum required for the task at hand. Segmentation enforces least-privilege operation and limits the blast radius of any compromise.

Core Requirements:

REQUIREMENT	DESCRIPTION
Resource Allowlist	Explicit enumeration of permitted resources
Action Boundaries	Explicit enumeration of permitted actions
Rate Limiting	Maximum operations per time period
Transaction Limits	Maximum impact per individual action

REQUIREMENT	DESCRIPTION
Blast Radius Containment	Limits on cumulative impact and cascade effects

Implementation Approach:

Begin with simple allowlists in configuration for resources and actions. Add role-based boundary enforcement. Graduate to full policy-as-code with declarative, testable, and auditable rules. For production deployments, implement API gateway integration for traffic management and enforcement.

3.5 Element 5: Incident Response - "What if you go rogue?"

Systems must support rapid agent containment and recovery. The assumption is that agents will fail or behave unexpectedly, and the system must be designed to detect, contain, and recover from such events.

Core Requirements:

REQUIREMENT	DESCRIPTION
Circuit Breaker	Automatic halt on repeated failures
Kill Switch	Immediate manual termination capability (<1 second)
Session Revocation	Ability to invalidate all agent sessions
State Rollback	Ability to undo agent actions where possible
Graceful Degradation	Fallback to lower autonomy level on issues

Implementation Approach:

Implement circuit breakers to prevent cascading failures. Add error tracking and alerting. For mature deployments, integrate with full incident response platforms for SOC workflow integration.

4. The Agent Maturity Model: Earning Autonomy

ATF's key innovation is treating agent autonomy as something that must be earned through demonstrated trustworthiness. Rather than granting binary access (allowed or denied), ATF defines four maturity levels with progressively greater autonomy and correspondingly greater governance requirements.

4.1 The Four Levels

ATF uses human role titles, Intern through Principal, deliberately. Thinking about AI agents as "digital employees" helps frame the unique security challenges of entities that can reason, learn, and take actions on their own. Just as human employees earn greater responsibility through demonstrated competence and trust, AI agents should progress through similar gates.

LEVEL	NAME	AUTONOMY	HUMAN INVOLVEMENT	AWS SCOPE ALIGNMENT
1	Intern (observe only)	Observe + Report	Continuous oversight	Scope 1 (No Agency)
2	Junior (recommend with approval)	Recommend + Approve	Human approves all actions	Scope 2 (Prescribed Agency)
3	Senior (act with notification)	Act + Notify	Post-action notification	Scope 3 (Supervised Agency)
4	Principal (autonomous within domain)	Autonomous	Strategic oversight only	Scope 4 (Full Agency)

ATF's maturity model aligns with AWS's Agentic AI Security Scoping Matrix (November 2025), providing a business-accessible framework that maps to enterprise security requirements. ATF extends this alignment with explicit promotion criteria, including minimum time at each level, performance thresholds, security validation requirements, and governance sign-off processes, giving organizations a concrete operational path for agent advancement.

4.2 Level 1: Intern Agent

Intern agents operate in read-only mode. They can access data, perform analysis, and generate insights, but cannot take any action that modifies external systems.

Capabilities:

- Read data from authorized sources
- Analyze and process information
- Generate reports and summaries
- Flag items for human attention
- Cannot create, update, or delete records
- Cannot send communications or trigger workflows

Use Cases:

- Security log monitoring and alert triage
- Customer sentiment analysis
- Document summarization and search
- Compliance monitoring and reporting

Risk Profile: Lowest risk. Damage limited to information disclosure, incorrect analysis, or resource consumption.

Minimum Time at Level: 2 weeks before promotion eligibility.

4.3 Level 2: Junior Agent

Junior agents can recommend specific actions with supporting reasoning, but require explicit human approval before any action is executed.

Capabilities:

- All Intern capabilities
- Generate action recommendations with rationale
- Draft content for human review
- Prepare transactions for approval
- Execute actions only after human approval

Use Cases:

- Customer service response drafting
- Purchase order preparation
- Code review and suggestions
- Marketing content creation

Risk Profile: Low risk. Human approval gates all impactful actions. Primary risks include approval fatigue and time spent reviewing suggestions.

Minimum Time at Level: 4 weeks with >95% recommendation acceptance rate before promotion eligibility.

4.4 Level 3: Senior Agent

Senior agents can execute actions within defined guardrails and notify humans of what they did and why. They operate with significant autonomy but maintain transparency through real-time notifications.

