



id: SI_009

model: anthropic/claude-opus-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: samarbeid

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: syklisk

Instans SI_009

Seeds: strategisk | syklisk | samarbeid | sårbarhet | implementer

Kjerneanalyse: Kontroll som syklisk samarbeidsprosess, ikke statisk arkitektur

De fleste rammeverk behandler agentkontroll som et engangsproblem — definer policyer, implementer, ferdig. **Det er feil.** Agentisk KI i et heterogent systemlandskap krever en syklisk kontrollmodell der sårbarhetene kontinuerlig avdekkes, evalueres og adresseres i takt med at agentene lærer og miljøet endres.

Strategisk dissens: Ikke bygg policy-sandkasser — bygg et handlingsbudsjett

Policy-sandkasser gir falsk trygghet. Eldre fagsystemer uten fullverdig integrasjonsstøtte kan ikke håndheve sandkasseregler pålitelig. I stedet: **implementer et syklisk handlingsbudsjett-system.**

- Hver agent får et kvantifisert handlingsbudsjett per syklus (f.eks. maks 50 skriveoperasjoner/time, 0 irreversible handlinger uten eskalering).
- Budsjettet justeres syklisk basert på observert atferd — ukentlig review i samarbeid mellom sikkerhets-, fag- og KI-team.
- Overskridelse trigger automatisk degradering til read-only-modus, ikke bare logging.



Sårbarhetskartlegging: De reelle blindsonene

Blindsone	Konkret risiko	Tiltak
Eldre fagsystemer uten audit-trail	Agent endrer data uten sporbarhet — rollback umulig	Implementer proxy-lag med write-ahead-log foran hvert fagsystem
RBAC uten kontekstsensitivitet	Agent med riktig rolle gjør riktig handling i feil kontekst (f.eks. endrer vedtak kl. 03:00 uten saksbehandler)	Innfør tidsvindu- og kontekst-attributter via en lettvekts OPA (Open Policy Agent)-instans ved siden av IAM
Kjedet API-kall mot eksterne tjenester	Sensitiv informasjon aggregeres på tvers av kall og eksponeres i mellomlagring	Implementer ephemeral data-kontekst som nullstilles etter hver agentsyklus
Samarbeid mellom agenter	Én agent eskalerer til en annen som har høyere rettigheter — uintendert privilege escalation	Flat tillitsmodell: ingen agent kan tildele eller arve rettigheter fra annen agent

Implementeringsmodell: 3-syklus pilot

Syklus 1 (måned 1-3): Read-only agenter. Full logging via sentralisert proxy. Kartlegg reell systemsårbarhet. **Syklus 2 (måned 3-6):** Skriveoperasjoner med handlingsbudsjett. Human-in-the-loop for irreversible handlinger. Ukentlig samarbeidsmøte mellom sikkerhet og fagseksjon. **Syklus 3 (måned 6-9):** Begrenset autonomi for validerte handlingsmønstre. Automatisert anomalideteksjon basert på baseline fra syklus 1-2.

Neste steg

1. **Implementer proxy-lag med write-ahead-log** foran de 3 mest kritiske fagsystemene innen 4 uker — dette er ufravikelig grunnmur for alt annet.
2. **Etabler tverrfaglig syklisk review-team** (sikkerhet + fag + KI) med ukentlig kadense fra dag 1 — kontroll uten samarbeid skaper blindsoner.
3. **Deploy OPA som policy-motor** ved siden av eksisterende IAM — dette gir kontekstsensitiv autorisasjon uten å erstatte RBAC, og kan stå på 2-3 uker.