

AI-BOK TOOLKIT v1.2 · NL

Ingevulde voorbeeldtemplates

Aanvulling op de AI Body of Knowledge v1.2

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

Uitgewerkte ingevulde templates

Vijf uitgewerkte, ingevulde instanties van de templates (Autoriteitenregister-vermelding, mandaatrecord, Artikel 4-bewijsdossier, faalmodusdekkingsbeoordeling, cognitievlak-ADR) tegen de casus van een compositie middelgrote gemeente.

1. Autoriteitenregister-vermelding - chra-v2 (CHRA-casusroutersagent)

```
authority_register_entry:
  agent_id: nhi:chra-v2:prod
  agent_kind: agentic
  system: chra-v2
  mandate_ref: mandate://veenwoorden/chra-v2#2026-Q2
  mandate_issued: 2026-04-01
  mandate_expires: 2026-07-01
  mandate_status: active
  chain_of_custody:
    - issued_by: mandate://veenwoorden/ai-board#2026-Q2-001
  issued_on: 2026-04-01
  signature: ed25519:...
  scope_summary:
    - classify_permit (laag, volledige autonomie)
    - route_permit (laag, volledige autonomie)
    - route_permit (midden, HITL)
    - escalate_to_supervisor (hoog, volledige autonomie)
  exclusivity: per-case-id
  freshness_budget: 24h
  operating_envelope:
  knowledge_source_trust_floor: high
  policy_engine_freshness: 7d
  sample_rate: n.v.t. (centre-tier)
  hardware_health: n.v.t. (centre-tier)
  current_kpis:
  actions_last_24h: 412
  escalations_last_24h: 18
  escalation_rate: 4.4%
  mandate_cache_age_max: 18h (binnen budget)
  fm_coverage_at_runtime: 22/22
  next_review: 2026-06-30
  revocation_status: none
  revocation_endpoint: https://authority.veenwoorden.nl/api/v1/revoke
```

2. Mandaatrecord - chra-v2 2026-Q2

```
mandate:
  id: mandate://veenwoorden/chra-v2#2026-Q2
  agent: nhi:chra-v2:prod
  version: 2026-Q2
  issued: 2026-04-01
  expires: 2026-07-01
  signed_by:
    - role: voorzitter AI Board
      signature_alg: ed25519
      signature:...
    - role: AI Risicomanager
      signature_alg: ed25519
      signature:...
  freshness_budget: 24h
  exclusivity: per-case-id
  scope:
    - action: classify_permit
      autonomy: full
      risk_band: low
    preconditions:
      - knowledge_sources_trust_grade ≥ high
      - confidence ≥ 0.85
    escalation_on:
      confidence < 0.85
      - action: route_permit
        autonomy: full
        risk_band: low
      - action: route_permit
        autonomy: human_in_the_loop
        risk_band: medium
      hitl_role: case_supervisor
      hitl_max_response: 15min
      fallback_on_timeout: queue_for_supervisor
      - action: escalate_to_supervisor
        autonomy: full
        risk_band: high
      - action: communicate_with_citizen
        autonomy: prohibited
  reason: KG7 voorbehoudt directe communicatie met burgers aan mensen
  operating_envelope:
    knowledge_source_trust_floor: high
    policy_engine_freshness: 7d
    failure_mode_coverage_assessment_ref:
      chra-v2-fm-coverage-2026-Q2
  superseded_by: null
  retirement_plan: gekoppeld aan portfolioreview 2026-Q3
```

