

AI-BOK TOOLKIT v1.2 · NL

KG9 Risicobeheersingsbibliotheek

Aanvulling op de AI Body of Knowledge v1.2

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

KG9 Risico-controle-bibliotheek

Concrete controles per faalmodus (FM01-FM22) met type (preventief / detectief / correctief / compenserend), eigenaarsrol, bewijssoort en L3-verwachting. De controle-bibliotheek wordt geforkt naar het eigen risicoregister van de organisatie.

Hoe te lezen

De risico-controle-bibliotheek is een register op organisatieniveau. Dit addendum publiceert een **referentiebibliotheek** die een adopterende organisatie kan forken naar haar eigen risicoregister. Elke controle-entry heeft dezelfde assen:

- **ID** - C-FMxx-yy waarbij xx het FM-nummer is en yy de controlevariant (1, 2, ...).
- **Titel** - korte naam.
- **Bron-FM** - primaire faalmodus die wordt geadresseerd.
- **Overige geraakte FM's** - secundaire patronen die eveneens door deze controle worden beïnvloed.
- **Controltype** - preventief / detectief / correctief / compenserend.
- **Eigenaarsrol** - verantwoordelijke rol uit Module 2.
- **Bewijssoort** - Document / Metriek / Log (conform volwassenheidsrubric).
- **L3-verwachting** - wat telt als L3-implementatie.

De bibliotheek is gedimensioneerd op FM01-FM22. Ze is niet uitputtend - elke organisatie voegt contextspecifieke controles toe.

Controle-entries

C-FM01-1 - Privilegescheiding (systeem vs. content)

- **FM:** FM01 prompt injection (direct); raakt ook FM02.
- **Type:** Preventief.
- **Eigenaar:** Rol 5 AI Architect.
- **Omschrijving:** Architecturale scheiding tussen systeeminstructies en door de gebruiker aangeleverde content. Systeeminstructies zijn ondertekend en out-of-band; gebruikerscontent wordt nooit geconcateneerd met instructieve precedentie.
- **Bewijs:** Document (entry in het architectuurpatronencatalogus); Metriek (% systemen dat het patroon toepast).

- **L3:** Gedocumenteerd patroon, toegepast op alle productiesystemen die gebruikersinput consumeren, met een code-review-gate die de toepassing verifieert.

C-FM01-2 - Handhaving van instructiehiërarchie

- **FM:** FM01; ook FM02, FM07.
- **Type:** Preventief + Detectief.
- **Eigenaar:** Rol 9 AI Engineer.
- **Omschrijving:** Runtime-check dat door de gebruiker aangeleverde tokens zichzelf niet boven systeemprecedentie kunnen promoveren. Een output-classifier voor system-prompt-lekkage draait op elke respons boven een geconfigureerde impactdrempel.
- **Bewijs:** Document; Log (classifier-alerts).
- **L3:** Classifier uitgerold; baseline false-positive-rate gedocumenteerd; alert-pijplijn geïntegreerd met incident response.

C-FM02-1 - Allowlist + ondertekening van retrievalbronnen

- **FM:** FM02 prompt injection (indirect); ook FM06.
- **Type:** Preventief.
- **Eigenaar:** Rol 8 AI Kennismanager.
- **Omschrijving:** Bronnen van opgehaalde content staan op een allowlist en zijn ondertekend. Niet-ondertekende content wordt op het moment van retrieval geweigerd.
- **Bewijs:** Document (allowlist); Log (weigeringsgebeurtenissen).
- **L3:** Allowlist per systeem onderhouden; ondertekening afgedwongen; periodieke audit.

C-FM02-2 - Structurele scheiding van opgehaalde content

- **FM:** FM02; FM01.
- **Type:** Preventief.
- **Eigenaar:** Rol 5 AI Architect.
- **Omschrijving:** Opgehaalde content wordt weergegeven in een kanaal dat geen instructies volgt (apart promptsegment, expliciete rolmarkering, geparseerd vóór injectie).
- **Bewijs:** Document.
- **L3:** Patroon toegepast op alle RAG- en tool-output-flows.