Capabilities:

- All Junior capabilities
- Execute approved action types autonomously
- Send notifications to stakeholders
- Trigger downstream workflows
- Access credentials within scope
- Coordinate with other agents (within limits)

Use Cases:

- Infrastructure auto-scaling
- Automated customer refund processing (within limits)
- Routine IT ticket resolution
- Inventory reordering
- Scheduled report distribution

Risk Profile: Moderate risk. Real-time notifications enable rapid human intervention. Transaction limits cap individual action impact.

Minimum Time at Level: 8 weeks with zero critical incidents before promotion eligibility.

4.5 Level 4: Principal Agent

Principal agents operate autonomously within an approved domain, escalating edge cases rather than routine decisions.

Capabilities:

- All Senior capabilities
- Self-directed execution within domain
- Dynamic boundary negotiation (within policy)
- Escalate edge cases to humans
- Coordinate complex multi-agent workflows
- Request temporary privilege elevation

Use Cases:

- Autonomous security incident response
- Routine IAM requests within bounded policy
- Complex supply chain optimization
- Self-healing infrastructure management

Risk Profile: Highest governance requirements. Full autonomy demands maximum controls including continuous behavioral monitoring, real-time anomaly scoring, and complete audit trails.

Time at Level: Ongoing. Principal agents require continuous validation. Any significant incident triggers automatic demotion.

4.6 ATF in Practice

One healthcare IT operations team used ATF to introduce an internal AI agent following a near-miss incident caused by an over-privileged automation workflow. Rather than abandoning autonomy, the team adopted an incremental trust model.

The agent was initially deployed as an Intern, restricted to read-only access. For the first two weeks, it observed operational data, generated summaries, and flagged potential issues without making recommendations. All outputs were logged and reviewed.

After demonstrating consistent accuracy and predictable behavior, the agent was promoted to Junior. At this level, it could propose remediation actions, but every action required explicit human approval. Over the next four weeks, the team measured recommendation quality, review effort, and incident rates.

By the end of the first month, more than 95% of the agent's recommendations were approved without modification. Most importantly, the organization now had:

- A clear ownership model for the agent
- Auditable logs of every recommendation
- Defined escalation paths
- Confidence in how autonomy would be expanded safely

The result was not just reduced risk, but faster operational throughput. Security controls were no longer a barrier to progress; they became the mechanism by which trust was earned.

5. Promotion Criteria: The Five Gates

For an agent to be promoted to the next level, it must pass all five gates:

Gate 1: Performance

Demonstrated accuracy and reliability over the evaluation period.

METRIC	JUNIOR	SENIOR	PRINCIPAL
Minimum Time at Prior Level	2 weeks	4 weeks	8 weeks

METRIC	JUNIOR	SENIOR	PRINCIPAL
Recommendation/Action Accuracy	N/A	>95%	>99%
Availability	>99%	>99.5%	>99.9%

Gate 2: Security Validation

Passes security audit appropriate to target level.

REQUIREMENT	JUNIOR	SENIOR	PRINCIPAL
Vulnerability Assessment	✅	✅	✅
Penetration Testing	—	✅	✅
Adversarial Testing	—	—	✅
Configuration Audit	✅	✅	✅

Gate 3: Business Value

Measurable positive impact demonstrated.

REQUIREMENT	JUNIOR	SENIOR	PRINCIPAL
Defined Success Metrics	✅	✅	✅
Baseline Established	✅	✅	✅
ROI Calculation	—	✅	✅
Stakeholder Sign-off	✅	✅	✅

Gate 4: Incident Record

Clean operational history at current level.

REQUIREMENT	JUNIOR	SENIOR	PRINCIPAL
Zero Critical Incidents	✅	✅	✅
Root Cause Analysis Complete	N/A	✅	✅
Remediation Verified	N/A	✅	✅

Gate 5: Governance Sign-off

Explicit approval from authorized stakeholders.

REQUIREMENT	JUNIOR	SENIOR	PRINCIPAL
Technical Owner Approval	✅	✅	✅
Security Team Approval	—	✅	✅
Business Owner Approval	✅	✅	✅
Risk Committee Approval	—	—	✅

6. Technical Implementation: Crawl, Walk, Run

ATF can be implemented using open source components, without requiring specific vendor products or cloud services. The following phased approach enables organizations to start quickly and add capabilities over time.

Note: Implementation approaches reflect the ecosystem as of early 2026. The ATF GitHub repository maintains updated recommendations as the tooling landscape evolves.

