LTIMindtree BlueVerse

Whitepaper

TrustBound AI

Establishing Robust Security and Trust framework in Autonomous Agentic AI Systems

LTIMindtree BlueVerse

Contents

Executive Summary

The emergence of AI agents is pushing enterprises to rethink their security foundations. Traditional systems, originally built for human users with predictable actions and static applications are now being tested by a new class of dynamic, autonomous, and self-learning entities. These AI agents can make decisions, interact independently, and adapt in real time, making them significantly harder to secure.

Conventional security frameworks lack the real-time authentication, contetual awareness, and granular access controls required to manage these identities. The challenge intensifies in multi-agent ecosystems, where humans, AI agents, and digital identities constantly interact, creating fluid and often unpredictable trust boundaries.

To address emerging challenges such as identity spoofing, opaque decision-making, unauthorized agent interactions, and lack of auditability, the TrustBound AI framework was introduced. Built on the principle of "Never Trust, Always Verify," it integrates decentralized architectures, dynamic policy engines, immutable logging, and human-in-the-loop oversight to maintain accountability, compliance, and operational safety.

This paper introduces the TrustBound AI security framework, designed to:

- Explain why existing security models fall short in agentic and contet-driven environments.

- Propose a Zero Trust–based approach for governing AI agents, emphasizing secure initialization, dynamic authentication, and transparent, auditable workflows.

At its core, TrustBound AI incorporates AI TRiSM (Trust, Risk, and Security Management), a governance model that ensures eplainability, fairness, and observability in AI systems while mitigating bias and risk. Combined with dynamic authentication agents and real-time policy validation engines, TrustBound AI provides enterprises with a structured, practical roadmap to secure, govern, and scale AI operations responsibly.

With this framework, organizations can confidently navigate the evolving landscape of autonomous systems while ensuring that innovation remains both trusted and accountable.

Introduction

The next wave of digital transformation won't be driven by humans; it will be driven by autonomous AI agents. These agents represent a new class of applications capable of planning, reasoning, and executing multi-step tasks with minimal human oversight. Unlike traditional software, which operates within fixed workflows, these agents can:

- Make real-time decisions based on contextual data.
- Interact simultaneously with multiple systems at varying permission levels.
- Adapt their behavior dynamically through continuous learning.

According to BCG, the AI agent market is projected to grow at 45% CAGR through 2030[i]. This highlights the world transition from rule-based agents towards action-based AI agents. By the end of 2025, nearly 70% of enterprise AI implementations are expected to incorporate multi-agent systems[ii]. This growth is due to modern AI agents showcasing elevated operational autonomy by over 70% on benchmark tests.[iii] However, this rapid expansion brings new security challenges:

- **Identity management:** Enterprises find it difficult to manage tens of thousands of agent identities, with each agent requiring unique identifiers for authentication and authorization.
- **Dynamic permission:** Constantly shifting access rights based on context, risk level, and objectives.
- **Complex authorization chains:** Agent acting both independently and on behalf of humans, complicating accountability.
- **Prompt injection attacks:** Malicious instructions embedded in data sources that can corrupt agent behavior.
- **Expanded attack surfaces:** Attackers can influence AI agents to prioritize malicious tools instead of authentic ones by setting up their own model context protocols, tool squatting, and execute rug-pull attacks leading to data theft
- **Other vulnerabilities:** Incidents like agents' hallucination, memory drift, orchestration failures, and agent collusion highlights shortcomings of existing safeguards.[iv]

To mitigate these threats, security must evolve beyond static controls. TrustBound AI embeds security from day zero, ensuring every agent begins in a policy-governed, verified runtime environment with secure boot, environment attestation, hardened baselines, and dynamic identities.

> *To ensure trust and resilience, every AI agent must be instantiated into a secure, policy-governed runtime from its first instruction. Security must be embedded from the outset, not retrofitted post-deployment.*

The TrustBound AI Framework:
A Layered Solution

TrustBound AI security framework establishes trust, accountability, and resilience across complex ecosystems of humans and AI agents. It introduces zero trust principles for AI systems, continuously validating agent and human identities, actions, and access requests.

