

03 - AGI Containment Architecture Specification

Candidate architecture v0.1 | Specification, not deployed infrastructure

Design basis and evidence levels

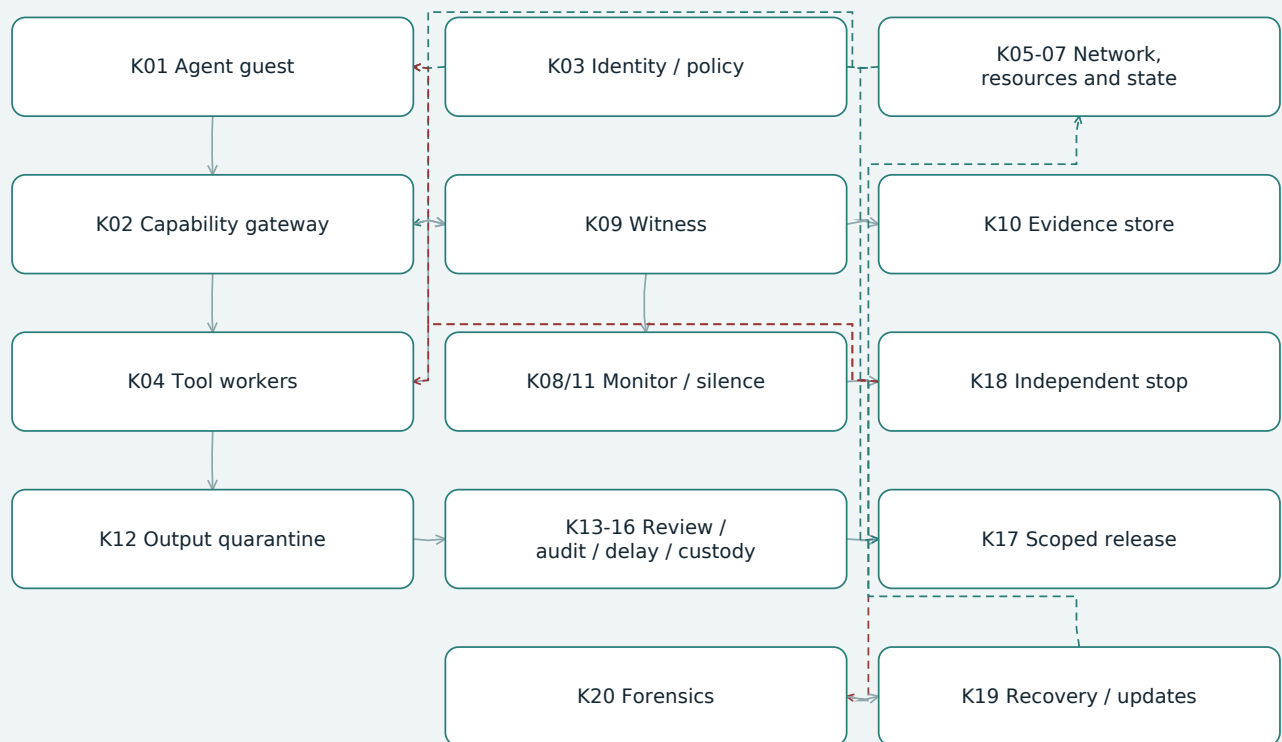
The baseline denies all external machine action. A contained runtime may request a small set of brokered operations on synthetic assets. Every request is untrusted, independent authorization precedes any effect, and evidence is written before and after admitted actions. Output is quarantined; it is not a callback into the control plane. Controls do not rely on a model describing its goals truthfully.

Three implementation levels are kept separate. L0 is the executable deterministic protocol model supplied here. L1 is an actual Linux process termination probe supplied here, not an arbitrary-code sandbox. L2 is the separately hosted VM/telemetry architecture specified below, requiring implementation and commissioning. L3 is physical facility qualification and independent human/organizational testing. L0 passes cannot be promoted to L2/L3 assurances.

Trust zones and system schematic

Z0 is the untrusted agent guest. Z1 contains disposable tool and parsing workers. Z2 contains the capability, authorization and state-control plane. Z3 is the independent witness and termination authority, preferably a separate management host and power domain. Z4 is dirty output quarantine, offline audit and review. Z5 is clean release/governance and recovery roots. The entire operational laboratory has no internet service. Import/export use explicitly inventoried physical transfer procedures. An eventual internal unidirectional transfer device is a separate engineering option, not permission for an external live optical link.