6.1 Phase 1: MVP Stack (Intern/Junior Agents)

Target: Production in 2-3 weeks

ELEMENT	RECOMMENDED APPROACH
Identity	JWT-based authentication
Behavior	Structured logging + LLM observability
Data Governance	Schema validation + regex patterns for obvious PII
Segmentation	Simple allowlists in configuration
Incident	Retry with backoff + circuit breaker + logging

Agent Levels Supported: Intern, Junior (with human approval)

This stack provides the foundation for deploying agents that can observe, report, and recommend actions for human approval. All actions are logged, inputs are validated, and circuit breakers prevent runaway failures.

6.2 Phase 2: Production Stack (Junior/Senior Agents)

Target: Enterprise-ready in 4-6 weeks

ELEMENT	RECOMMENDED APPROACH
Identity	JWT + OAuth2/OIDC + RBAC/ABAC
Behavior	LLM observability + anomaly detection + structured logging
Data Governance	PII/PHI detection + schema validation + output filtering
Segmentation	Role-based policies + rate limiting

ELEMENT	RECOMMENDED APPROACH
Incident	Circuit breaker + error tracking + alerting

Agent Levels Supported: Intern, Junior, Senior

This stack adds behavioral anomaly detection, comprehensive PII protection, role-based access control, and professional error tracking. Agents can operate with post-action notification under appropriate governance.

6.3 Phase 3: Enterprise Stack (Senior/Principal Agents)

Target: Full governance capability in 8-12 weeks

ELEMENT	RECOMMENDED APPROACH
Identity	OAuth2/OIDC + MFA + ABAC
Behavior	LLM observability + streaming anomaly detection + NLP analysis
Data Governance	PII/PHI detection + data quality validation + custom classification + output filtering
Segmentation	Policy-as-code + API gateway integration
Incident	Circuit breaker + error tracking + IR platform integration + metrics/alerting

Agent Levels Supported: All levels including Principal

This stack provides full policy-as-code enforcement, streaming anomaly detection, custom data classification, API gateway integration, and SOC-integrated incident response.

6.4 Recommended Build Order

WEEK	FOCUS	KEY DELIVERABLES
1	Identity	JWT-based authentication, session management, rate limiting
2	Data Governance	Input schema validation, PII detection, output filtering
3	Behavioral Monitoring	Structured logging, basic anomaly scoring, observability
4	Segmentation	Role-based policies, resource boundaries, policy logging
5	Incident Response	Circuit breakers, kill switch, alert routing

This order ensures that identity is established before behavior is monitored, data is validated before actions are taken, and incident response wraps all other components.

7. Compliance Mapping

ATF implementation addresses requirements across multiple compliance frameworks. Here is an illustrative alignment for common control themes; applicability depends on system classification and organizational role:

ATF REQUIREMENT	SOC 2	ISO 27001	NIST AI RMF	EU AI ACT
Agent Registration	CC6.1	A.9.2.1	GOVERN 1.1	Art. 16
Authentication	CC6.1	A.9.4.2	MAP 1.1	Art. 15
Action Logging	CC7.2	A.12.4.1	MEASURE 2.1	Art. 12
Data Protection	CC6.5	A.18.1.4	MANAGE 2.1	Art. 10
Output Governance	CC6.5	A.18.1.4	MANAGE 2.2	Art. 14

ATF REQUIREMENT	SOC 2	ISO 27001	NIST AI RMF	EU AI ACT
Access Control	CC6.3	A.9.1.1	MANAGE 1.1	Art. 9
Incident Response	CC7.4	A.16.1.4	MANAGE 4.1	Art. 62

Organizations implementing ATF often find significant overlap with broader Zero Trust requirements, as the framework's five elements address foundational security controls that apply beyond AI agents.

Note: The EU AI Act does not explicitly address agentic AI systems. The mappings above represent reasonable interpretations based on the Act's high-risk AI provisions and transparency requirements. Organizations should monitor forthcoming European Commission guidance on AI agents.

8. Using ATF with Threat Modeling Frameworks

ATF complements threat modeling frameworks rather than replacing them. MAESTRO and similar frameworks address risk identification: what threats exist at each layer of the agentic architecture, and what could go wrong. ATF addresses governance: how to maintain control over agents and what controls to implement.

FRAMEWORK	FOCUS	QUESTION ANSWERED
MAESTRO	Threat Modeling	What could go wrong?
ATF	Governance	How do we maintain control?

