xage SECURITY

Securing AI Systems with Zero Trust

Mapping Xage Capabilities to the MITRE ATLAS Framework

Artificial intelligence systems are increasingly embedded in enterprise platforms, operational environments, and critical infrastructure. As these systems expand in scope and connectivity to key data and systems, they introduce new attack surfaces that require enforceable, policy-driven security controls.

MITRE's Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) provides a comprehensive framework for understanding threats targeting AI systems. Similar in structure to MITRE ATT&CK, ATLAS catalogs adversary tactics and techniques specific to AI environments, including model access, execution abuse, privilege escalation, and data exfiltration.

ATLAS demonstrates how familiar adversarial behaviors apply directly to AI models, training pipelines, inference systems, and supporting infrastructure.

Mapping Xage to MITRE ATLAS

Xage demonstrates particularly strong coverage across the following tactic categories:

- **Initial Access**
- **AI Model Access**
- **Execution**
- **Privilege Escalation**
- **Credential Access**
- **Lateral Movement**
- **Collection**
- **Command and Control**
- **Exfiltration**

These categories span much of the attack chain, from initial compromise through system control and data extraction. Addressing these areas reduces risk across AI systems and the enterprise environments that support them.

xage SECURITY

Key AI Security Challenges Addressed by Xage

As organizations deploy AI models and autonomous agents into production environments, new risk scenarios emerge that extend beyond traditional IT threats. AI systems often operate with broad access to data, APIs, and infrastructure, making privilege control and containment essential. Securing these environments requires clear enforcement boundaries around how models and agents access, process, and transmit information.

Xage addresses several of the most critical AI-specific risk scenarios:

- **Containing Rogue or Compromised AI Agents:** Limiting or containing the impact of rogue agents or compromised large language models, preventing them from accessing or affecting systems beyond their authorized scope.
- **Preventing AI-Driven Privilege Escalation:** Stopping users from leveraging over-privileged LLMs or AI agents to escalate privileges or gain unauthorized access to systems and sensitive resources.
- **Blocking AI-Enabled Data Exfiltration:** Restricting unwanted data access and preventing potential data exfiltration through privileged AI components, APIs, or service accounts.

Overall Coverage

Xage's Zero Trust architecture aligns closely with the adversarial behaviors defined in MITRE ATLAS. **The platform provides coverage for more than two-thirds of the techniques identified in the framework**, with protections spanning a broad portion of the AI attack lifecycle from initial access through command and control and data exfiltration.

In particular, Xage delivers **comprehensive coverage across high-impact tactic categories including privilege escalation, credential access, lateral movement, collection, command and control, and exfiltration.** These areas represent some of the most operationally significant stages of AI system compromise, where controlling identity, access, and infrastructure movement is critical.

2/3

Overall technique coverage

Xage mapped its Zero Trust Access and Protection capabilities directly to the MITRE ATLAS matrix and developed a detailed heatmap linking each ATLAS technique to corresponding mitigations. Techniques are color coded to indicate coverage levels, with green representing full coverage and yellow representing supporting coverage.