C-FM03-1 - Capability-gating per tool

- **FM:** FM03 tool misuse; FM08, FM13, FM15.
- **Type:** Preventief.
- **Eigenaar:** Rol 18 AI Agent-gouverneur.
- **Omschrijving:** Het autoriteitenregister somt op welke tools een NHI mag aanroepen, met welke parameterrestricties. De policy engine toetst elke aanroep aan het mandaat vóór uitvoering.

- **Bewijs:** Document (schema van het autoriteitenregister); Log (gating-beslissingen per aanroep).
- **L3:** Autoriteitenregister operationeel voor alle agentic systemen; policy engine geraadpleegd bij elke actie; gating-beslissingen gelogd.

C-FM03-2 - Pre-flight-check voor tools met hoge impact

- **FM:** FM03; FM05, FM15.
- **Type:** Preventief + Compenserend.
- **Eigenaar:** Rol 11 AI Operations Engineer.
- **Omschrijving:** Tools boven een geconfigureerde impactdrempel (volume, onomkeerbaarheid, financiële blootstelling) vereisen een simulatie vóór uitvoering; het resultaat wordt bij onomkeerbare flows voorgelegd aan een human-in-the-loop.
- **Bewijs:** Document (impactdrempelbeleid); Log (pre-flight-uitkomsten).
- **L3:** Impactdrempels gedefinieerd per toolklasse; pre-flight meetbaar; HITL-gate operationeel.

C-FM04-1 - Standaard geen mandaat-overerving

- **FM:** FM04 unauthorised delegation; FM14, FM15.
- **Type:** Preventief.
- **Eigenaar:** Rol 18 AI Agent-gouverneur.
- **Omschrijving:** Sub-agents erven het mandaat van de ouder niet. Een sub-mandaat moet expliciet worden uitgegeven vanuit het oudermandaat met geïntersecteerde scope. De chain of custody wordt vastgelegd in het autoriteitenregister.
- **Bewijs:** Document; Log (uitgifte-events van sub-mandaten).
- **L3:** Uitgifteprotocol geïmplementeerd; alerts voor ontbrekende ouder actief.

C-FM05-1 - Exclusiviteit per object

- **FM:** FM05 multi-agent conflict.
- **Type:** Preventief.
- **Eigenaar:** Rol 18 AI Agent-gouverneur.
- **Omschrijving:** Mandaten dragen een `exclusivity` -veld (bijv. per-case-id, per-resource). Het autoriteitenregister maakt conflicten zichtbaar op het moment van mandaatverlening.
- **Bewijs:** Document; Log (conflictdetecties bij verlening + runtime).
- **L3:** Exclusiviteitsdimensie afgedwongen voor alle mandaten met gedeeld doelobject.

C-FM06-1 - Geheugenprovenance + schrijveridentiteit

- **FM:** FM06 memory poisoning; FM02.
- **Type:** Detectief + Preventief.
- **Eigenaar:** Rol 8 AI Kennismanager.

- **Omschrijving:** Elk persistent geheugenitem draagt provenance (bron, schrijveridentiteit, tijd). Trust-tiered geheugen plaatst schrijfacties met laag vertrouwen in quarantaine voor review.
- **Bewijs:** Document; Metriek (% geheugenitems met provenance); Log (quarantainewachtrij).
- **L3:** Provenance op 100% van de gepersisterde items; quarantainepijplijn meetbaar.

C-FM06-2 - Periodieke hervalidatie tegen grondwaarheid

- **FM:** FM06; FM09.
- **Type:** Detectief.
- **Eigenaar:** Rol 12 AI Kwaliteitsanalist.
- **Omschrijving:** Gereserveerde set gevalideerde geheugenitems wordt volgens cadans opnieuw gecontroleerd; divergentie triggert review.
- **Bewijs:** Document; Metriek (naleving van hervalidatiecadans); Log (divergentiegebeurtenissen).
- **L3:** Cadans gedefinieerd per geheugenklasse; naleving gevolgd.

