TECHNICAL DATA SHEET / TECHNOLOGY PREVIEW

CLAWOLF AIS

AI & Agent Security
AI-domain security, trust and runtime governance - independently operable, natively connected.

AIS is designed to protect AI models, applications, agents, retrieval, memory, tools and connected resources. Its governing chain connects accountable identity and purpose to permitted action, independently observed effect and recovery evidence. Standalone AIS owns AI-domain analysis and protection; AS-OS adds institutional context and coordinated enterprise response.

At a glance	Technical product profile
Protection scope	AI discovery and posture; identity/delegation; prompt and multimodal inspection; RAG/memory; MCP/tools; model supply chain; data/output and governed response.
Trust mechanisms	Causal evidence integrity, semantic effect assurance, joint-outcome defense, causal decontamination, mission preservation and change-bound assurance.
Adaptive protection	CLAWOLF Immunity: reproduce, generate, qualify, approve, verify and recover guardrail changes.
Adoption	Independent AIS installation, UI and licence; optional authenticated native AS-OS integration.
Deployment design	Air-gapped and connected on-premises, VMware Kubernetes private cloud and customer-account EKS/AKS/GKE.
Product status	In development and controlled validation. Not generally available or certified for customer production. Target release: 30 November 2026, subject to acceptance.

This sheet describes the committed product design and identifies the narrower current validation scope. Design coverage is not a claim that every capability is released. Reader map: architecture 2 | estate 3 | content 4 | tools/data 5 | signatures 6 | Immunity 7 | governance 8 | response 9 | integration 10 | deployment 11 | operations 12 | evaluation 13.

01 / PRODUCT ARCHITECTURE

One AI-domain authority. Optional enterprise reach.

AIS protects the AI domain without requiring an AS-OS runtime, database, service or licence.

IDENTITY	AUTHORITY	ACTION	EFFECT	EVIDENCE

Functional boundary	AIS responsibility
Estate and reachability	Attributable AI assets, dependency versions, effective permissions, resources and trust boundaries.
Activity inspection	Prompts, retrieved content, memory, agent messages, tool definitions/arguments/results, artifacts and resource use.
Mission integrity	Relate task purpose and delegated authority to observed behavior and cumulative outcomes.
Bounded enforcement	Apply only supported controls through explicitly enrolled bindings; expose bypass and observation limits.
Assurance	Adversarial qualification, guardrail engineering and invalidation when relevant versions or permissions change.
Evidence and integration	Retain origin/derivation, authenticated findings and decision/action/verification lineage.

Adoption mode	Authority and value
Standalone AIS	AI-domain policy, analysis, approvals, authorized enforcement, verification and recovery. It must remain useful without the rest of the portfolio.
AIS + AS-OS	AIS contributes AI evidence and scoped control options. AS-OS governs the separately assigned enterprise response with institutional policy, Decision DNA and validation context.
AIS + CNDA	Separately adopted domain products can add identity, data, behavior, insider, supply-chain and cloud context. Other CNDA products remain roadmap.

Every consequential action has one assigned authority and one executor. Integration does not create a second decision maker, automatic licence or permission to act outside the enrolled scope.

Discover what acts, who owns it and what it may change

Discovery is the entry point; protection depends on attributable inventory and current execution authority.

Capability	Design specification	Operational significance
Estate discovery	Inventory models, apps, agents, MCP servers, tools, skills, dependencies and ownership; retain source, observation time and versions.	Separate observed assets from inferred or stale records. Broad automatic shadow-AI discovery remains a design scope, not a released connector claim.
Posture and reachability	Assess configuration, permissions, reachable resources, dependencies and attack paths.	Distinguish a possible path from an observed exploit or confirmed exposure.
Agent identity	Associate human sponsors, agents, services and workloads with their effective identity and credentials.	A name supplied in a prompt is not authenticated identity.
Delegation	Bind actions to scope, target and validity; handle expiry, suspension, revocation and changed principals.	Prevent a handoff from silently broadening authority.
AI bill of materials	Retain model, tool, component and dependency versions with their provenance.	Invalidate affected assurance when a component changes.

Current implementation footprint

The candidate contains tenant-bound AI-agent registration, enrolled source identities, authority delegation, typed findings, inventory/evidence surfaces and signed response contracts. That footprint does not prove every discovery source or every AI platform is integrated.