xage SECURITY

Reconnaissance	Resource Development	Initial Access	AI Model Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Lateral Movement	Collection	AI Attack Staging	Command and Control	Exfiltration	Impact
Active Scanning	Acquire Infrastructure	AI Supply Chain Compromise	AI Model Inference API Access	AI Agent Clickbait	AI Agent Context Poisoning	AI Agent Tool Invocation	Corrupt AI Model	AI Agent Tool Credential Harvesting	Cloud Service Discovery	Phishing	AI Artifact Collection	Craft Adversarial Data	AI Service API	Exfiltration via AI Agent Tool Invocation	Cost Harvesting
Gather RAG-Indexed Targets	Acquire Public AI Artifacts	Drive-by Compromise	AI-Enabled Product or Service	AI Agent Tool Invocation	AI Agent Tool Data Poisoning	LLM Jailbreak	Delay Execution of LLM Instructions	Credentials from AI Agent Configuration	Discover AI Agent Configuration	Use Alternate Authentication Material	Data from AI Services	Create Proxy AI Model	Reverse Shell	Exfiltration via AI Inference API	Data Destruction via AI Agent Tool
Gather Victim Identity Information	Develop Capabilities	Evade AI Model	Full AI Model Access	Command and Scripting Interpreter	LLM Prompt Self-Replication	Valid Accounts	Evade AI Model	OS Credential Dumping	Discover AI Artifacts		Data from Information Repositories	Generate Deepfakes		Exfiltration via Cyber Means	Denial of AI Service
Search Application Repositories	Establish Accounts	Exploit Public-Facing Application	Physical Environment Access	LLM Prompt Injection	Manipulate AI Model		False RAG Entry Injection	RAG Credential Harvesting	Discover AI Model Family		Data from Local System	Generate Malicious Commands		Extract LLM System Prompt	Erode AI Model Integrity
Search Open AI Vulnerability Analysis	LLM Prompt Crafting	Phishing		User Execution	Modify AI Agent Configuration		Impersonation	Unsecured Credentials	Discover AI Model Ontology			Manipulate AI Model		LLM Data Leakage	Erode Dataset Integrity
Search Open Technical Databases	Obtain Capabilities	Prompt Infiltration via Public-Facing Application			Poison Training Data		LLM Jailbreak		Discover AI Model Outputs			Verify Attack		LLM Response Rendering	Evade AI Model
Search Open Websites/Domains	Poison Training Data	Valid Accounts			Prompt Infiltration via Public-Facing Application		LLM Prompt Obfuscation		Discover LLM Hallucinations						External Harms
Search Victim-Owned Websites	Publish Hallucinated Entities				RAG Poisoning		LLM Trusted Output Components Manipulation		Discover LLM System Information						Spamming AI System with Chaff Data
	Publish Poisoned Datasets						Manipulate User LLM Chat History		Process Discovery						
	Publish Poisoned Models						Masquerading								
	Retrieval Content Crafting						Virtualization/ Sandbox Evasion								
	Stage Capabilities														

Xage MITRE ATLAS Coverage

The MITRE ATLAS matrix defines adversarial tactics across the AI attack lifecycle. Xage addresses these techniques through a set of core Zero Trust capabilities that control access, limit privilege, segment infrastructure, protect data and continuously enforce policy across AI environments and the resources that AI connects to. The following sections outline how these capabilities align directly to high-impact ATLAS tactics.

1. Identity as the Primary Security Boundary: Xage enforces strict identity verification for every user, service, workload, and device requesting access to AI systems. Access is granted only after authentication and policy evaluation.

This model directly mitigates unauthorized model access and abuse of valid accounts within Initial Access, Credential Access, and AI Model Access tactics. By making identity the primary control boundary rather than network location, Xage reduces reliance on implicit trust. Xage protects identities, manages credentials, and access policies ensuring that only authorized LLMs, agents and users have access to data and resources. In addition, Xage layers phishing-resistant MFA and just-in-time context-based controls.

2. Enforcing Least Privilege in AI Environments: Authentication alone is not sufficient. Xage applies centrally managed least privilege policies that tightly control access to models, training datasets, modification capabilities, inference APIs, and resources including data and sensitive systems.

Each authenticated entity is limited to only what is required for its function. This reduces privilege escalation risk and minimizes unnecessary exposure of sensitive data or model artifacts. If an account or service is compromised, the operational impact is contained. With Xage, there are no standing privileges or unprotected credentials that attackers on AI systems could utilize. Credentials to data and resources are never exposed to AI Models and Agents, rather injected by Xage just-in-time with context-based privilege controls.

3. Jailbreak-Proof Protection for LLMs and Agentic AI: As enterprises deploy LLMs and autonomous AI agents, traditional guardrails alone are not enough. Xage enforces input and output guardrails to detect prompt injection, jailbreak attempts, and unsafe responses, along with retrieval filtering to prevent sensitive data from being exposed through manipulated RAG queries. This protects against common attack patterns designed to trick models into leaking data or taking unintended actions.

But Xage goes further. Instead of relying solely on language-based controls, Xage applies granular role, attribute, and privilege-based access enforcement directly to the underlying data, tools, and systems. Even if a model is manipulated through natural language, it cannot access resources or execute actions beyond explicitly authorized policies. By combining guardrails with identity-based, zero-trust access control for AI resources, Xage delivers comprehensive, jailbreak-resistant protection for LLMs and Agentic AI.

4. Microsegmentation Across AI Infrastructure: AI systems typically consist of distributed components such as model servers, training pipelines, storage systems and repositories, and APIs.