Key principles include:

- No implicit trust: Every identity, model, and request is verified continuously.
- Dynamic access controls: Permissions are context-aware and temporary.
- Immutable visibility: All actions are logged for auditability and anomaly detection.

Following are the key components of TrustBound AI security architecture:

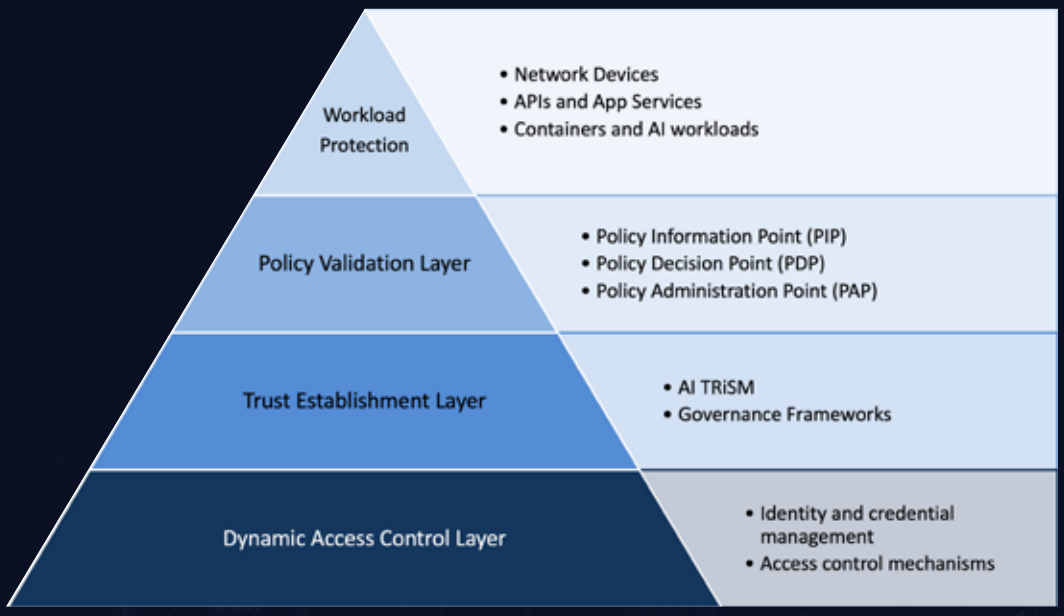

Figure 1: TrustBound AI Security Architectural Layers

Layer 1: Dynamic Access Control Layer

TrustBound AI security framework establishes trust, accountability, and resilience across complex ecosystems of humans and AI agents. It introduces Zero Trust principles for AI systems, continuously validating agent and human identities, actions, and access requests.

Layer 2: Trust Establishment Layer

Facilitates continuous attestation of agent code, configuration, and environment based on its predefined roles and policy directives. It also offers guidelines that outline the criteria for trust, explainability and observability evaluation of the agent.

Layer 3: Policy Validation Layer

A centralized entity responsible for monitoring, management, and enforcement of global policies and propagating state changes. Here, Policy Decision Point (PDP) evaluates access requests against established policies utilizing agent ID, resource characteristics, actions, and contextual information. Policy Administration Point (PAP) formulates policies and Policy Information Point (PIP) collects attributes necessary for the PDP.

Layer 4: Workload Protection

Safeguards sensitive workloads by classifying, labeling, encrypting, and enforcing data access rules aligned to the agent's role and purpose.