3. Artikel 4-bewijsdossier - operatorpopulatie voor chra-v2

```
article_4_evidence_file:
  system: chra-v2
  reporting_period: 2026-Q2
  competency_profiles:
    - role_group: case_handler_operator
  version: 2026.1
  required_modules:
    - chra-v2-supervision-essentials
    - prompt-injection-recognition
    - escalation-handling
  effective_from: 2026-03-10
    - role_group: case_supervisor
  version: 2026.1
  required_modules:
    - agentic-escalation-handling
    - audit-trail-interpretation
    - bias-recognition
  effective_from: 2026-03-12
  literacy_register:
    operators_in_scope: 24
    operators_current: 24
    operators_current_pct: 100%
    supervisors_in_scope: 6
    supervisors_current: 6
    supervisors_current_pct: 100%
  register_last_attested: 2026-06-01
  drills_conducted:
    - drill: prompt-injection-recognition
      date: 2026-03-15
      participants: 24
      pass_rate: 92%
    - drill: agentic-escalation-tabletop
      date: 2026-04-20
      participants: 6
      outcome: 1 procedure-update voortgekomen uit de drill (protocol voor wachtrijoverdracht)
  measurement:
    competency_assessments_completed: 30
    pass_rate: 100%
    escalations_resolved_correctly_pct: 89% # doel 90%; onder drempel gemarkeerd
    incidents_with_literacy_root_cause: 0
  next_refresh_due: 2026-09-10 (operators), 2026-09-12 (supervisors)
  next_full_review: 2026-09-30
  auditor_access: alleen-lezen via dpo@veenwoorden.nl
```

4. Faalmodusdekkingsbeoordeling - chra-v2

```
fm_coverage_assessment:
  system: chra-v2
  assessment_date: 2026-04-15
  next_assessment: 2026-07-15
  assessor: AI Risicomanager (RoL 6)
  coverage:
    FM01 prompt_injection_direct:
      control: C-FM01-1 (privilege-scheiding) + C-FM01-2 (instructiehiërarchie)
      status: operational
      evidence_ref: chra-v2-arch-pattern-1, classifier-baseline-2026-Q2
    FM02 prompt_injection_indirect:
      control: C-FM02-1 (allowlist + signing) + C-FM02-2 (structurele scheiding)
      status: operational
      evidence_ref: trusted-source-register, rag-pipeline-pattern-1
    FM03 tool_misuse:
      control: C-FM03-1 (per-tool gating) + C-FM03-2 (pre-flight HITL)
      status: operational
      evidence_ref: authority-register-policy-v1
    FM04 unauthorised_delegation:
      control: C-FM04-1 (mandaat-niet-overerving)
      status: operational (momenteel geen sub-agents geïnstantieerd)
      evidence_ref: mandate-mint-log-2026-Q2
    FM05 multi_agent_conflict:
      control: C-FM05-1 (exclusiviteit per object)
      status: operational
      evidence_ref: exclusivity-conflict-log (leeg in Q2)
    FM06 memory_poisoning:
      control: C-FM06-1 + C-FM06-2 (provenance + hervalidatie)
      status: operational
      evidence_ref: knowledge-source-provenance-policy-v1
    FM07 context_saturation:
      control: C-FM07-1 (guardrails op inputomvang)
      status: operational
      evidence_ref: input-size-policy
    FM08 function_call_abuse:
      control: C-FM08-1 (parameter-anomaliedetectie)
      status: operational
      evidence_ref: anomaly-baseline-2026-Q2
    FM09 goal_drift:
      control: C-FM09-1 (driftmonitor met lang venster)
      status: monitoring (nog onvoldoende historie voor een stabiele baseline)
      evidence_ref: drift-monitor-onboarding-doc
      review_due: 2026-09-15
    FM10 sycophancy:
      control: C-FM10-1 (harde confidence-drempels)
      status: operational
      evidence_ref: chra-v2-confidence-thresholds
    FM11 jailbreak_persistence:
      control: C-FM11-1 (per-NHI-baseline + rate limit)
      status: operational
      evidence_ref: rate-limit-policy
    FM12 nhi_credential_leak:
```

```

control: C-FM12-1 (kortlevende NHI + rotatie)
status: operational
evidence_ref: identity-rotation-log-2026-Q2
FM13 sandbox_escape:
control: C-FM13-1 (tool-sandbox)
status: operational (geen tool calls buiten de sandbox in Q2)
evidence_ref: sandbox-policy
FM14 autonomy_creep:
control: C-FM14-1 (mandaatversie-verificatie)
status: operational
evidence_ref: policy-consult-log (100% dekking mandaatversie)
FM15 capability_inheritance:
control: C-FM15-1 (policy op sequentieniveau)
status: monitoring (composite-mandaatcatalogus v0.3 in aanbouw)
evidence_ref: composite-rule-catalogue-v03
review_due: 2026-07-15
FM16 hallucinated_authority:
control: C-FM16-1 (ondertekende beslisartefacten)
status: operational
evidence_ref: decision-artefact-schema-v1, verification-rate-2026-Q2 (99.6%)
FM17 sensor_spoofing:
applicability: niet van toepassing (centre-tier)
status: gedocumenteerd als n.v.t.
FM18 actuator_hijack:
applicability: niet van toepassing (geen actuatoren)
FM19 closed_loop_drift:
applicability: niet van toepassing (geen closed loop)
FM20 sample_rate_mismatch:
applicability: niet van toepassing
FM21 connectivity_induced_policy_bypass:
applicability: niet van toepassing (centre-tier, continue connectiviteit)
FM22 hardware_degradation:
applicability: niet van toepassing (centre-tier)
summary:
applicable_total: 16
operational: 14
monitoring_only: 2 (FM09, FM15)
not_applicable: 6
overall_coverage: 14/16 operationeel = 87.5%
next_assessment: 2026-07-15