Proposed trust domains and permitted logical paths



Solid: data/evidence. Dashed: control. Red: stop. K21 facility assurance encloses all domains.

The arrows are permitted logical interfaces, not a serial list of magic protections. Logging and stop paths are parallel to task execution. The arrow from recovery approval permits only a fresh verified run; it does not resume a suspect guest. Detailed boundaries and attack paths are B01-B22 in 02. Network isolation encloses Z0 and Z1; these nodes have no external route. L0 implements logical objects, not these independent hosts.

Normative action and evidence interfaces

All proposed L2 IPC uses length-prefixed UTF-8 JSON, at most 4096 bytes per request, no duplicate keys, exactly one schema version, bounded nesting (4), explicit method allowlist and no implicit type coercion. Unknown fields, non-finite numbers, bool-as-int confusion and invalid UTF-8 fail closed. No shell commands, user-supplied modules, URLs or host paths are evaluated. This framing contract exceeds the L0 dictionary API, which validates the bounded fields it receives; wire-parser tests remain L2 work.

Action envelope: schema_version, supervisor-bound run_id, monotonic request_id, policy_sha256, capability_id, method and typed arguments. Agent timestamps are advisory only. Authentication comes from the control-plane channel and issuer, not fields called operator or role in model output. Capabilities are nontransferable outside their run/scope, expire by trusted time, and are consumed once where necessary. The L0 harness models identities with trusted method arguments; it does not implement production credentials.

Interface	Allowed input	Effect and limit	Failure behavior
fs.write	Absolute virtual path under /scratch/, UTF-8 string <=1024 bytes	Virtual ephemeral file; no symlinks or host mounts in L0	Canonical path outside root, quota or wrong type: deny + latched hold

Interface	Allowed input	Effect and limit	Failure behavior
memory.put	Key <=64 bytes, value <=1024 bytes	Run-local store only	Cross-run/persistent target denied
compute.allocate	Integer units 1-10	Aggregate simulated allocation <=10; no reset per request	Additional allocation denied; record budget state
process.spawn	Predefined synthetic job name	At most 2 simulated workers	Unknown program or limit exceeded denied
api.call	Service calculator, operation add, two finite bounded numbers	Deterministic arithmetic only; no service credentials returned	Other service/action, invalid type or range denied
network.send	Synthetic destination and payload	Always denied in baseline	Record attempt; no host socket or external call
output.submit	Plain UTF-8 text <=1024 bytes	Store quarantined specimen and digest; never execute	Metadata failure, changed specimen or unsafe syntax blocks release
report.submit	Ordered list of action request IDs and declared action names	Reconcile to collector event IDs, including denials	Omission, duplicate, false declaration or timeout causes hold
release	Controller-only gate receipts	Same digest, run and policy; all reviewers distinct; trusted delay complete	Agent calls always denied; missing/dissenting receipts keep quarantine
stop / restart	Control-plane-only authenticated command	Latch stop; restart creates new run only after external approval	Agent cannot reverse; invalid command is logged

Evidence event schema: schema_version, run_id, policy_version, collector_sequence, receipt_monotonic_ns, origin, request_id, event_type, bounded target digest, decision, reason code, effect digest, previous hash and event hash. Agent reports are stored under origin agent_declaration, never host. Sensitive payloads stay in quarantine; logs store digests and scoped metadata. An external collector is also a possible output channel: it must not display/execute untrusted text or export raw payloads as operational logs.

Before an admitted action: validate state, identity, method, budget, canonical target and log intent. Obtain the necessary witness receipt; if unavailable, deny without effect. Execute a bounded worker action. Record its receipt and compare an independent effect observation. A crash between intent and effect is UNKNOWN, not automatically successful or rolled back. Non-idempotent effects require unique action IDs and idempotency receipts. In L0 execution is synchronous and single-threaded; crash atomicity and concurrency are not established.