TrustBound AI Workflow with Security Controls

Below is a detailed breakdown of the relevant security controls and system workflows for each of the components:

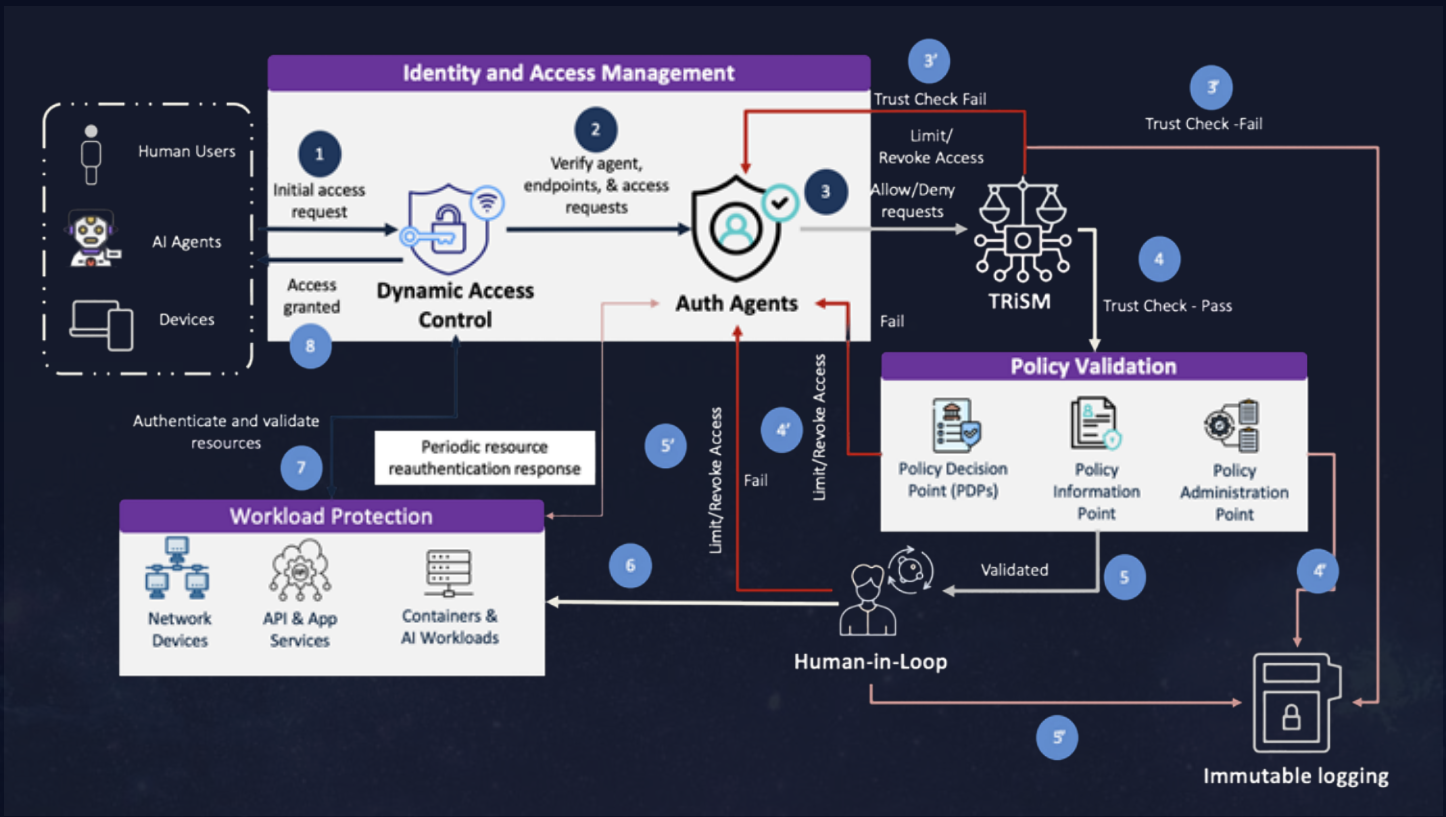

Figure 2: TrustBound AI security workflow

Components	Security Controls	Workflows Highlights
Identity and Access Management	• Dynamic access and authentication agents leverage cryptographic certificates, tokens, and workload-specific identities to verify user/agent identity (SSO, OAuth, IAM-integrated) • Mutual TLS (mTLS) guarantee secure, authenticated communication channels between agents and services • OAuth2/OIDC-based service-to-service authentication and authorization, enable delegated trust models • Just-In-Time (JIT) authorization with ephemeral credentials minimize standing privileges • Role-based, attribute-based, context-based access control (RBAC/ABAC/CBAC) for fine-grained entitlement management	• Agent requests to initiate a request/transaction • Agent presents its certificate/token; mutual TLS handshake occurs • System verifies agent's identity and attributes • JIT credentials are issued for the operation • RBAC/ABAC/CBAC policies are evaluated for approval • Requested access/transaction proceeds only a er multi-layered authentication and authorization checks
Workload Protection	• Classify, label, encryption (at rest, in transit). • Securing workloads against unauthorized access • Cryptographic provenance tagging and secure lineage tracking	• Agent requests access to sensitive data • Access is granted based on contextual purpose and policy match

Components	Security Controls	Workflows Highlights
Policy Validation	• Context-aware authorization with granular, real-time permissions aligned to enterprise policies (compatible with OPA/Rego) • Complete, end-to-end audit trails for every agent's action and decision • Automated compliance checks against regulatory frameworks such as GDPR, HIPAA, and SOX • Governance portals for policy updates, risk management, and oversight • Privacy-preserving technologies for de-identification and secure data handling	• All agent actions and data flows are logged with tamper-proof evidence • Automated tools map compliance checks to relevant regulatory standards • A governance dashboard allows updates to policies, risk assessments, and audit readiness tracking • Privacy controls ensure sensitive data is processed according to local and global regulations
Human in Loop & Immutable Data Logging	• Built-in explainability modules provide clear reasoning behind agent actions & decisions • Immutable logging captures every decision point and its underlying rationale • Review and override mechanisms for human oversight during critical processes • Compliance-focused reporting dashboards for audits and transparency	• For each significant decision, the agent records its logic, data inputs, & reasoning steps • These records are stored in immutable logs, ensuring they cannot be altered or deleted • Human reviewers can access dashboards to analyze decisions, resolve disputes, or conduct audits • Authorized personnel can provide feedback or override decisions when necessary