Required exposure states

Configured, observed, stale, offline, unsupported and unavailable coverage must remain distinguishable. Absence of a source observation cannot be presented as a clean or fully discovered AI estate.

Protection applies to enrolled traffic, assets and actions that traverse the supported inspection/enforcement boundary. Out-of-band activity and unsupported platforms require separate visibility or integration.

03 / RUNTIME CONTENT PROTECTION

Inspect the content an agent actually consumes

Direct and indirect instructions may arrive through documents, retrieved text, tool responses or non-text media. A trusted file type does not make an instruction trustworthy.

Input family	Current candidate inspection path	Limit / interpretation
Text and structured content	UTF-8 plain text, Markdown, JSON, YAML and HTML. HTML extraction inspects comments/attributes as well as visible text.	Bounded parsing; malformed/oversized inputs fail or are rejected.
Documents	PDF and Office Open XML documents, presentations and spreadsheets.	Pinned extraction dependencies; declared format/size limits and isolated execution.
Images	PNG, JPEG and TIFF using pinned OCR and language data.	Text extraction, not a general claim of complete visual-semantic understanding.
Audio / video	WAV and MPEG audio; MP4 audio transcription using pinned local assets.	Transcript inspection, not comprehensive interpretation of every sound or video frame.
Code / tool results	Text/code and supported structured tool output can be inspected for policy-relevant instructions and sensitive data.	Inspection does not authorize executing arbitrary uploaded code.
Retrieval / memory	Maintain source, version, access scope and derivation context as content enters subsequent tasks.	A retrieval result is not evidence of authorized downstream use.

Policy outcomes

The implementation distinguishes allow, allow with sanitized content, deny, require approval and unavailable. Task continuation is separately recorded as continue, continue with sanitized content, stop or not applicable.

Detection scope

Current analysis includes deterministic policy/pattern checks and pinned tools. It is not a claim of universal semantic detection, complete multilingual robustness or immunity to novel frontier attacks. Supported parsing does not itself establish attack-detection efficacy.

Constrain the path from an allowed call to its consequences

Tool names and successful replies are insufficient. Bind the actual server, schema, principal, target and downstream effect.

Control area	Technical design / candidate contract
MCP enrollment	Bind server name/version/identity, tool name, tool-schema/catalog/capability digests and current authority. Changes invalidate the affected binding.
MCP transport	Candidate adapters include allowlisted stdio and enrolled streamable HTTP. Remote MCP uses an exact HTTPS /mcp endpoint and a tenant-owned credential reference.
HTTP tools	Constrained HTTPS origin, exact paths, allowed methods, approved resolved addresses, deadlines and effect destinations. Mutations carry an idempotency key.
Remote authentication	Candidate contracts support API-key references, mTLS and OAuth2 client credentials with scoped trust/issuer/audience inputs. Live-provider compatibility is deployment specific.
Supply-chain inspection	Pinned ModelScan, Syft, Grype and YARA paths retain artifact/tool/rules or vulnerability-database identity and genuine report evidence.
Data / output	Evaluate secrets, sensitive records, destination context and approved transmission; distinguish masking, denial, approval and authorized use.
RAG / memory	Retain source lineage, tenant/task scope and access; quarantine contaminated derivatives and do not reactivate stale context after restart.

Effect-aware protection

An allowed read must not silently become an unauthorized export or mutation. AIS separates inspection evidence, authorized execution and independently observed postconditions. Unobservable effects remain unverified.

The source contains supported binding contracts and controlled paths; it does not establish native integration with every MCP implementation, model repository, database, SaaS application or AI-provider service.

05 / SIGNATURE TRUST MECHANISMS

Protect the mission and its evidence

Six signature mechanisms define the AIS trust architecture. Their design scope is broader than the currently accepted controlled cases.