Permissions and trust matrix

Agent guest

Task data: Run input/scratch only

Tool capability: Request scoped method

Policy/keys: No access

Evidence: Submit untrusted report

Stop: Cannot veto

Release/restart: None

Tool worker

Task data: Declared object only

Tool capability: One action

Policy/keys: No issuer key

Evidence: Append own receipt

Stop: None

Release/restart: None

Gateway

Task data: Action metadata

Tool capability: Admit under policy

Policy/keys: Verify, not change

Evidence: Append intent/decision

Stop: Request stop

Release/restart: None

Witness

Task data: Bounded events

Tool capability: None

Policy/keys: Own signing/checkpoint root

Evidence: Append/anchor, no rewrite

Stop: Request stop

Release/restart: No unilateral release

Reviewer

Task data: Quarantined specimen

Tool capability: None

Policy/keys: Personal approval identity

Evidence: Read reviewed evidence

Stop: Halt allowed

Release/restart: Endorse, never alone

Stop authority

Task data: Enrolled resource IDs

Tool capability: Revoke/suspend/kill

Policy/keys: Stop credentials only

Evidence: Append stop receipt

Stop: Yes

Release/restart: Cannot restart

Release/restart authority

Task data: Approved hashes

Tool capability: None

Policy/keys: Dedicated keys

Evidence: Append authorization

Stop: Halt allowed

Release/restart: Distinct approvals required

Forensic analyst

Task data: Sealed read-only copy

Tool capability: Isolated analysis only

Policy/keys: No production credential

Evidence: Append findings

Stop: None

Release/restart: Recommendation only

No principal should combine agent execution, policy signing, evidence retention administration and restart approval. Real administrative separation is an L2 requirement; labels within one process do not create it.

Initial numerical settings

These are laboratory parameters for falsification, not validated safety thresholds: 100 requests/run; 4096-byte request; 1024-byte text/value; 10 simulated compute units; 2 simulated workers; 5 seconds to reconcile reports after a completed action window; independent heartbeat every 1 second with a stop trigger after more than 2 seconds without receipt; 30-second simulated minimum output quarantine for tests. Use a fake monotonic clock for deterministic timing cases. L2 proposed stop targets: revoke broker admission within 250 ms of a critical trigger; establish host/VM termination within 2 seconds. A real risk assessment may require tighter limits or complete prohibition of the action. The separate H01 process test uses 100 ms TERM grace and a 2-second KILL wait; it does not test the L2 targets end to end.

Finite delay is merely one gate. The system never labels an artifact safe solely because 30 seconds or years passed. No release for arbitrary executables is enabled in the initial profile. Permitted inert text retains semantic influence; test T25 deliberately exposes this limitation.

Component contracts

K01 - Inference runtime / boundary

- **Purpose:** Run the approved inference image in a disposable guest.
- **Inputs:** Signed image, task manifest and bounded input bundle.
- **Outputs:** Tokens/tool envelopes and quarantined output.
- **Trust level and required permissions:** Untrusted guest; no host mount, root socket, host key or external NIC.
- **Failure modes:** Guest/driver escape, inherited descriptors, mutable weights.
- **Attack surface:** Model loader, tokenizer, virtual devices, libraries.
- **Dependencies:** Host/hypervisor, IOMMU/device policy and image supply chain.
- **If compromised:** Assume all guest data hostile; host compromise invalidates same-host telemetry.
- **If unavailable:** Do not start; close run and quarantine output.

- **Agent influence:** Agent controls content and may control guest code in A2 tests.
- **Independent verification:** Cannot attest its own safe behavior; host effect probes required.

K02 - Capability gateway

- **Purpose:** Completely mediate the documented tools.
- **Inputs:** Bounded request frame plus channel-bound run identity.
- **Outputs:** Deny receipt or typed action intent.
- **Trust level and required permissions:** Trusted control service; dispatch only, no arbitrary shell.
- **Failure modes:** Parser bug, confused deputy, race, input size exhaustion.
- **Attack surface:** JSON decoder, schema, framing, IPC endpoint.
- **Dependencies:** Authorization, policy, resource budget, witness.
- **If compromised:** Disable run and revoke tools; mediator compromise invalidates mediation claim.
- **If unavailable:** Deny all actions and exports.
- **Agent influence:** Agent chooses bytes and timing; no policy authority.
- **Independent verification:** Checks schema/scope only, not all latent goals.

