

Security for AI workflows

Platform safeguards and security program | 29 September 2026

Triage builds Integrity, a platform for evaluating AI workflows against an organization's policies. Integrity connects policy, observed activity, review decisions, and supporting evidence so teams can understand how an AI system behaves and decide where intervention is appropriate.

The information passing through an AI workflow can include proprietary instructions, conversation context, model responses, and tool activity. Our security architecture protects that information through scoped authorization, encrypted content storage, controlled processing, and traceable operations.

Core protections

Area	How Triage protects the boundary
Access	Workspace membership and service authorization bind access to the relevant organization, project, deployment, and session.
Data protection	HTTPS protects hosted connections. AES-256-GCM envelopes protect saved provider credentials and retained session content.
Customer control	Customers select the workflow, governing policies, integration mode, and supported retention settings.
Operational security	Version-controlled changes, automated checks, vulnerability monitoring, and recorded remediation support release decisions.
Oversight	Oneleet maintains the SOC 2 Security control inventory, policy program, monitoring, and supporting evidence.

This overview is written for security, engineering, and procurement teams evaluating Triage. The companion Security Architecture and Data Handling document describes processing boundaries in detail. The Evaluation Plan provides a concrete route to assessing product value with a bounded dataset.

Document owner: Triage Security | Public customer information | Version 2.0
Security review contact: info@triage-sec.com

How Integrity fits your environment

Integrity receives the activity that a customer application sends through an enabled integration. The application can use the platform to observe policy judgments and, where the selected deployment supports it, apply an intervention before continuing a workflow. Triage does not obtain access to an internal repository, tool, or business system simply because a workspace has been created.

Separate administration from workflow processing

The control plane manages workspace membership, projects, deployment configuration, credentials, and evidence access. The workflow processing path authenticates service requests and evaluates the submitted activity. This separation allows administrative permissions and application credentials to serve different purposes.

Keep the governing context with the decision

Requests and retained evidence carry organization, project, deployment, and session context. Integrity uses the selected policy and runtime version when evaluating activity. Recorded outcomes distinguish completed review, unavailable review, and uncertain operations so a missing result is not presented as an approval.

Use the mode appropriate to the workflow

Mode	Customer experience
Observe	Review activity and inspect judgments while the application continues its normal workflow. This mode supports measurement and policy tuning.
Supported enforcement	The integration applies the selected intervention behavior. Tool execution still depends on the customer application's permissions and enforcement of the integration contract.
Evaluation workspace	A separate project uses public or synthetic examples and controlled tool fixtures to establish usefulness before connecting internal systems.

Customers retain control of their application's tools, cloud credentials, network permissions, and downstream effects. Integrity adds policy evaluation and evidence to that environment; the application's existing access controls remain in place.

Availability is specific to the enabled deployment and integration. Production use is subject to qualification of the proposed workload and the agreed deployment scope.

Protecting customer information

Workspace and service authorization

Human access is governed by authenticated workspace membership and the permissions needed for the requested operation. Machine requests use deployment credentials. The service checks the association between the credential, project, organization, and deployment; a record identifier alone does not authorize access to its contents.

Shared hosted services use logical tenant isolation. Authorization checks apply when information is requested, including evidence access and supported export workflows. Customer-model credentials and Integrity reviewer grants are separate from the platform credential used to admit workflow traffic.

Encryption and credential handling

Hosted traffic uses HTTPS. Saved provider credentials and supported retained session content use AES-256-GCM authenticated encryption. Protected session records are bound to their expected record context, helping prevent encrypted content from being substituted into another customer's record.

Runtime services obtain sensitive configuration through managed secret storage. Secrets are excluded from public release artifacts and ordinary customer evidence. Production database connections use certificate-verifying TLS.

Purpose and retention

Triage processes Customer Data to provide the configured service under the customer agreement and documented instructions. Those terms do not authorize unrelated model training or cross-customer reuse. Customer-specific adaptation, when supported and authorized, has a separate preparation, evaluation, and activation lifecycle.

Retention settings govern how long supported evidence remains available in the application. End-of-service handling includes the relevant application records, runtime storage, connected providers, and backups. A customer review establishes the applicable retention and cleanup requirements for each layer.

The architecture guide includes the data inventory, encryption scope, provider roles, and retention boundaries used in a technical review.