Mechanism	Technical behavior	Example outcome
Causal Evidence Integrity	Trace origin and derivation across retrieval, copies, summaries and messages. Do not treat repeated derivatives as independent corroboration.	Five agents quoting one poisoned document still contribute one origin.
Semantic Effect Assurance	Compare permitted target, operation and data scope with separately observed consequences.	A tool reply marked successful cannot conceal an unauthorized downstream effect.
Joint-Outcome Defense	Assess cumulative access, authority and effects across related agents, branches and tasks.	Individually permitted reads cannot reset a shared disclosure or resource boundary.
Causal Decontamination	Trace/quarantine suspect memory and derivatives while preserving unaffected context and unresolved scope.	Removing an original document does not leave its poisoned summaries active.
Mission-Preserving Defense	Continue only the safe part of a task under existing authority and trustworthy dependencies. Stop or escalate otherwise.	Contain the dangerous branch without granting new privileges to finish the task.
Continuous Assurance Invalidation	Bind assurance to model, tool, permission, policy and context versions; recheck affected evidence after change.	An older approval cannot certify a changed agent or tool version.

What the current evidence establishes

Controlled cases exercise lineage, scoped effects, response verification, selected branch/authority controls and guardrail qualification. They do not prove complete multi-agent semantic reasoning, all autonomous threat classes or every third-party action.

Causal cleanup concerns context and evidence derivatives. It does not reverse an already executed payment, data disclosure or other external transaction; remediation requires separate supported authority and controls.

06 / CLAWOLF IMMUNITY

Qualified protection instead of unchecked self-modification

Immunity connects reproducible weaknesses to tested guardrail candidates, authorized promotion and verifiable recovery.

Stage	Required behavior	Retained artifact
Discover and reproduce	Identify a genuine weakness within authorized scope; preserve versions and original observations.	Finding, attack inputs, target and evidence references.
Generate	Derive a constrained guardrail candidate from the finding. Generation cannot authorize activation.	Candidate rule/template and provenance.
Qualify	Test the original attack, held-out attacks and legitimate work against the intended scope.	Qualification result, coverage, limits and clean-task regressions.
Approve and promote	Bind current policy/approval and delegated authority to the exact candidate and target.	Approved version and activation intent.
Verify and recover	Observe the protective effect independently; roll back to the approved prior version when required.	Observed effect, verification and rollback result.
Refresh assurance	Invalidate evidence affected by a relevant change; preserve finding-to-protection lineage.	Current assurance revision and remaining gaps.

Implementation model

The present candidate uses deterministic guardrail templates, bounded qualification jobs and explicit publish/rollback controls. That is a governed engineering loop, not a claim that an unconstrained model can rewrite security policy or deploy arbitrary defenses.

Scope of improvement

A passed qualification applies to its exact target, versions, declared cases and legitimate-task baseline. Novel attacks, unseen applications and changed permissions can invalidate that conclusion. Failed or incomplete qualification cannot promote a rule as verified protection.

The broad Immunity vision includes future application areas. Robotics, quantum systems, embedded manufacturing and universal hyperscaler infrastructure protection are not delivered compatibility claims in this preview.

Make governance operational across the AI lifecycle

Ownership, risk, trust and policy remain connected to actual runtime control, evidence and retirement.

Governance responsibility	Product design / candidate behavior	Decision enabled
Ownership and purpose	Register use cases, accountable owners, AI assets, intended purpose and risk boundaries.	Who is accountable, and what was the AI authorized to do?
Risk, approval and exceptions	Record assessment, approver separation and time-bound exceptions; review after relevant changes.	Is the current operation permitted under the organization decision?
Trustworthiness	Retain technical assessments, evidence, limits and distinct author/approver records.	What supports trust for this exact system version?
Policy-to-control mapping	Connect applicability, approved packs and control evidence to enforcement/monitoring, with explicit gaps.	Which obligations have operating controls and which remain unfulfilled?
Change and retirement	Track changed assets/policies, invalidate stale assurance and govern retention, hold, disposition and retirement.	What must be re-evaluated, retained or safely decommissioned?

Regional context

The candidate includes organization/jurisdiction facts, source lifecycle and applicability handling for US, EU, Middle East/GCC and Türkiye policy contexts. Stale, future, superseded, restricted or unsupported content must not appear as current effective enforcement.

Framework alignment

The design relates ownership and lifecycle to NIST AI RMF GOVERN, MAP, MEASURE and MANAGE, and relates threat/assurance work to OWASP agentic risks and MITRE ATLAS. These are design mappings, not third-party certification or full legal conformity.

Governance, identity and lifecycle work remains under final acceptance. A mapped obligation, uploaded policy or evidence record alone cannot establish that the required control is implemented and effective.