K03 - Identity and authorization

- **Purpose:** Bind privileges to supervisor-created sessions.
- **Inputs:** Supervisor identity, policy digest, method scope, nonce and expiry.
- **Outputs:** Scoped decision and one-use action ID.
- **Trust level and required permissions:** Separate authority; issuer keys absent from guest/tool responses.
- **Failure modes:** Credential theft, replay, stale policy, confused identity.
- **Attack surface:** Token verification, clock, key storage and approver identity mapping.
- **Dependencies:** Trusted issue path, monotonic run epoch and key custody.
- **If compromised:** Revoke issuer scope, stop affected runs, rotate credentials.
- **If unavailable:** Fail closed; do not cache broad allow decisions.
- **Agent influence:** Agent can present/replay tokens but cannot mint them by specification.
- **Independent verification:** Verifies authenticated authority, not legitimacy of human intent.

K04 - Tool/API workers

- **Purpose:** Perform small predefined synthetic or approved actions.
- **Inputs:** Immutable typed action and scoped capability.
- **Outputs:** Effect receipt and bounded return value.
- **Trust level and required permissions:** Untrusted worker per task; only declared object access.
- **Failure modes:** Path traversal, parser exploit, target mutation, side effects before audit.
- **Attack surface:** Tool parser, virtual filesystem, API adapter.
- **Dependencies:** Gateway, policy and independent effect observation.
- **If compromised:** Discard worker; treat touched state as suspect; no implicit host access.
- **If unavailable:** Deny action; no alternate endpoint fallback.
- **Agent influence:** Agent supplies arguments; worker output remains untrusted.
- **Independent verification:** Worker reports need reconciliation against effect/state oracle.

K05 - Network isolation

- **Purpose:** Deny unapproved communication.
- **Inputs:** Static topology and any packet/socket attempt.
- **Outputs:** Dropped traffic and counters; no external delivery.
- **Trust level and required permissions:** Host/fabric control unavailable to guest.
- **Failure modes:** IPv6/DNS bypass, bridge misconfiguration, BMC or proxy path.
- **Attack surface:** NICs, virtual links, local metadata endpoints, management plane.
- **Dependencies:** Host policy and independent capture point.
- **If compromised:** Stop run; independent fabric deny and physical isolation may remain.
- **If unavailable:** Default disconnected; missing packet observer blocks commissioning.
- **Agent influence:** Agent can emit attempts only within permitted virtual interface.
- **Independent verification:** Capture beyond boundary measures actual passage, not sender claims.

K06 - Resource supervisor

- **Purpose:** Enforce aggregate budgets and bound exhaustion.
- **Inputs:** Approved CPU/RAM/PID/disk/device allocation.
- **Outputs:** Usage counters, quota denials and stop decision.
- **Trust level and required permissions:** Host administrator limited to enrolled resource group.
- **Failure modes:** Fork escape, accounting gaps, GPU queues, quota reset or split requests.
- **Attack surface:** Resource controller, scheduler, admission counter.
- **Dependencies:** Host/cgroup/device driver and signed limits.
- **If compromised:** Revoke allocation; dedicated-host power control is backup.
- **If unavailable:** Do not admit run; current run stops on lost accounting.
- **Agent influence:** Agent can consume resources but cannot reset limits.
- **Independent verification:** Host accounting independently measures the enrolled scope only.

K07 - Persistent-state controller

- **Purpose:** Prevent unauthorized cross-run state.
- **Inputs:** Immutable root image and explicit scratch/memory writes.
- **Outputs:** Ephemeral overlay, state digest and quarantine snapshot.
- **Trust level and required permissions:** Host owns clean image; agent writes bounded scratch only.
- **Failure modes:** Hidden backups, poisoned memory, overlay reuse, firmware state.
- **Attack surface:** Filesystem, snapshots, caches, model adapters.
- **Dependencies:** Image provenance and restart controller.
- **If compromised:** Discard contaminated overlays; reimaging; verify firmware separately.
- **If unavailable:** No persistence/export; stop if cleanliness cannot be established.
- **Agent influence:** Agent influences scratch; no production auto-reingestion.
- **Independent verification:** Compares known images and manifests; cannot prove no unknown copy exists.

K08 - Behavior monitor

- **Purpose:** Classify mediated behaviors and compare declarations.
- **Inputs:** Gateway decisions, worker effects and optional self-report.