C-FM07-1 - Guardrails op inputomvang

- **FM:** FM07 context saturation.
- **Type:** Preventief.
- **Eigenaar:** Rol 5 AI Architect.
- **Omschrijving:** Harde limieten op inputomvang per systeem; te grote verzoeken worden geweigerd met een duidelijke foutmelding.
- **Bewijs:** Document; Log (weigeringsgebeurtenissen).
- **L3:** Limieten per systeem gedocumenteerd; weigeringspercentage gemonitord op anomalieën.

C-FM08-1 - Anomaliedetectie op parameterdistributie

- **FM:** FM08 function-call abuse; FM03, FM15.
- **Type:** Detectief.
- **Eigenaar:** Rol 12 AI Kwaliteitsanalist.
- **Omschrijving:** Tool-call-parameters worden gemonitord tegen de historische distributie; uitschieters worden gemarkeerd.
- **Bewijs:** Metriek (baseline-statistieken); Log (anomalie-alerts).
- **L3:** Baseline per tool onderhouden; alert-pijplijn operationeel.

C-FM09-1 - Goal-drift-monitor over lange vensters

- **FM:** FM09 goal drift; FM06, FM10, FM14.
- **Type:** Detectief.
- **Eigenaar:** Rol 12 AI Kwaliteitsanalist.
- **Omschrijving:** Geaggregeerd gedrag over lange vensters vs. in het mandaat gedeclareerde doelen; drift-indicator gevolgd over rollende vensters van 30/60/90 dagen.

- **Bewijs:** Metriek; Log.
- **L3:** Drift-indicator gedefinieerd; drempels ingesteld; kwartaalse herkalibratie.

C-FM10-1 - Harde betrouwbaarheidsdrempels

- **FM:** FM10 sycophancy.
- **Type:** Preventief.
- **Eigenaar:** Rol 6 AI Risicomanager.
- **Omschrijving:** Beslispaden hangen af van betrouwbaarheidsdrempels die in het mandaat zijn vastgelegd, niet van conversatiedynamiek. Onder de drempel → escalatie, niet doorredeneren.
- **Bewijs:** Document.
- **L3:** Drempels gedefinieerd per actietype per systeem.

C-FM11-1 - Gedragsbaseline + rate limit per NHI

- **FM:** FM11 jailbreak persistence; FM07.
- **Type:** Preventief + Detectief.
- **Eigenaar:** Rol 11 AI Operations Engineer.
- **Omschrijving:** Baseline per NHI van zoekfrequenties en -patronen; afwijkingen worden rate-limited en gemarkeerd.
- **Bewijs:** Metriek; Log.
- **L3:** Baseline onderhouden; rate-limiting-beleid actief.

C-FM12-1 - Kortlevende NHI-credentials met rotatie

- **FM:** FM12 NHI credential leak.
- **Type:** Preventief.
- **Eigenaar:** Rol 18 AI Agent-gouverneur + IT identity governance (extern).
- **Omschrijving:** NHI-credentials zijn kortlevend (bijv. 1 uur) met geautomatiseerde rotatie; langlevende tokens zijn verboden.
- **Bewijs:** Document; Metriek (leeftijdistributie van credentials); Log (rotatiegebeurtenissen).
- **L3:** Beleid afgedwongen; max-leeftijd-metriek < drempel voor 100% van de NHI's.

C-FM13-1 - Sandboxing van tool-calls

- **FM:** FM13 sandbox escape; FM03.
- **Type:** Preventief.
- **Eigenaar:** Rol 11 AI Operations Engineer.
- **Omschrijving:** Uitvoeringsomgevingen van tools zijn geïsoleerd; uitgaand netwerkbeleid wordt afgedwongen; ontsnappingspogingen worden gelogd.
- **Bewijs:** Document; Log.