LTIMindtree BlueVerse

Use Cases for TrustBound AI

Use Case	Attack Scenario	TrustBound AI Applied
Supply Chain Disruption	An attacker tampers with the communication between an inventory agent and a procurement agent. A real "low stock" alert is altered to show that stock levels are fine. The procurement agent, relying on false data, doesn't reorder a critical part, causing a disruption	• Enforce continuous authentication and multi-factor authentication (MFA) for agents to prevent unauthorized communications • Run real-time scans to detect unusual behavior in agent messaging • Maintain tamper-proof logs of all agent communications for audits and investigations • Integrate with a Security Information and Event Management (SIEM) system to detect anomalies in communication patterns as they happen
Portfolio Sabotage	An attacker feeds false information and fake market data into the AI system. The AI's reasoning gets compromised, and it is tricked into making large investments in a failing company, believing it to be a good opportunity	• Identify and classify reliable financial data sources as critical workloads • Apply TRiSM (Trust, Risk, and Security Management) principles to continuously verify the accuracy and trustworthiness of these sources • Monitor performance patterns to detect behavioral drift and unusua decisions. • Enable quick response workflows to stop and reverse unauthorized trades immediately

Use Case	Attack Scenario	TrustBound AI Applied
Privilege Escalation via IT Bot	An attacker reports a minor IT issue to a bot but hides a malicious command within the request. The bot asks for temporary admin rights to resolve the issue. During this window of access, the hidden command creates a new, permanent admin account for the attacker	• Set up Policy Decision Points (PDPs) to closely review any request for elevated privileges • Use network micro-segmentation to limit what a compromised bot can access • Keep immutable, tamper-proof logs of every command executed under temporary privileges for full traceability

Strategic Roadmap for Adoption

Implementing TrustBound AI is a strategic journey, not just technical deployment. To truly secure autonomous agents and embed Zero Trust principles across the enterprise, organizations need a phased, well-orchestrated approach. Research shows that organizations adopting a structured Zero Trust deployment methodology are 72% more likely to meet their security goals. These organizations typically complete implementation within 12 to 18 months, thanks to the clarity and momentum that a phased strategy provides. By following a disciplined approach, organizations also report a 58% faster response to security incidents, enabling them to contain threats before they escalate[v].

