



id: SI_001

model: anthropic/claude-opus-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: tillit

05_handling: implementer

01_perspektiv: operasjonell

02_tidhorisont: langsiktig

Instans SI_001

Seeds: operasjonell | langsiktig | tillit | sårbarhet | implementer

Kjerneanalyse: Tillit er en driftsegenskap, ikke en policy-erklæring

Dissenspoeng først: Mainstream-tilnærmingen sier «start med governance-rammeverk, deretter implementer». Jeg snur det. **Implementer en minimal agent i produksjonslikt miljø innen uke 6 — fordi sårbarhetene dere trenger å forstå er operasjonelle, ikke teoretiske.**

Sårbarhetslaget ingen snakker om

Deres eldre fagsystemer har implisitte tillitsmodeller. De ble bygget med antagelsen «et menneske sjekker skjermbildet før innsending». Når en agent kaller et SOAP-endepunkt i et fagsystem fra 2009, finnes ingen semantisk validering av om handlingen er rimelig — kun om den er teknisk gyldig. Dette er den farligste blindsonen: **agenten kan utføre tusenvis av teknisk gyldige, men operasjonelt meningsløse handlinger** før noen oppdager det. Fragmentert logging gjør at gjennomsnittstid til deteksjon (MTTD) kan bli uker.

Operasjonell kontrollarkitektur

Implementer et **Agent Gateway**-lag (tilsvarende API Gateway, men handlingsorientert):

- Handlingsklassifisering** — Hver agenthandling tagges: `lesbar`, `reversibel-skriv`, `irreversibel-skriv`. Irreversible handlinger krever menneskelig godkjenning i pilotperioden. Null unntak.



2. **Tillitsbudsjett per sesjon** — Ikke bare rate-limiting, men et kumulativt risikobudsjett. Agent som har endret 3 dataposter får lavere budsjett for neste handling. Inspirert av Google's BeyondCorp, men for agenter.
3. **Replaybar hendelseslogg** — Sentraliser all agentinteraksjon i en append-only event store (f.eks. Apache Kafka med immutable topics). Dette løser fragmentert logging og gir rollback-grunnlag.
4. **Policy-sandkasse via OPA (Open Policy Agent)** — Kobles til Agent Gateway. RBAC i eksisterende IAM dekker «hvem», OPA dekker «hva, når, hvor mye, i hvilken kontekst». Kan implementeres uten å bytte IAM-plattform.

Langsiktig tillitsmodell

Tillit til agentene må opptjenes operasjonelt over tid. Konkret: definer tillitsnivåer (T0–T3) der agenten starter på T0 (kun leseaksess) og eskaleres basert på revisjonsdata etter 30/60/90 dager. Dokumenter eskaleringskriteriene nå.

Undervurderte risikoer

- **Prompt injection via eksterne API-responser** — agenten parser data fra tjenester dere ikke kontrollerer. Sanitize alt.
- **Kaskadesvikt** — én feilhandling trigger automatisk kompenserende handlinger som selv feiler. Implementer circuit breakers per integrasjon.
- **Nøkkelrotasjon** — agentenes service accounts må ha 24-timers token-levetid i pilot. Ikke 90 dager.

Neste steg

1. **Uke 1-2:** Implementer Agent Gateway som proof-of-concept med handlingsklassifisering mot ett fagsystem. Mål MTTD for simulerte feilhandlinger.
2. **Uke 3-4:** Deploy OPA med 5 konkrete policies (maks antall skriveoperasjoner, blokkering av irreversible handlinger, dataklassifiseringskontroll) og test mot reelle scenarier.
3. **Uke 5-6:** Gjennomfør en red team-øvelse der sikkerhetsteamet forsøker å få agenten til å eksfiltrere data eller utføre kaskadehandlinger. Resultatene definerer resten av piloten.