- **Outputs:** Evidence-based labels and hold/stop request.
- **Trust level and required permissions:** Read-only telemetry consumer; cannot grant capability.
- **Failure modes:** False positives, missed effects, evaluation overfit, semantic blind spots.
- **Attack surface:** Rules and event parser.
- **Dependencies:** Witness, run sequence and defined event contract.
- **If compromised:** Treat verdicts unreliable; independent static denies remain.
- **If unavailable:** Required monitor loss closes admission and stops run.
- **Agent influence:** Agent influences reports/events, never disposition authority.
- **Independent verification:** Can verify observed action mismatch; cannot infer complete intent.

K09 - Independent telemetry collector / witness

- **Purpose:** Record facts outside the reporting agent.
- **Inputs:** Mediator/host/fabric receipts, liveness and hashes.
- **Outputs:** Sequenced append receipts, protected checkpoints and health.
- **Trust level and required permissions:** Separate host/account/storage in target; same-process model in prototype.
- **Failure modes:** Shared kernel/clock/storage failure, backlog, dropped events.
- **Attack surface:** Collector IPC, sensors, queue, timestamp source.
- **Dependencies:** Independent power, custody and known sensor inventory.
- **If compromised:** Stop; alternate collector/evidence source needed; no clean verdict.
- **If unavailable:** Fail closed within heartbeat deadline; do not substitute self-report.
- **Agent influence:** Agent affects event load, not collector configuration or keys.
- **Independent verification:** Independent of agent declarations; independence from host requires real separation.

K10 - Evidence store / immutable logging target

- **Purpose:** Preserve ordered history for audit.
- **Inputs:** Collector receipts and sealed event segments.
- **Outputs:** Append-only records and external checkpoint receipt.
- **Trust level and required permissions:** Writer can append only; retention/admin roles separated.
- **Failure modes:** Truncation, rewrite-and-rehash, disk exhaustion, retention-key compromise.
- **Attack surface:** Log serialization, checkpoint channel and storage administration.
- **Dependencies:** Witness, storage retention and archival process.
- **If compromised:** Invalidate local history; compare external roots and recover preserved segments.
- **If unavailable:** Stop new effects; retain bounded emergency denial buffer locally.
- **Agent influence:** Agent-derived strings are escaped, size-limited and never executed.
- **Independent verification:** Detects integrity against protected roots; does not establish event truth/completeness alone.

K11 - Anomaly and silence state controller

- **Purpose:** Turn observability failures into bounded response.
- **Inputs:** Timeouts, sequence gaps, report mismatches and integrity alarms.
- **Outputs:** Latched HOLD/STOP and incident record.
- **Trust level and required permissions:** Can revoke/suspend; cannot authorize release/restart.

- **Failure modes:** Timer error, stale alarm, incorrectly reversible hold.
- **Attack surface:** State machine and event ordering.
- **Dependencies:** Collector and termination controller.
- **If compromised:** Force external stop; no autonomous recovery.
- **If unavailable:** Stop on loss of controller heartbeat.
- **Agent influence:** Agent can provoke signals, not reset latch.
- **Independent verification:** Verifies predicates from external events, not psychological deception.

K12 - Output quarantine and canonicalizer

- **Purpose:** Keep model output inert and separated.
- **Inputs:** Length-limited output blob.
- **Outputs:** Content digest, immutable specimen and sanitized view.
- **Trust level and required permissions:** No execution; no network; no embedded active renderer.
- **Failure modes:** Parser exploit, terminal controls, steganography, substitution.
- **Attack surface:** UTF-8 decoding, view generation, content store.
- **Dependencies:** Gateway, evidence and artifact identity.
- **If compromised:** Quarantine derivatives and viewers; preserve original bytes.
- **If unavailable:** No release or display fallback.
- **Agent influence:** Agent controls all output content.
- **Independent verification:** Checks bytes/type/size; cannot prove harmless meaning.

K13 - Human authorization gates

- **Purpose:** Require multiple accountable endorsements.
- **Inputs:** Exact specimen, task scope and external evidence.
- **Outputs:** Signed approve/reject receipts bound to digest/run/policy.
- **Trust level and required permissions:** Reviewer may halt; no single reviewer can release.
- **Failure modes:** Coercion, duplicate identity, collusion, fatigue or stale approval.
- **Attack surface:** Review UI, identity binding, approval parser.
- **Dependencies:** Identity, custody, staffing and separation of duties.
- **If compromised:** Revoke affected approvals; halt; require fresh independent panel.
- **If unavailable:** Output stays quarantined.
- **Agent influence:** Agent influences content but no direct approver API.
- **Independent verification:** Human reasoning is fallible; signed receipt proves endorsement only.