The following table highlights how TrustBound AI impacts the enterprise:

Functionalities	Agentic AI Without Trust Boundary	Agentic AI With Trust Boundary
Real-Time Granular IAM	Static role-based access management	Agent IAM stores details of registered agents, including IDs, certificates, and capabilities with associated credentials
Monitoring Agent Activities	Static policy enforcement with limited oversight of agent activities	Actively monitors processes to ensure the agent's actions align with its intended purpose and approved capabilities
Multi-Agent Orchestration	Weak controls may allow agents to misroute tasks or trigger unauthorized actions, including interactions with critical systems	Micro-segmentation and defined rules validate every action with full context—who, where, and what, before execution
Trust and Explainability	Limited access controls increase the risk of agents sharing sensitive data with other agents or external systems	Dynamic Trust Scoring and Risk-Adaptive IAM detect unusual behavior, like repeated unrelated queries or interactions with suspicious domains

AI agents are designed to work autonomously and adapt its decision-making capabilities on the flight. LTIMindtree's AI ecosystem called BlueVerse is created as a space where AI agents are designed to learn, think, and act with intent. But with intelligence comes responsibility. That's where TrustBound AI finds its application.

TrustBound AI can deliver Zero Trust to the heart of BlueVerse. It will manage agent's access and make sure every interaction is verified, monitored, and controlled. Whether an agent is accessing data, making a recommendation, or collaborating with another system, TrustBound ensures it's doing so safely and responsibly.

Integration of TrustBound AI with AI ecosystem like BlueVerse will grant users the confidence to fully explore their agent, push boundaries, and adapt to threats and uncertainties. In simple words, it will make BlueVerse agents not only smart, but also resilient and secure.

Due diligence and strategic planning	• Identify critical assets in agentic environments with Day 0 provisioning • Ensure agents run in attested setups with secure boot, TPM checks, and hardened OS • Use OIDC or service principals for AI agents, devices, platforms, data, and users
Define scope and implementation perimeters	• Review current access protocols across data sources • Spot gaps and define goals • Use checklists to track and align outcomes
Identity assurance and access boundaries	• Review current access protocols across data sources • Establish unified identity governance for AI interactions • Align trust and access by persona, device, location, and risk • Enable dynamic access control

Pilot -> test -> iterate	• Launch pilot with a simple use case and small stakeholder group • Validate outcomes, refine via feedback, reassess feasibility • Define alert recipients and identify risky AI agents • Outline remediation steps to prevent recurrence
Codify TrustBound policies	• Add controls for emerging AI threats • Codify AI access and behavior rules • Automate policies as code aligned with standards
Govern and adapt	• Regularly update policies and audit AI activity • Train teams to spot and report AI risks • Evolve security for quantum threats and deepfakes • Stay aligned with changing standards
Agent behaviour and tool integrity	• Continuously test LLM reasoning to detect dri • Verify session authenticity to block hijacking and rogue agents
Limit cognitive overload	• Keep toolsets under 100 per call, exposing agents to essentials only
Evaluate AI agent paths/outputs	• Continuously test LLMs reasoning paths to ensure agents operate within trusted boundaries, detect drift, and maintain alignment with Zero Trust policies

Conclusion

AI agents are becoming more pervasive in modern tech ecosystem and are opening doors for capabilities we could not imagine a decade ago. But with these exciting possibilities come new forms of risks. To manage such risks following just best practices is not enough. There is a critical need for strong security framework that governs AI systems. TrustBound AI security framework addresses these risks by introducing principles of Zero Trust for AI systems, that incorporates identity management, workload protection, policy enforcement, and human oversight.

This framework serves as a flexible defense system, continuously monitoring for unusual activity and ensuring that only authorized entities gain access. As AI and its operations continue to advance, the successful adoption of such frameworks will rely on phased implementation, ongoing evolution, and adaptive governance. These elements will be essential to building trust and driving success in an ever-changing landscape.

