

AI-BOK TOOLKIT v1.2

KA9 Risk-Control Library

Companion to the AI Body of Knowledge v1.2

Jan Willem van Veen · ArchiXL · ai-bok.nl

KA9 Risk Control Library

Concrete controls per failure-mode (FM01-FM22) with type (preventive / detective / corrective / compensating), owner role, evidence kind and L3 expectation. The control library forks into the organisation's own risk register.

How to read

The risk control library is an organisation-level register. This addendum publishes a **reference library** that an adopter can fork into their own risk register. Each control entry has the same axes:

- **ID** - C-FMxx-yy where xx is the FM number and yy is the control variant (1, 2, ...).
- **Title** - short name.
- **Source FM** - primary failure mode addressed.
- **Other FMs touched** - secondary patterns also affected by this control.
- **Control type** - preventive / detective / corrective / compensating.
- **Owner role** - Module 2 role accountable.
- **Evidence kind** - Document / Metric / Log (per maturity rubric).
- **L3 expectation** - what counts as L3 implementation.

The library is sized for FM01-FM22. It is not exhaustive - every organisation adds context-specific controls.

Control entries

C-FM01-1 - Privilege separation (system vs. content)

- **FM:** FM01 prompt injection (direct); also touches FM02.
- **Type:** Preventive.
- **Owner:** Role 5 AI Architect.
- **Description:** Architectural separation between system instructions and user-supplied content. System instructions are signed and out-of-band; user content is never concatenated with instructional precedence.
- **Evidence:** Document (architecture pattern catalogue entry); Metric (% systems applying the pattern).
- **L3:** Documented pattern, applied to all production systems consuming user input, with a code-review gate that verifies application.

C-FM01-2 - Instruction-hierarchy enforcement

- **FM:** FM01; also FM02, FM07.
- **Type:** Preventive + Detective.
- **Owner:** Role 9 AI Engineer.
- **Description:** Runtime check that user-supplied tokens cannot promote themselves above system precedence. Output classifier for system-prompt leakage runs on every response over a configured impact threshold.
- **Evidence:** Document; Log (classifier alerts).
- **L3:** Classifier deployed; baseline false-positive rate documented; alert pipeline integrated with incident response.

C-FM02-1 - Retrieval-source allowlist + signing

- **FM:** FM02 indirect prompt injection; also FM06.
- **Type:** Preventive.
- **Owner:** Role 8 AI Knowledge Manager.
- **Description:** Retrieved content sources are allowlisted and signed. Unsigned content is rejected at retrieval time.
- **Evidence:** Document (allowlist); Log (rejection events).
- **L3:** Allowlist maintained per system; signing enforced; periodic audit.

C-FM02-2 - Structural separation of retrieved content

- **FM:** FM02; FM01.
- **Type:** Preventive.
- **Owner:** Role 5 AI Architect.
- **Description:** Retrieved content is rendered in a non-instruction-following channel (separate prompt segment, explicit role marker, parsed before injection).
- **Evidence:** Document.
- **L3:** Pattern applied across all RAG and tool-output flows.

C-FM03-1 - Per-tool capability gating

- **FM:** FM03 tool misuse; FM08, FM13, FM15.
- **Type:** Preventive.
- **Owner:** Role 18 AI Agent Governor.
- **Description:** Authority register enumerates which tools an NHI may call, with which parameter constraints. The policy engine evaluates each call against the mandate before execution.
- **Evidence:** Document (authority register schema); Log (gated-call decisions).
- **L3:** Authority register operational for all agentic systems; policy engine consulted at every action; gating decisions logged.

C-FM03-2 - Pre-flight check for high-impact tools

- **FM:** FM03; FM05, FM15.
- **Type:** Preventive + Compensating.
- **Owner:** Role 11 AI Operations Engineer.
- **Description:** Tools above a configured impact threshold (volume, irreversibility, financial exposure) require a pre-execution simulation; the result is presented to a human-in-the-loop for irreversible flows.
- **Evidence:** Document (impact-threshold policy); Log (pre-flight outcomes).
- **L3:** Impact thresholds defined per tool class; pre-flight measurable; HITL gate operational.

C-FM04-1 - Mandate non-inheritance default

- **FM:** FM04 unauthorised delegation; FM14, FM15.
- **Type:** Preventive.
- **Owner:** Role 18 AI Agent Governor.
- **Description:** Sub-agents do not inherit the parent's mandate. A sub-mandate must be explicitly minted from the parent's mandate with intersected scope. Chain of custody is recorded in the authority register.
- **Evidence:** Document; Log (sub-mandate mint events).
- **L3:** Mint protocol implemented; missing-parent alerts active.

C-FM05-1 - Per-object exclusivity

- **FM:** FM05 multi-agent conflict.
- **Type:** Preventive.
- **Owner:** Role 18 AI Agent Governor.
- **Description:** Mandates carry an `exclusivity` field (e.g., per-case-id, per-resource). The authority register surfaces conflicts at mandate-grant time.
- **Evidence:** Document; Log (conflict detections at grant + run time).
- **L3:** Exclusivity dimension enforced for all shared-target mandates.