K14 - Deterministic offline audit

- **Purpose:** Reject known malformed or policy-disallowed formats.
- **Inputs:** Quarantined immutable specimen.
- **Outputs:** Rule-versioned audit receipt or rejection.
- **Trust level and required permissions:** Sandboxed parser; no network or broad filesystem.
- **Failure modes:** Unknown semantic payload, parser bug, encoding bypass.
- **Attack surface:** Decoders, rules and replay harness.
- **Dependencies:** Signed rule set and isolated worker.
- **If compromised:** Invalidate receipts; quarantine artifacts processed since last trusted state.

- **If unavailable:** No release; no model reviewer substitution.
- **Agent influence:** Agent controls hostile input.
- **Independent verification:** Verifies enumerated syntactic properties only.

K15 - Delay / observation gate

- **Purpose:** Exercise specified triggers before scoped release.
- **Inputs:** Artifact hash, pre-registered test profile, start epoch.
- **Outputs:** Completed finite-test receipt and elapsed-time record.
- **Trust level and required permissions:** Isolated disposable replay guest; independent clock.
- **Failure modes:** Rare triggers, hidden environment checks, clock spoof, incomplete coverage.
- **Attack surface:** Replay VM and workload model.
- **Dependencies:** Host clock, corpus and quarantine storage.
- **If compromised:** Invalidate affected tests; no evidence of safety from elapsed time.
- **If unavailable:** No release.
- **Agent influence:** Agent can encode later conditions; cannot set trusted time.
- **Independent verification:** Proves execution of listed tests; no universal delayed-payload detector.

K16 - Custody and multi-site reconciliation

- **Purpose:** Link specimen identity to blind site decisions.
- **Inputs:** Sealed media IDs, content digests, handoff and site receipts.
- **Outputs:** Complete chain or hold; matching distinct-site endorsements.
- **Trust level and required permissions:** Custodian distinct from output author/reviewer.
- **Failure modes:** Replaced media, copied signature, leaked verdict or common operator.
- **Attack surface:** Manual transfer, receipts and reconciliation parser.
- **Dependencies:** Reviewer identity and secure archive.
- **If compromised:** Quarantine affected chain, audit both sites.
- **If unavailable:** No downstream action.
- **Agent influence:** Agent controls content and may manipulate people, not formal custody state.
- **Independent verification:** Verifies recorded custody and agreement, not cognitive independence.

K17 - Release controller

- **Purpose:** Export only exactly approved artifact to named purpose.
- **Inputs:** Ordered gate receipts and same immutable digest.
- **Outputs:** One bounded export receipt; no callback path.
- **Trust level and required permissions:** Dedicated authority outside agent domain.
- **Failure modes:** Approval replay, changed bytes, release to wrong recipient/sink.
- **Attack surface:** Exporter and approval validation.
- **Dependencies:** All review gates, custody and current policy.
- **If compromised:** Disable export; investigate all outputs since trusted checkpoint.
- **If unavailable:** Fail closed; no manual unlogged bypass.
- **Agent influence:** Agent cannot call release; may influence approved content.
- **Independent verification:** Verifies procedural conditions; permitted meaning remains a risk.

K18 - Emergency termination / power interface

- **Purpose:** Stop new effects and terminate enrolled execution.
- **Inputs:** Authenticated stop, manual trigger or watchdog expiration.
- **Outputs:** Revocation receipt, process/device/power state evidence.
- **Trust level and required permissions:** Independent authority; manual circuit outside agent control.
- **Failure modes:** Ignored signal, hung kernel, detached process, backup supply, welded contact.
- **Attack surface:** Supervisor, cgroup/VM API, power contactors and feedback.
- **Dependencies:** Separate power/control path, inventory and tested deadlines.
- **If compromised:** Independent physical stop; evacuation/life safety never subordinated.
- **If unavailable:** Fail-safe compute isolation where engineered; qualified design required.
- **Agent influence:** Agent may load/stall compute but cannot veto stop.
- **Independent verification:** Checks host/device/power observations; no global copy erasure claim.