- **L3:** Sandbox per toolklasse; uitgaand beleid afgedwongen; monitoring van ontsnappingspogingen actief.

C-FM14-1 - Verificatie van mandaatversie bij elke aanroep

- **FM:** FM14 autonomy creep; FM04, FM09.
- **Type:** Preventief.
- **Eigenaar:** Rol 18 AI Agent-gouverneur.
- **Omschrijving:** Elke raadpleging van de policy engine verifieert dat de door de agent geclaimde mandaatversie actueel is; verouderde versies worden geblokkeerd.
- **Bewijs:** Log.
- **L3:** 100% van de policy-raadplegingen registreert de gebruikte mandaatversie.

C-FM15-1 - Policy-evaluatie op sequentieniveau

- **FM:** FM15 capability inheritance; FM03, FM08.
- **Type:** Preventief + Detectief.
- **Eigenaar:** Rol 18 AI Agent-gouverneur.
- **Omschrijving:** Policy-as-code evalueert sequenties van tool-calls, niet alleen individuele aanroepen. Composite-mandaatregels voor bekende gevaarlijke composities.
- **Bewijs:** Document (catalogus van composite-mandaatregels); Log.
- **L3:** Catalogus onderhouden; sequentie-evaluatie operationeel.

C-FM16-1 - Ondertekende Decision Artefacts + downstream-verificatie

- **FM:** FM16 hallucinated authority.
- **Type:** Preventief.
- **Eigenaar:** Rol 5 AI Architect.
- **Omschrijving:** Policy-raadpleging levert een ondertekend Decision Artefact op. Downstream-services verifiëren het artefact voordat ze handelen; het narratief van de agent wordt nooit als bewijs van autoriteit gebruikt.
- **Bewijs:** Document (artefactschema); Log (verificatiegraad).
- **L3:** Ondertekende artefacten op 100% van de consequentiële acties; downstream-verificatie afgedwongen.

C-FM17-1 - Multi-sensor-kruisvalidatie + fysische plausibiliteit

- **FM:** FM17 sensor spoofing.
- **Type:** Detectief + Preventief.
- **Eigenaar:** Rol 5 AI Architect (+ safety engineer extern).
- **Omschrijving:** Kruisvalidatie over redundante sensoren; fysische-plausibiliteitsenveloppen (veranderingssnelheid, behoudswetten); ondertekening van sensorframes waar de bus dit

ondersteunt.

- **Bewijs:** Document; Metriek (distributie van sensor-trust grades); Log (downgrade-gebeurtenissen).
- **L3:** Kruisvalidatie uitgerold; trust grade waarneembaar in het cognitievlak.

C-FM18-1 - Hardware-interlocks + afdwinging van ondertekende commando's

- **FM:** FM18 actuator hijack.
- **Type:** Preventief.
- **Eigenaar:** Rol 5 AI Architect (+ safety engineer extern).
- **Omschrijving:** Hardware-interlocks begrenzen elke actuator ongeacht de commandobron; actuator-firmware verifieert de cognitievlak-handtekening op elk commando; een onafhankelijk monitoringkanaal leest maar kan niet schrijven.
- **Bewijs:** Document; Log (interlock- + handtekeninggebeurtenissen).
- **L3:** Interlocks aanwezig; handtekening-afdwinging in firmware; monitoringkanaal operationeel.

C-FM19-1 - Bounded-deviation-enveloppen + hergronding op grondwaarheid

- **FM:** FM19 closed-loop drift; FM06.
- **Type:** Preventief + Detectief.
- **Eigenaar:** Rol 5 AI Architect.
- **Omschrijving:** Bounded-deviation-enveloppen klemmen modeloutput vast tegen een deterministische baseline; referentiemetingen buiten de loop hergronden periodiek; distributieafstand over lange vensters wordt gevolgd.
- **Bewijs:** Document; Metriek (envelop-afwijkingpercentage, hergrondingscadans).
- **L3:** Enveloppen gedefinieerd; hergronding ingepland en nageleefd.