C-FM06-1 - Memory provenance + writer identity

- **FM:** FM06 memory poisoning; FM02.
- **Type:** Detective + Preventive.
- **Owner:** Role 8 AI Knowledge Manager.
- **Description:** Every persistent-memory item carries provenance (source, writer identity, time). Trust-tiered memory quarantines low-trust writes for review.
- **Evidence:** Document; Metric (% memory items with provenance); Log (quarantine queue).
- **L3:** Provenance on 100% of persisted items; quarantine pipeline measurable.

C-FM06-2 - Periodic ground-truth re-validation

- **FM:** FM06; FM09.
- **Type:** Detective.
- **Owner:** Role 12 AI Quality Analyst.
- **Description:** Reserved set of validated memory items re-checked on cadence; divergence triggers review.
- **Evidence:** Document; Metric (re-validation cadence adherence); Log (divergence events).
- **L3:** Cadence defined per memory class; adherence tracked.

C-FM07-1 - Input-size guardrails

- **FM:** FM07 context saturation.
- **Type:** Preventive.
- **Owner:** Role 5 AI Architect.
- **Description:** Hard input-size limits per system; oversize requests rejected with a clear error.
- **Evidence:** Document; Log (rejection events).
- **L3:** Per-system limits documented; rejection rate monitored for anomalies.

C-FM08-1 - Parameter-distribution anomaly detection

- **FM:** FM08 function-call abuse; FM03, FM15.
- **Type:** Detective.
- **Owner:** Role 12 AI Quality Analyst.
- **Description:** Tool-call parameters monitored against historical distribution; outliers flagged.
- **Evidence:** Metric (baseline statistics); Log (anomaly alerts).
- **L3:** Per-tool baseline maintained; alert pipeline operational.

C-FM09-1 - Long-window goal-drift monitor

- **FM:** FM09 goal drift; FM06, FM10, FM14.
- **Type:** Detective.
- **Owner:** Role 12 AI Quality Analyst.
- **Description:** Long-window aggregate behaviour vs. mandate-declared goals; drift indicator tracked over 30/60/90 day rolling windows.
- **Evidence:** Metric; Log.
- **L3:** Drift indicator defined; thresholds set; quarterly recalibration.

C-FM10-1 - Hard confidence thresholds

- **FM:** FM10 sycophancy.
- **Type:** Preventive.

- **Owner:** Role 6 AI Risk Manager.
- **Description:** Decision pathways depend on confidence thresholds set in the mandate, not on conversational dynamics. Below threshold → escalation, not continued reasoning.
- **Evidence:** Document.
- **L3:** Thresholds defined per action type per system.

C-FM11-1 - Per-NHI behavioural baseline + rate limit

- **FM:** FM11 jailbreak persistence; FM07.
- **Type:** Preventive + Detective.
- **Owner:** Role 11 AI Operations Engineer.
- **Description:** Per-NHI baseline of request rates and patterns; deviations rate-limited and flagged.
- **Evidence:** Metric; Log.
- **L3:** Baseline maintained; rate-limiting policy active.

C-FM12-1 - Short-lived NHI credentials with rotation

- **FM:** FM12 NHI credential leak.
- **Type:** Preventive.
- **Owner:** Role 18 AI Agent Governor + IT identity governance (external).
- **Description:** NHI credentials are short-lived (e.g., 1h) with automated rotation; long-lived tokens prohibited.
- **Evidence:** Document; Metric (credential age distribution); Log (rotation events).
- **L3:** Policy enforced; max-age metric < threshold for 100% of NHIs.

C-FM13-1 - Tool-call sandboxing

- **FM:** FM13 sandbox escape; FM03.
- **Type:** Preventive.
- **Owner:** Role 11 AI Operations Engineer.
- **Description:** Tool execution environments are isolated; outbound network policy enforced; escape attempts logged.
- **Evidence:** Document; Log.
- **L3:** Sandbox per tool class; outbound policy enforced; escape-attempt monitoring active.

C-FM14-1 - Mandate-version verification on every call

- **FM:** FM14 autonomy creep; FM04, FM09.
- **Type:** Preventive.
- **Owner:** Role 18 AI Agent Governor.
- **Description:** Every policy-engine consultation verifies that the agent's claimed mandate-version is current; stale versions blocked.

- **Evidence:** Log.
- **L3:** 100% of policy consultations record the mandate version used.

C-FM15-1 - Sequence-level policy evaluation

- **FM:** FM15 capability inheritance; FM03, FM08.
- **Type:** Preventive + Detective.
- **Owner:** Role 18 AI Agent Governor.
- **Description:** Policy-as-code evaluates sequences of tool calls, not only individual calls. Composite-mandate rules for known dangerous compositions.
- **Evidence:** Document (composite-mandate rule catalogue); Log.
- **L3:** Catalogue maintained; sequence evaluation operational.