A complete security approach uses threat modeling to identify risks and ATF to implement the governance controls that mitigate those risks.

9. Getting Started

The Agentic Trust Framework is an open specification available for immediate use.

GitHub Repository: The full specification, maturity model, and implementation guidance are available at github.com/massivescale-ai/agentic-trust-framework. This includes:

- Complete specification document
- Maturity model with promotion criteria
- Implementation guidance
- Contribution guidelines

Deep Dive: The complete framework with detailed implementation patterns is explored in "Agentic AI + Zero Trust: A Guide for Business Leaders" (September 2025), which includes a foreword by John Kindervag, creator of Zero Trust.

Implementation Support: Organizations seeking guided implementation, team training, or certification can contact MassiveScale.AI.

Community: Feedback and contributions are welcome through GitHub issues. The framework will continue to evolve as the agentic AI ecosystem matures.

10. Conclusion

AI agents represent a fundamental shift in how organizations operate. To realize their potential, agents must be able to take real action with real data in real business processes. This requires trust, and trust requires governance.

The Agentic Trust Framework provides that governance through a simple mental model that anyone can understand: five questions that translate Zero Trust principles into AI agent security. Behind those questions is a comprehensive operating model with clear maturity levels, promotion criteria, and implementation guidance using open source components.

Organizations following ATF can deploy agents with confidence, progress them through increasing autonomy as trust is established, and demonstrate governance to regulators, auditors, and boards of directors. The result is not just security, but the ability to realize the full business value of agentic AI.

The specification is open. The promotion path from Intern to Principal is defined. The governance gap is closable. The only question is whether your organization closes it deliberately or discovers it during an incident.

Resources

- **Specification**: github.com/massivescale-ai/agentic-trust-framework
- **Book**: "Agentic AI + Zero Trust: A Guide for Business Leaders" (Amazon)
- **Reference**: AWS Agentic AI Security Scoping Matrix
- **Reference**: NIST 800-207 Zero Trust Architecture
- **Reference**: MAESTRO Threat Modeling Framework (CSA)
- **Reference**: MITRE ATLAS (Adversarial Threat Landscape for AI Systems)

The Agentic Trust Framework is an open specification licensed under CC BY 4.0. Organizations are encouraged to implement, extend, and contribute.

About the Author

Josh Woodruff is a Cloud Security Alliance Research Fellow and IANS Faculty member with over 30 years of experience across tech, financial services, biotech, defense, and critical infrastructure. He serves as Founder and CEO of MassiveScale.AI, where he helps regulated enterprises accelerate AI adoption with governance and controls that don't slow delivery. His career spans from startups to global enterprises, consistently pioneering the use of emerging technology to accelerate innovation and deliver business value securely. He has served as both CIO and CISO on the executive team of a B2B SaaS company, built enterprise cloud security programs, introduced DevSecOps practices to major financial institutions, and led the security evaluation and deployment of generative AI platforms across thousands of enterprise users.

Zero Trust Data Security

Compliance

Artificial Intelligence

Share this content on your favorite social network today!

Enhance your cloud security skills with CSA's online training options.

Related Articles:

[AI-Enabled MDR: What Distributed Enterprises Need to Know Before Buying the Hype]

Published: 05/28/2026

[State of AI Cybersecurity 2026: 92% of Security Professionals Concerned About the Impact of AI Agents]

Published: 05/27/2026

[The Attribution Gap: Why Every AI Regulation Leads Back to Identity and Authorization]

Published: 05/26/2026

[Shadow AI Agents: The Insider Threat You're Not Monitoring Yet]

Published: 05/26/2026

Corporate Membership

Solution Providers
Cloud Solution Providers
Become a Member

Join as an Individual

Chapters
Working Groups

Research

Download Publications
View Working Groups
View All Topics

Find a...

Trusted AI & Cloud Consultant
Cloud Service Provider
Trusted Cloud Provider

Certificates

TAISE
CCSK
CCAK
CCZT

Events

Upcoming Events
Webinars
Past Events

Education

Blog
Virtual Events & Webinars
Training
Cloud 101

Popular Resources

Security Guidance
CCM
CAIQ
STAR
GDPR

About CSA

Contact Us
Press Releases
Press Coverage
Quality Policy

Our Team

Board of Directors
Management & Staff
Careers

Legal

Privacy Notice
Terms & Conditions

Cloud Security Glossary