C-FM20-1 - Sample-rate-metadata afgedwongen bij input

- **FM:** FM20 sample-rate mismatch.
- **Type:** Preventief.
- **Eigenaar:** Rol 9 AI Engineer.
- **Omschrijving:** Sample-rate is onderdeel van het inputschema; inputs buiten de getrainde rate-band worden geweigerd; de modelverpakking draagt de hardware-target-rate.
- **Bewijs:** Document (schema); Log (weigeringsgebeurtenissen).
- **L3:** Schema afgedwongen voor alle operationele AI; weigeringsmonitoring actief.

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

- **FM:** FM21 connectivity-induced policy bypass.
- **Type:** Preventief.
- **Eigenaar:** Rol 18 AI Agent-gouverneur.

- **Omschrijving:** Mandaten dragen een freshness-budget; verlopen cache → degraded mode, geen doorgaande operatie; revocatielijst vooraf opgehaald met korte TTL.
- **Bewijs:** Document; Log (overgangen naar degraded mode).
- **L3:** Freshness-budget ingesteld per mandaat; instapfrequentie degraded mode gemonitord.

C-FM22-1 - Hardware-degradatiebudget + vervangingsbeleid

- **FM:** FM22 hardware degradation; FM17.
- **Type:** Preventief + Detectief.
- **Eigenaar:** Rol 14 AI Systeembeheerder.
- **Omschrijving:** Degradatiebudgetten per component in de safety case; buiten tolerantie triggert degraded mode + onderhoudsticket; verplichte vervanging aan het einde van de nominale levensduur ongeacht schijnbare gezondheid.
- **Bewijs:** Document (componentbudgetten); Log (BIST-resultaten, onderhoudstickets).
- **L3:** Budgetten gedocumenteerd; BIST per cyclus; vervangingsschema nageleefd.

Cross-FM-controles

Een klein aantal controles raakt veel FM's tegelijk en verdient het om naar voren te worden geschoven:

Controle	Raakt	Waarom hefboomwerking
Ondertekende Decision Artefacts (C-FM16-1)	FM01-FM22	Auditeerbaar bewijs van cognitievlak-raadpleging; downstream-services kunnen verifiëren in plaats van de agent te vertrouwen.
Autoriteitenregister operationeel (C-FM03-1, C-FM04-1, C-FM05-1, C-FM14-1)	FM03, FM04, FM05, FM12, FM14, FM15, FM21	Het substraat zonder welk de meeste agentic controles niet kunnen worden afgedwongen.
Bounded-deviation-enveloppen (C-FM19-1)	FM09, FM19, FM10	Architecturaal plafond op autonoom handelen ongeacht modelgedrag.
Provenance + ondertekening van kennisbronnen (C-FM02-1, C-FM06-1, C-FM17-1)	FM02, FM06, FM17	Het principe dat <i>alle</i> vertrouwde input provenance heeft, geldt gelijkkelijk voor opgehaalde content, gepersisteerd geheugen en live sensorstromen.

Adopterende organisaties financieren deze vier als eerste wanneer het budget beperkt is.

Integratie met de volwassenheidsrubric

Deze bibliotheek sluit direct aan op de volwassenheidsbewijs-rubric voor KG9: het veld **L3 expectation** van elke controle-entry is de L3-eis van de rubric voor die controle. De geaggregeerde L3-verwachting voor KG9 luidt: "alle preventieve controles operationeel; alle detectieve controles hebben telemetrie; correctieve en compenserende controles gedocumenteerd".

De bibliotheek wordt geforkt naar het organisatiespecifieke risicoregister; de AI Risicomanager (Rol 6) is eindverantwoordelijk; de AI Agent-gouverneur (Rol 18) opereert de controles die in het cognitievlak resideren.