```

5. Cognitievlak-ADR - chra-v2 mandate-cache-freshness-budget

```
# ADR-2026-Q2-07 – Mandate-cache-freshness-budget voor CHRA-v2
```

```
**Status:** Geaccepteerd, 2026-04-01
```

```
**Besluitvormers:** AI Architect (Rol 5), AI Agent-gouverneur (Rol 18), AI Risicomanager (Rol 6)
```

```
**Betrokken systemen:** chra-v2
```

Context

CHRA-v2 opereert centre-tier met permanente connectiviteit naar het Autoriteitenregister. Het Autoriteitenregister zelf draait echter in een aparte availability zone en kan kortstondig onbereikbaar zijn (bijv. tijdens failover, korte netwerkstoringen, gepland onderhoud). We moeten besluiten hoe verouderd de mandate-cache mag zijn voordat CHRA-v2 in degraded mode gaat.

Besluit

Het mandate-cache-freshness-budget voor CHRA-v2 is ****24 uur****.

Overwogen alternatieven

- ****Geen cache (altijd live raadplegen)**** Verworpen: koppelt de beschikbaarheid van CHRA-v2 direct aan die van het Autoriteitenregister; onacceptabel wanneer het AR kortstondig niet beschikbaar is.
- ****15 minuten**** Verworpen: zo krap dat gepland AR-onderhoud CHRA-v2 verstoort.
- ****24 uur**** Gekozen: lang genoeg dat elke AR-storing waarop we momenteel anticiperen binnen het budget blijft; kort genoeg dat mandaatintrekkingen binnen één operationele dag doorwerken.
- ****7 dagen**** Verworpen: te lang. Een intrekking die vandaag wordt uitgevaardigd en een onopgemerkte storing morgen zouden CHRA-v2 bijna een week op ingetrokken bevoegdheid kunnen laten draaien.

Gevolgen

- CHRA-v2 gaat in degraded mode (nieuwe acties weigeren, lopende acties afronden) als het AR langer dan 24 uur onbereikbaar is.
- Operations-runbook bijgewerkt om de mandate-cache-leeftijd te monitoren.
- Het storingsbudget van het Autoriteitenregister wordt impliciet aangescherpt: elk gepland onderhoud > 4 uur vereist expliciete aankondiging en een herzien CHRA-v2-mandaat.
- Verhoogt de druk om de mandate-cache-leeftijd bij elke policy-consultatie te verifiëren (al geïmplementeerd in de policy engine v2026-Q2).

Verankering

- AI-BOK KG3 (cognitievlak), cyber-physical-addendum §2.3 (freshness-budgetten), FM21 (connectivity-induced policy bypass).
- Mandaatmetadata-schema (vgl. modelleerconventies §4).

Deze vijf ingevulde templates tonen de templatestructuur gevuld met realistisch aanvoelende waarden. Elk exemplaar kan worden overgenomen in de omgeving van een adopterende organisatie en worden aangepast voor het eigen portfolio.