



id: SI_009

model: anthropic/claude-opus-4.6

seeds:

04_risiko: usikkerhet

03_relasjon: tillit

05_handling: implementer

01_perspektiv: økonomisk

02_tidhorisont: kortsiktig

Instans SI_009

Seeds: økonomisk | kortsiktig | tillit | usikkerhet | implementer

Dissens: Ikke bygg arkitekturen først — bygg tillit gjennom kontrollert implementering

Mainstream-rådet er å etablere Zero Trust, OPA, sentralisert logging og full IAM-redesign før første agent går live. Med 6–9 måneders pilotvindu og begrensede ressurser er dette en oppskrift på at ingenting blir implementert. Kostnaden ved å vente er reell: dere mister læringsdata, organisatorisk momentum og politisk vindu.

Økonomisk realisme som styrende prinsipp

En mellomstor offentlig virksomhet har typisk 2–4 dedikerte sikkerhetsressurser. Full Zero Trust-arkitektur koster 8–15 MNOK og 18+ måneder. **Dere har ikke råd til dette som forutsetning.** I stedet: invester 500k–1M i en minimal kontrollamme som vokser iterativt.

Kortsiktig implementeringsstrategi (0–9 mnd)

Måned 1–2: Én agent, ett system, human-in-the-loop

- Velg én prosess med lav integritetsrisiko (f.eks. statusoppslag, rapportgenerering — ikke dataendring)
- Agenten foreslår, menneske godkjenner. Hvert kall logges med prompt, respons og tidsstempel i en enkel append-only-logg (selv en PostgreSQL-tabell duger)



- Service account med hardkodet scope: kun GET mot spesifikt API-endepunkt, token med 1-times levetid

Måned 3-5: Utvid med policy-guard

- Implementér OPA som sidecar på agentens execution-lag — kun for denne agenten. Ikke enterprise-utrulling
- Definer 10-15 eksplisitte deny-regler fremfor å forsøke komplett allow-modell
- Innfør circuit breaker: >5 kall/minutt → automatisk stopp

Måned 6-9: Kontrollert skrivetilgang

- Tillat muterende operasjoner kun med compensating action definert på forhånd
- Alle skriveoperasjoner via mellomlag som skriver til staging-tabell, ikke direkte mot fagsystem

Usikkerhet som permanent tilstand

Dere vil aldri eliminere emergent behavior-risiko fullstendig. Den farligste blindsonen er **latent kobling mellom legacy-systemer** — en agent som oppdaterer felt X i monolitten kan trigge batch-jobb Y som ingen husker eksisterer. Modeller dette med pre-mortem: «agenten endret noe, og 48 timer senere oppdaget vi...»

Tillit bygges med beviser, ikke arkitekturdokumenter

Ledelse, datatilsyn og fagsystem-eiere trenger observerbare bevis på kontroll. En fungerende audit-logg fra pilot med 500 reelle transaksjoner er mer tillitvekkende enn en 80-siders sikkerhetsarkitektur.

Neste steg

1. **Denne uken:** Identifiser én konkret prosess med lav risiko og høy synlighet. Definer scope som 3-5 tillatte API-kall. Estimer kostnad for minimal logging-pipeline (trolig <100 timer).
2. **Innen 30 dager:** Opprett dedikert service account med eksplisitt deny-all-default, whitelist kun pilot-endepunkter. Implementér kill-switch (manuell og automatisk ved anomali).
3. **Innen 60 dager:** Gjennomfør pre-mortem-workshop med systemforvaltere for legacy-systemene agenten berører — kartlegg skjulte triggere, batch-jobber og koblinger som ikke finnes i dokumentasjonen.