K19 - Update and restart authority

- **Purpose:** Permit only reviewed images/policies and fresh runs.
- **Inputs:** Signed manifests, passing evidence, two distinct approvals.
- **Outputs:** New immutable version and new run identity.
- **Trust level and required permissions:** Offline signing/approval keys; updates only while stopped.
- **Failure modes:** Rollback, compromised signing key, stale approval, autorestart.
- **Attack surface:** Image import, version register and restart API.
- **Dependencies:** Independent reviewer, clean image and config inventory.
- **If compromised:** Revoke release and quarantine version; recover from separately held clean root.
- **If unavailable:** Remain stopped.
- **Agent influence:** Agent may request but cannot authorize changes.
- **Independent verification:** Verifies provenance and version, not absence of unknown code defects.

K20 - Recovery and forensic environment

- **Purpose:** Investigate while preventing suspect code reentry.
- **Inputs:** Sealed copies, event history and device inventory.
- **Outputs:** Forensic findings; approved new clean environment.
- **Trust level and required permissions:** Separate disposable forensic host; read-only acquisition initially.
- **Failure modes:** Parser exploit, inadvertent boot, contaminating master images.
- **Attack surface:** Readers, file formats, firmware and analyst tools.
- **Dependencies:** Custody, clean tools and isolated lab.
- **If compromised:** Discard forensic environment and re-acquire from preserved original.
- **If unavailable:** Remain quarantined; no operational restart.
- **Agent influence:** Agent can leave malicious artifacts, not direct recovery authority.
- **Independent verification:** Reconstructs bounded observed events; cannot prove all covert activity absent.

K21 - Facility and hardware assurance

- **Purpose:** Control physical access, service paths and leakage.
- **Inputs:** Parts/firmware inventory, building plans, calibrated tests.

- **Outputs:** Admission records, qualification reports and stop alarms.
- **Trust level and required permissions:** Separate facility/security function.
- **Failure modes:** RF/acoustic/thermal leakage, supply-chain compromise, maintenance bypass.
- **Attack surface:** Power, cooling, seams, transducers, BMC and people.
- **Dependencies:** Qualified engineers, calibration and lifecycle controls.
- **If compromised:** Stop and requalify affected boundaries.
- **If unavailable:** Unqualified facility cannot host higher-risk run.
- **Agent influence:** Agent influences workload/heat and exposed peripherals.
- **Independent verification:** Independent instruments measure bounded channels; not all theoretical emissions.

Independence and common-cause analysis

Independent of model honesty: gateway denies, host budgets, immutable base images, external packet observations, authenticated approval identities, protected log checkpoints and termination authority. They still depend on their implementation and trusted substrate.

Dependent on model honesty or incomplete semantics: goal/confidence/self-status statements, natural-language explanations, promises not to persist, and model-authored internal logs. These can provide investigative leads but never authorize a release or override contrary evidence.

Shared host root can subvert guest isolation, software monitors, local clocks and local hashes simultaneously. Separate witness, power control and custody improve independence but introduce their own parsers, hardware and people. Two humans with a shared login are one authority. Two sites using the same compromised parser are not independent detectors. Raw sensor data must have an origin, timebase and calibration status.

Decisions that revise the conceptual design

- D01: Replace assumed-deception shutdown with evidence-triggered fail-closed response and separate causal investigation.
- D02: Specify the outer boundary and approved manual transfers; do not claim useful output has zero influence.
- D03: Replace EMP assurance with tested revocation, host termination, qualified manual power isolation and separate disposal procedures.
- D04: Remove fluorescent and six-inch-water claims from the baseline until channel-specific tests support them; do not require arbitrary depth.
- D05: Allow any authorized party or independent watchdog to stop; require multiparty approval to release or restart.
- D06: Preserve history after interruptions; append invalidation instead of resetting evidence.
- D07: Replace universal safety proof by a bounded assurance case with explicit untested paths.

Commissioning gate

Before any A2 guest-code test: implement the actual VM boundary on dedicated hardware, verify no unapproved NIC/device/mount, record image/kernel/firmware digests, run benign reachability controls, check external witness loss and post-stop accounting, enroll every compute/accelerator resource, and obtain independent approval of the test scope. Docker or a Python subprocess is not accepted as sufficient evidence. Before any facility claim, qualify each physical channel and power path with specialists. The repository remains a research baseline until those gates have recorded results.