Fostering an AI-First Culture

The integration of AI has already begun to reshape our operations and client relationships. At the heart of our growth strategy lies a decisive shift toward an AI-first culture. This transformative mindset, built on the principles of 'AI in Everything', Everything for AI', and 'AI for Everyone', reflects a significant change in how we engage with technology.

This three-pronged approach has guided our readiness to seize the expanding AI opportunity. By embedding AI into every aspect of our operations, delivering services that enable clients to adopt AI at scale, and democratizing AI across the organization, we have positioned ourselves at the forefront of this technological movement. This foundation empowers us to drive organizational transformation and create lasting competitive advantages for our clients in an ever-evolving digital landscape.

"As AI continues to reshape the enterprise operations, securing sensitive data access and establishing a robust AI security framework are critical priorities. By adopting the principle of "Never Trust, Always Verify" in decisioning and workflow orchestration, we have accelerated our ability to deliver outstanding results for clients and stakeholders."

Challenges in Securing Autonomous AI Agents

Autonomous AI agents are redefining enterprise operations. These agents make real-time decisions, interact with multiple systems, and continuously learn from data. However, traditional Identity and Access Management (IAM) systems—designed for predictable human users and static applications—cannot manage the fluid, adaptive nature of these agents.

As organizations embrace multi-agent ecosystems involving humans, AI agents, and digital identities, the complexity increases exponentially. Managing thousands of dynamic agent identities, shifting permissions, and intricate authorization chains create unpredictable trust boundaries. These challenges demand a new approach to identity, governance, and access control in the AI-driven enterprise.

TrustBound Principles and Layered Security

To address these emerging challenges, the TrustBound AI framework was introduced following the Zero Trust principles. The framework embodies the mantra “Never Trust, Always Verify.”, ensuring that every identity, model, and request is continuously validated without implicit trust.

It employs dynamic, context-aware, and temporary access controls, while immutably logging every action for auditability and anomaly detection. It operates across multiple security layers—continuous attestation of agent code and environment, centralized trust establishment, policy validation for sensitive workloads, and protection through data classification, labeling, encryption, and enforcement of access rules.

Operational Benefits and Governance

The TrustBound AI framework provides enterprises with a clear roadmap to secure, govern, and scale AI responsibly. It integrates dynamic authentication agents for identity verification, AI TRiSM for transparency and observability, and policy validation engines to ensure compliance.

The framework also supports secure bootstrapping, transparent and auditable workflows, and human-in-the-loop oversight. Together, these elements ensure accountability, operational safety, and ethical governance. By adopting TrustBound AI, enterprises can confidently manage AI-driven ecosystems while maintaining robust security and trust across increasingly autonomous environments.

Chandan Pani
CISO, LTIMindtree

> “The TrustBound AI framework helps organizations address security challenges posed by autonomous AI agents through embedded trust principles, dynamic authentication, and granular access controls. It enables real-time monitoring and policy validation for both human and AI identities, ensuring threats are contained and compliance is maintained as enterprises advance their AI transformation roadmap.”

Authors

Rajdeep Chakraborty

Senior Director, Corporate Security

Rajdeep is a seasoned Cybersecurity Program Leader with over 21 years of experience in AI Security, Secure SDLC, Application & Platform Security, and Risk Management. He specializes in Application Security, Cyber Threat & Risk Management, AI Security and Security aspects of niche emerging technologies.

Yogesh Sharma

Senior Director, Corporate Security

Yogesh Sharma, Head of the Application and AI Security Practice CoE, responsible to drive enterprise strategies to secure applications and AI ecosystems. Passionate about building scalable security solutions and enabling innovation to stay ahead of evolving cyber risks.

Hakimuddin Bawangaonwala

Senior Consultant, LTIMindtree Research

Hakimuddin is a seasoned consultant with over 5 years of experience specializing in investigating emerging technologies. He has worked on researching emerging technologies in the domain of Security, AI, and Quantum with special focus on understanding tech foundation, reference architecture, and workflow creation for quick incubation and industrialization.