Operating and maintaining the service

Controlled changes

Code changes are managed through version control and pull requests. Automated workflows exercise application behavior, tenant boundaries, SDKs, web components, and container packaging. Release images are built from identified source revisions and checked before publication. Deployment verification establishes which revision is actually serving traffic.

Vulnerability management

Triage uses source, dependency, container, and infrastructure monitoring to identify security issues. Findings are evaluated for affected components, exposure, severity, and remediation. A fix is tracked through review and release verification; internal review is complemented by the scheduled independent penetration test.

Monitoring and service continuity

Operational monitoring covers service availability, dependencies, and key workflow paths. Synthetic canaries verify that a representative request can complete and that its outcome is retained. These checks provide evidence beyond an isolated health endpoint.

The hosted environment uses automated database backups and encrypted storage snapshots. Runtime state backup jobs have completion monitoring and failure alarms. Recovery planning considers application state, databases, runtime evidence, credentials, and the versions required to read restored data.

Incident handling

Our incident response policy defines investigation, containment, recovery, and follow-up responsibilities. Security concerns are received at info@triage-sec.com. Customer notification and cooperation obligations are established in the applicable agreement; security evidence can be exchanged through an agreed confidential channel.

People and supplier controls

The security program includes employee training, device monitoring, access reviews, documented policies, and vendor risk assessment. Oneleet provides the control inventory and evidence workflow. Program reviews track implementation and outstanding work against the systems and people in scope.

Assurance and customer review

Triage organizes its security program around the SOC 2 Security criteria and uses Oneleet to manage controls, policies, monitoring, and evidence. This provides a structured basis for customer diligence and independent assessment.

Assessment status

Assessment	Status on 29 September 2026
SOC 2 Type II	Readiness and evidence review are in progress. An auditor-issued Triage report is not yet available.
Independent penetration test	Scheduled. The report will be available for review after testing and the applicable disclosure process.
Customer technical review	The security overview, architecture and data-handling guide, and evaluation plan are available now. Additional evidence is reviewed against the proposed deployment.

A SOC 2 observation period and report date will be communicated once confirmed with the assessment team. Infrastructure-provider reports are considered during supplier review and are separate from Triage's own assessment.

Information available for diligence

- Service architecture, processing routes, and the providers relevant to the proposed deployment.
- Authorization boundaries, encryption, retention, and customer responsibilities.
- Applicable policies, control evidence, and security remediation relevant to the intended use.
- Deployment scope, qualification evidence, incident cooperation, and contractual data-protection terms.

For an internal integration, Triage and the customer agree the permitted data, participants, model routes, retention, and required security evidence. Any material open item relevant to that deployment is addressed through that review.

Send a questionnaire or review request to info@trriage-sec.com. Current materials are published at trust.trriage-sec.com. Detailed operating evidence can be shared under appropriate confidentiality arrangements.

Customer control and next steps

Responsibilities at the integration boundary

Party	Responsibility
Triage	Operate the agreed service, enforce platform authorization, protect stored credentials and evidence, maintain security monitoring, and provide relevant review materials.
Customer	Authorize submitted data, manage workspace access, protect integration keys, configure least-privilege tools, and select appropriate application safeguards.
Joint review	Confirm the data flow, providers, permitted use, retention, incident contacts, workload qualification, and acceptance criteria.

A practical evaluation path

Start with a representative workflow and a fixed set of public or synthetic cases. A coding assistant can be evaluated on preserving authorization requirements while completing a bug fix. A support assistant can be evaluated on following an approved support policy while interpreting retrieved content. Both produce concrete evidence of legitimate task completion, policy adherence, intervention rate, latency, and available usage.

The Evaluation Plan describes these workflows and the artifacts to retain. Results from the agreed cases guide the next integration decision. Customer data and internal systems are introduced only within the approved scope.

Document set

- Security Overview: the platform safeguards, operating program, and assessment status.
- Security Architecture and Data Handling: technical boundaries, data categories, providers, and lifecycle controls.
- Evaluation Plan: representative workflows, measurement criteria, and a structured evaluation readout.

Contact

Trust Center: <https://trust.triage-sec.com>
Security and diligence: info@triage-sec.com
Technical documentation: <https://triage-sec.com/docs>

This document describes the service as of its publication date. Deployment-specific commitments and processing instructions are recorded in the applicable customer agreement and data processing terms.