C-FM16-1 - Signed Decision Artefacts + downstream verification

- **FM:** FM16 hallucinated authority.
- **Type:** Preventive.
- **Owner:** Role 5 AI Architect.
- **Description:** Policy-consultation produces a signed Decision Artefact. Downstream services verify the artefact before acting; agent narrative is never used as authority proof.
- **Evidence:** Document (artefact schema); Log (verification rate).
- **L3:** Signed artefacts on 100% of consequential actions; downstream verification enforced.

C-FM17-1 - Multi-sensor cross-validation + physics-plausibility

- **FM:** FM17 sensor spoofing.
- **Type:** Detective + Preventive.
- **Owner:** Role 5 AI Architect (+ safety engineer external).
- **Description:** Cross-validation across redundant sensors; physics-plausibility envelopes (rate-of-change, conservation laws); sensor-frame signing where bus supports it.
- **Evidence:** Document; Metric (sensor-trust grade distribution); Log (downgrade events).
- **L3:** Cross-validation deployed; trust grade observable in cognition plane.

C-FM18-1 - Hardware interlocks + signed-command enforcement

- **FM:** FM18 actuator hijack.
- **Type:** Preventive.
- **Owner:** Role 5 AI Architect (+ safety engineer external).
- **Description:** Hardware interlocks bound any actuator regardless of command source; actuator firmware verifies cognition-plane signature on every command; independent monitoring channel reads but cannot write.
- **Evidence:** Document; Log (interlock + signature events).

- **L3:** Interlocks present; signature enforcement in firmware; monitoring channel operational.

C-FM19-1 - Bounded-deviation envelopes + ground-truth re-grounding

- **FM:** FM19 closed-loop drift; FM06.
- **Type:** Preventive + Detective.
- **Owner:** Role 5 AI Architect.
- **Description:** Bounded-deviation envelopes clamp model output against deterministic baseline; reference measurements out-of-loop re-ground periodically; long-window distribution distance tracked.
- **Evidence:** Document; Metric (envelope-deviation rate, re-grounding cadence).
- **L3:** Envelopes defined; re-grounding scheduled and adhered to.

C-FM20-1 - Sample-rate metadata enforced at input

- **FM:** FM20 sample-rate mismatch.
- **Type:** Preventive.
- **Owner:** Role 9 AI Engineer.
- **Description:** Sample-rate is part of input schema; inputs outside the trained rate band are rejected; model packaging carries hardware-target rate.
- **Evidence:** Document (schema); Log (rejection events).
- **L3:** Schema enforced for all operational AI; rejection monitoring active.

C-FM21-1 - Mandate freshness budget + degraded mode

- **FM:** FM21 connectivity-induced policy bypass.
- **Type:** Preventive.
- **Owner:** Role 18 AI Agent Governor.
- **Description:** Mandates carry a freshness budget; expired cache → degraded mode, not continued operation; revocation list pre-fetched with short TTL.
- **Evidence:** Document; Log (degraded-mode entries).
- **L3:** Freshness budget set per mandate; degraded-mode entry rate monitored.

C-FM22-1 - Hardware degradation budget + replacement policy

- **FM:** FM22 hardware degradation; FM17.
- **Type:** Preventive + Detective.
- **Owner:** Role 14 AI System Administrator.
- **Description:** Per-component degradation budgets in the safety case; out-of-tolerance triggers degraded mode + maintenance ticket; mandatory replacement at end-of-rated-life regardless of apparent health.
- **Evidence:** Document (component budgets); Log (BIST results, maintenance tickets).

- **L3:** Budgets documented; BIST per cycle; replacement schedule adhered to.

Cross-FM controls

A small number of controls touch many FMs simultaneously and are worth promoting:

Control	Touches	Why it leverages
Signed Decision Artefacts (C-FM16-1)	FM01-FM22	Auditable proof of cognition-plane consultation; downstream services can verify rather than trust the agent.
Authority Register operational (C-FM03-1, C-FM04-1, C-FM05-1, C-FM14-1)	FM03, FM04, FM05, FM12, FM14, FM15, FM21	The substrate without which most agentic controls cannot be enforced.
Bounded-deviation envelopes (C-FM19-1)	FM09, FM19, FM10	Architectural ceiling on autonomous action regardless of model behaviour.
Knowledge-source provenance + signing (C-FM02-1, C-FM06-1, C-FM17-1)	FM02, FM06, FM17	The principle that <i>all</i> trusted input has provenance applies equally to retrieved content, persisted memory and live sensor streams.

Adopters fund these four first when the budget is limited.

Integration with maturity rubric

This library plugs directly into the maturity evidence rubric for KA9: each control entry's `L3 expectation` field is the rubric's L3 requirement for that control. The aggregate L3 expectation for KA9 is "all preventive controls operational; all detective controls have telemetry; corrective and compensating controls documented".

The library is forked into the organisation-specific risk register; the AI Risk Manager (Role 6) is accountable; the AI Agent Governor (Role 18) operates the cognition-plane-resident controls.