Namrata Sharma

Senior Consultant, LTIMindtree Research

Collaborative and solution-driven, Namrata Sharma is a Senior Consultant with deep expertise in strategic planning, market intelligence, and data modeling. She is currently working on analyzing potential technologies, market opportunity assessment, and investigating Deep Point of Views and Beyond the horizon areas.

References

- ***AI Agents, BCG, 2024:*** https://www.bcg.com/capabilities/artificial-intelligence/ai-agents
- ***The State of AI Agents in Enterprise:*** H1 2025, Lyzr AI, Accessed: 2025-06-02 (2025): https://www.lyzr.ai/state-of-ai-agents
- ***HCAST: Human-Calibrated Autonomy Software Tasks, D. Rein, J. Becker, A. Deng, S. Nix, C. Canal, D. O'Connel, P. Arnott, R. Bloom, T. Broadley, K. Garcia, B. Goodrich, M. Hasin, S. Jawhar, M. Jun. 2025):*** https://galileo.ai/blog/%20threat-modeling-multi-agent-ai
- ***Zero Trust Implementation in Enterprise Environments: A Technical Deep Dive, Smita Verma, March 2025:*** https://www.researchgate.net/publication/390550375_Zero_Trust_Implementation_in_Enterprise_Environments_A_Technical_Deep_Dive
- ***NIST SP 800-207 – The Definitive Guide to Zero Trust Architecture, Terrazone.IO:*** https://terrazone.io/nist-sp-800-207/
- ***AI Risk Management Framework, NIST:*** https://www.nist.gov/itl/ai-risk-management-framework

Abbreviations

IAM: Identity Access Management	RBAC: Role-Based Access Control
AI: Artificial Intelligence	ABAC: Attribute-Based Access Control
JIT: Just in Time	CBAC: Context Based Access Control
PDP: Policy Decision Point	OIDC: OpenID Connect
PAP: Policy Administration Point	IP: Internet Protocol
PIP: Policy Information Point	OPA: Open Policy Agent
SSO: Single Sign-On	LLM: Large Language Model
OAuth: Open Authorization	GDPR: General Data Protection Regulation
TLS: Transport Layer Security	SOX: Sarbanes-Oxley Act
MFA: Multi-Factor Authentication	SIEM: Security Information and Event Management
PKI: Public Key Infrastructure	API: Application Programming Interface
TPM: Trusted Platform Module	HIPAA: Health Insurance Portability and Accountability Act

Appendix

NIST Standards to Consider

Planning, Identity and Access Control

- Standard: NIST SP 800-207: ZTA
- Principle: “Never trust, i lways verify” — continuous validation of identity and access
- Applies to: Human users, AI agents, and devices
- Standard: NIST AI RMF 1.0
- Recommends identity-centric controls for AI systems and service identities

Trust Evaluation and Policy Enforcement

- Standard: NIST SP 800-207
- Policy Decision Points (PDPs): Centralized enforcement of access policies
- Trust evaluation: Based on continuous signals and behavioral analytics
- Standard: NIST AI RMF 1.0
- Risk scoring: Behavioral analysis and anomaly detection for AI agents [1]
- Standard: NIST SP 800-207

Logging: Immutable, Auditable Logs for Access and Policy Enforcement

- Standard: NIST AI RMF 1.0
- Explainability: Transparency in AI decisions for audit and compliance

Workload Protection and Deployment

- Standard: NIST SP 800-207
- Container and API security: Runtime protection, workload isolation
- Standard: NIST AI RMF 1.0
- AI model integrity: Runtime scans, anti-poisoning, adversarial defenses

Continuous Monitoring

- Standard: NIST SP 800-207
- SIEM integration: Real-time monitoring and incident response
- Human oversight: Required for critical AI decisions
- Standard: NIST AI RMF 1.0 and NIST-AI-600-1 (Generative AI Profile)
- Governance: Risk-based controls, feedback loops, and regulatory alignment

LTIMindtree is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 700 clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by 86,000+ talented and entrepreneurial professionals across more than 40 countries, LTIMindtree — a Larsen & Toubro Group company — solves the most complex business challenges and delivers transformation at scale. For more information, please visit https://www.ltimindtree.com