08 / GOVERNED RESPONSE

Contain within scope. Verify the result.

The current action catalog is intentionally explicit. Controls apply only to enrolled bindings and supported postconditions.

Action family	Candidate control	Completion boundary
Prompt protection	Enable the permitted prompt guard.	Active guard/version and resulting probe evidence, not only a request receipt.
Tool access	Suspend an enrolled tool binding.	Observe the disabled binding and denied invocation.
HTTP destination	Block the scoped egress destination.	Observe the changed policy and denial while the unaffected destination remains distinct.
Session / delegation	Revoke a session or suspend delegated authority.	Subsequent authorized use must reflect the revoked/suspended state.
Memory lineage	Quarantine the specified lineage and dependent context.	Retain affected/unaffected records and unresolved cleanup scope.
Task branch	Suspend the admitted branch.	Observe the branch gate; no unrelated branch may inherit the action.
Guardrail lifecycle	Publish or roll back an approved guardrail version.	Independent observation/probe of exact version and protective result.

Separation of duties

The runtime separates application admission, least-privilege execution, state observation and independent probes. An executor must not write its own independent verification. A probe does not mutate protected effect state.

Safe failure

Expired delegation, missing entitlement, stale policy, unavailable evidence, unsupported binding, failed execution or absent postcondition stays blocked, failed or unverified. A UI action or accepted command cannot manufacture a successful security outcome.

A customer may configure monitoring without enforcement. The interface must expose that distinction; visibility is not blocking coverage. Broad third-party containment and kill-switch behavior require an explicitly supported target contract.

09 / INTEGRATION & COMBINED VALUE

Connect AI protection to enterprise operations

Use AIS independently or add an authenticated, versioned native boundary to AS-OS. Integration extends context without duplicating execution authority.

Integration path	Technical behavior	Customer contribution
AI source admission	Enrolled producer identity, typed findings, evidence references, tenant/asset/task context and durable record.	Attributable findings rather than anonymous activity labels.
Tool interaction	Supported MCP and constrained HTTPS bindings retain actual request/result identities.	Connect declared authority to the interaction being inspected.
Native publisher	Versioned authenticated messages, durable outbox, replay/identity controls and acknowledgment/recovery.	Keep a recoverable boundary between independent products.
AS-OS decision path	AIS evidence enters institutional context, decision, policy/HITL and assigned response.	Coordinate AI risk with the wider security operation.
Action and observation	Signed assigned action, distinct executor/observer/probe and rollback evidence.	Follow authority through to the actual protective effect.
Evidence/reporting	Link selected findings, current outcomes and governed exports to tenant and decision identity.	Review the same chain across operations and assurance.

The product distinction

AIS protects AI and agent activity. AIS + AS-OS connects that protection to institution-wide governed, coordinated and outcome-proven security operations. AS-OS contributes enterprise context, Trust Engine, Validation Fabric and Decision DNA; AIS does not outsource its basic AI-domain analysis or protection.

CNDA roadmap contribution

Identity Defense Intelligence, Data & Database Intelligence, Behavioural Intelligence, Insider Defense Intelligence, Supply Chain Exposure, and Cloud Exposure & Attack In Motion Intelligence are separately adoptable roadmap products. Their future context is not present-day AIS coverage.

Controlled native-stage acceptance is not whole-product, live-provider, release or every-connector acceptance.

10 / DEPLOYMENT & CUSTODY

Four deployment targets. Independent customer control.

These are AIS delivery-design targets. AS-OS R5 package acceptance does not certify the separate AIS product or its deployment media.

Deployment target	Intended topology	Connectivity boundary
Air-gapped on-premises	AIS application, database, authoritative evidence store, trust material and analysis dependencies inside the customer boundary.	Offline operating path; controlled update and licence-artifact transfer.
Connected on-premises	Customer-owned Kubernetes, data services and enrolled integrations.	Approved egress only; optional customer-owned AI credentials.
VMware private cloud	Customer-managed Kubernetes on VMware infrastructure.	Customer controls ingress, storage, certificates and internal service paths.
BYOC	Customer cloud account using AWS EKS, Azure AKS or Google GKE.	Customer owns infrastructure/data plane; no required CLAWOLF central control plane.