Xage applies Zero Trust microsegmentation across hybrid, multi-cloud, on-premises, and edge environments to isolate these components into secure zones. Communication between systems is explicitly defined by policy and all access is controlled via defined paths.

This limits adversary movement between AI components and helps contain compromised services. It also reduces the ability to pivot between enterprise IT systems and AI pipelines.

5. Continuous Visibility and Real-Time Enforcement: Xage continuously evaluates access decisions and enforces policy in real time. All access activity is logged for governance, and session visibility supports investigation when needed.

This visibility and enforcement model helps prevent suspicious model access patterns, unauthorized execution behavior, command and control activity, and data exfiltration attempts.

MITRE ATLAS Alignment and AI Security with Xage

MITRE ATLAS provides a standardized framework for evaluating adversarial threats to AI systems. Mapping security controls directly to ATLAS techniques enables organizations to identify coverage gaps, align protections to recognized threat categories, and communicate AI security posture in quantifiable terms.

Let's look at a common attack chain scenario:

Indirect Prompt Injection via RAG → Sensitive Data Exfiltration

As organizations integrate AI assistants with internal document repositories, SaaS platforms, email systems, and plugins, attackers exploit Retrieval-Augmented Generation (RAG) workflows by embedding hidden malicious instructions inside retrievable content. When the AI assistant pulls that content into its context window, the model unknowingly processes the attacker's instructions alongside legitimate data.

The injected instructions can override system intent and cause the model to access sensitive internal systems (e.g., email, file storage, APIs) and include confidential information in its output. Because RAG systems inherently trust retrieved content, this attack vector has been widely demonstrated in security research and represents a realistic enterprise risk.

Xage Protects at Each Step in the Attack Chain

MITRE ATLAS Tactic	What Attacker Does	How Xage Mitigates
Reconnaissance	Identifies organization uses AI assistant with RAG and integrations	Zero Trust access to AI endpoints; authenticated API access only; no anonymous probing
Resource Development	Crafts malicious document with hidden prompt injection instructions	External creation cannot be prevented, but ingestion can be controlled
Initial Access (AI data supply chain)	Uploads malicious document into retrievable repository	Identity-bound uploads; RBAC enforcement; XPAM-controlled admin access; credential vaulting prevents misuse of repository credentials
ML Model Access	Poisoned document retrieved into model context	Retrieval guardrails enforce document eligibility policies before inclusion in context - source, content, history
ML Attack Staging	Hidden instructions override system prompt	Input filtering strips instruction-like patterns; system prompt isolation; strict separation between retrieved content and model directives

MITRE ATLAS Tactic	What Attacker Does	How Xage Mitigates
Credential Access	Model induced to access email, file stores, or APIs	No standing privileges for AI service accounts; least-privilege access policies; Just-in-Time (JIT) access required for sensitive systems; ephemeral, short-lived tokens instead of static API keys; XPAM credential vaulting
Privilege Escalation	Attempts to leverage overprivileged service accounts or plugins	Elimination of shared credentials; workload identity enforcement; approval-based elevation; continuous session validation
Exfiltration	Sensitive data embedded in output or sent via plugin	Output filtering + DLP-like inspection block PII, secrets, credentials; policy-based egress controls; expiring tokens prevent persistent misuse
Persistence	Attempts to maintain long-term access via stored credentials	Automatic credential rotation; no static keys; ephemeral JIT tokens; revocation on policy change
Impact	Data breach, compliance violation, reputational damage	Segmentation limits blast radius; centralized audit logging and session tracing enable rapid detection and containment

With coverage for more than two-thirds of the techniques defined in MITRE ATLAS, Xage demonstrates measurable alignment to the AI threat landscape. The platform delivers comprehensive protection across high-impact tactic categories such as privilege escalation, credential access, lateral movement, collection, command and control, and exfiltration. This breadth and depth of coverage highlights the effectiveness of applying Zero Trust principles to secure AI environments.

Effective AI security depends on identity-centric access control, granular authorization and privilege management, jailbreak-proof protections, segmentation, and continuous visibility and monitoring. Xage delivers these capabilities through a distributed Zero Trust architecture designed to secure hybrid environments, including AI workloads operating across enterprise IT, cloud platforms, edge systems, and operational technology networks.

Applying Zero Trust principles to AI systems provides a consistent and enforceable approach to managing access and reducing exposure across the AI lifecycle.