



id: SI_005

model: anthropic/claude-opus-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: makt

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: kortsiktig

Instans SI_005

Seeds: strategisk | kortsiktig | makt | sårbarhet | implementer

Makt uten kontroll er den reelle risikoen — ikke teknologien

La meg være direkte: Dere er i ferd med å delegere **makt** — evnen til å endre data, trigge prosesser, kalle eksterne tjenester — til en aktør uten dømmekraft. I et landskap med legacy-monolitter uten deterministiske kontrakter er dette ikke et arkitekturprosjekt. Det er et **sårbarhetsskapende** prosjekt dersom det gjøres feil.

Den kortsiktige sannheten

Innenfor 6–9 måneder kan dere **ikke** bygge Zero Trust, sentralisert policy-motor, konsolidert audit-pipeline og safe execution environments. Alle som sier noe annet, selger. Dere må velge **én kontrollflate** og implementere den knallhardt.

Min dissens: Ikke start med arkitekturprinsipper. Start med å kartlegge de 5–10 konkrete handlingene agenten skal utføre, og bygg et **eksplisitt allow-list-regime**. Alt annet er blokkert. Ingen default-tilgang, ingen implisitt tillit.

Konkret implementeringsmodell

1. Agent Proxy Gateway (uke 1–6) Sett en dedikert proxy foran alle agent-interaksjoner. Agenten snakker aldri direkte med fagsystemer. Proxyen har:

- Hardkodet allow-list per API-endepunkt og HTTP-metode



- Rate-limiting: maks 50 kall/minutt, maks 5 skriveoperasjoner/minutt
- Circuit breaker som utløses ved >3 feil på 60 sekunder
- Fullstendig request/response-logging til immutable store (f.eks. append-only PostgreSQL eller S3 med Object Lock)

2. Ephemeral credentials med 15-minutters TTL (uke 4-8) Ingen persistent service account. Agenten får scoped tokens fra HashiCorp Vault med eksplisitt policy-binding per sesjon. Nøkkelrotasjon hver 15. minutt. Ved anomali: revoke umiddelbart.

3. Kill switch og blast radius (uke 2-4) Implementer en **hard kill switch** — et enkelt API-kall eller Slack-kommando som terminerer alle aktive agent-sesjoner innen 2 sekunder. Definer blast radius: agenten får aldri tilgang til systemer med integritetsklasse høy (folkeregister, økonomi, helse) i piloten.

Sårbarhetene dere undervurderer

- **Latent coupling:** Agenten kaller API A som trigger webhook til system B som skriver til database C. Dere ser kun kall A. Kjedereaksjonen er usynlig uten end-to-end correlation IDs.
- **Emergent loops:** Agent leser feilstatus → forsøker retting → skaper ny feil → leser igjen. Uten loop-deteksjon (maks 3 retries per unik handling per 10-minuttersvindu) kan dette eskalere på sekunder.
- **Maktkonsentrasjon:** Én agent med bred tilgang er én kompromittert credential fra katastrofe.

Neste steg

1. **Denne uken:** Kartlegg eksakt hvilke 5-10 handlinger piloten krever. Definer allow-list. Alt utenfor er blokkert — ingen unntak.
2. **Innen 30 dager:** Implementer Agent Proxy Gateway med logging, rate-limiting og kill switch. Ingen agent-interaksjon uten denne proxyen.
3. **Innen 60 dager:** Gjennomfør en «red team»-øvelse der et internt team forsøker å få agenten til å eskalere privilegier, skape loops eller nå systemer utenfor scope. Dokumenter funnene som grunnlag for fase 2.