Architecture requirement	Technical boundary
Authoritative evidence	S3-compatible, signed-request object path with customer-managed encryption-key verification; PostgreSQL retains scoped metadata/lineage.
No local evidence fallback	Missing/unreachable authoritative storage blocks readiness or affected work. A local/PVC copy cannot silently become the authoritative evidence store.
Untrusted analysis	Separate digest-pinned OCI sandbox; no network, read-only root, dropped capabilities, non-root identity and bounded resources.
Native integration	Optional authenticated AS-OS boundary; independent AIS source, UI, identity, licence and storage responsibilities.

No managed-SaaS availability, cloud Marketplace listing, universal S3-provider compatibility, VMware/hyperscaler certification, OpenShift support or automatic multi-cloud DR is asserted. Final four-profile delivery qualification remains pending.

Technical controls and operating limits

Current values describe the controlled candidate, not a contractual production capacity or final compatibility matrix.

Area	Candidate specification
Application / persistence	TypeScript/Node application and operator UI; PostgreSQL 16 validation environment; separate migration, application, executor, observer and probe identities.
Local identity	scrypt password derivation; AES-256-GCM protection of TOTP secrets under an external key; atomically consumed TOTP counters for login/step-up.
Enterprise identity	Authorization-code OIDC with PKCE, state/nonce, issuer/audience, signature/time and one-use transaction checks. Required OIDC disables local fallback for that tenant.
Object integrity	Canonical retrieval verifies SHA-256 selected-object bytes. Reachable storage or a catalog row does not prove a specific object is available or intact.
Object encryption	Writes require the expected customer-managed encryption-key identity and aws:kms response semantics. An incompatible store must fail closed.
Analysis isolation	Current per-job sandbox: 2 CPU, 3 GiB memory, 32 processes, no network, read-only root and 128 MiB temporary filesystem.
Timeout ceilings	Current outer bounds: ordinary text 10 s; PDF/OOXML 120 s; audio/supply-chain 600 s; MP4 900 s. These are ceilings, not promised latency.
Readiness	Database, authentication/runtime-role dependencies and signed authoritative-store probe must satisfy readiness. Object-specific health is checked at retrieval/execution.

Operations to qualify

Customer topology must establish backups, restore with key custody, restart/recovery, retention/hold/disposition, upgrade/rollback, certificate/key rotation, concurrency, endurance and monitoring. The independent AIS EV/SA/OR and target-class delivery acceptance remain pending.

Live IdP, storage, model-provider and customer connector compatibility is established with the customer configuration at deployment. No universal rate, capacity, availability, RPO or RTO is claimed.

A clear boundary between design and availability

This technology-preview sheet supports technical evaluation. It is not an announcement of general availability or a replacement for the release-specific support matrix.

Capability group	Current evidence boundary
Analysis and inspection	Bounded controlled execution exists for selected content, scanner, tool and policy paths. No universal frontier-attack efficacy claim.
Response and evidence	Controlled scoped actions, persistent findings, authoritative evidence retrieval and independent effect/rollback paths have focused acceptance.
Native AS-OS integration	A bounded genuine controlled two-product journey has independent stage acceptance. Other products/providers remain separately qualified.
Governance / identity / lifecycle	Candidate implementation exists; complete independent final acceptance is still pending.
Remaining whole-product scope	Broad discovery, all committed detection/control coverage, operational target-class acceptance and four-profile customer delivery qualification remain in progress.
General availability	Not released. Owner-directed target: 30 November 2026; subject to completed product, assurance and delivery acceptance.

Licence and data lifecycle

AIS has its own signed tenant/installation-bound licence. Candidate checks cover product/version, increasing sequence, time, revocation and safe hold. Protected mutations require current entitlement; reads remain available during safe hold. Manufacturer private signing keys are not delivered to customers or resellers.

Evaluation materials

Request the current capability/format/binding matrix, exact-version evidence, data-flow and threat model, responsibility matrix and deployment bill of materials. Agree the customer workload, supported providers, evidence location, retention and success criteria before activation.

Framework references: NIST AI RMF 1.0 (nist.gov/itl/ai-risk-management-framework); OWASP GenAI Security Project (genai.owasp.org); MITRE ATLAS (atlas.mitre.org). Design mapping does not establish certification or legal compliance. Contact enterprise@clawolf.com / clawolf.com.