



id: SI_007

model: anthropic/claude-opus-4.6

seeds:

04_risiko: mulighet

03_relasjon: tillit

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: historisk

Instans SI_007

Seeds: strategisk | historisk | tillit | mulighet | implementer

Historisk mønster: Autonomi innført for raskt ender i tilbaketrekking

Hver gang offentlig sektor har innført autonome systemer uten tillitsarkitektur først — fra NAVs regelmotor-feil til Storbritannias Post Office-skandalen (Horizon) — har resultatet vært det samme: tap av tillit som tar tiår å gjenoppbygge. Mønsteret er klart: **teknisk kapabilitet uten etterprøvbar kontroll skaper institusjonell risiko som overstiger den operasjonelle gevinsten.**

Strategisk dissens: Ikke implementer agentisk KI — implementer en tillitsinfrastruktur

Mainstream-tilnærmingen sier «start med en pilot, iterer raskt.» Jeg er uenig. Med 6–9 måneders tidshorisont og begrenset sikkerhetskapasitet bør dere **ikke** slippe autonome agenter løs på produksjonssystemer. I stedet: bruk pilotperioden til å bygge kontrollplanet som muliggjør agentisk KI etterpå.

Konkret kontrollarkitektur: «Trust Broker»-modellen

Én ny komponent — en Trust Broker — mellom agenter og alle systemer:

1. **Policy-sandkasse via OPA (Open Policy Agent):** Definer handlingsrom som kode. Hvert API-kall fra en agent evalueres mot deklarativer regler. Eksempel:



`agent.may("update", "vedtak") = false` for irreversible handlinger. OPA kompensere for IAM-plattformens manglende finmasket polycystøtte.

2. **Tillitsgradering per handling:** Klassifiser alle agentoperasjoner i tre nivåer: observe (les), propose (foreslå med human-in-the-loop), execute (autonom). I piloten: **null handlinger på execute-nivå**. Dette er ikke en flaskehals — det er strategisk risikostyring.
3. **Unified audit log:** Innfør OpenTelemetry med obligatorisk correlation-ID per agentkjede. Alle fagsystemer instrumenteres med en tynn proxy som logger til én sentral hendelsesstrøm. Estimat: 4-6 uker for de 3-4 mest kritiske systemene.
4. **Nøkkelrotasjon og kortlevde tokens:** Agenter får SPIFFE-identiteter med maks 15 minutters levetid. Ingen statiske API-nøkler. Eldre fagsystemer uten tokenstøtte tilgjengeliggjøres kun via Trust Broker med rate-limiting (maks 50 kall/minutt).

Blindsoner dere undervurderer

- **Prompt injection via eksterne API-responser** — en agent som konsumerer tredjeparts-API kan manipuleres til å endre interne data
- **Kaskadeeffekter** — én feilhandling i et fagsystem uten rollback-støtte propagerer gjennom agentkjeden
- **Ansvarspulverisering** — uten eksplisitt logg over hvorfor agenten handlet, kan ingen holdes ansvarlig

Neste steg

1. **Uke 1-2:** Kartlegg alle systemintegrasjoner og klassifiser hver operasjon som observe/propose/execute. Ingen agenter får execute-rettigheter i piloten.
2. **Uke 3-8:** Implementer Trust Broker med OPA og SPIFFE. Deploy som obligatorisk gateway — ingen direkte agent-til-system-kommunikasjon.
3. **Uke 9-12:** Kjør agenter i propose-modus med menneskelig godkjenning. Mål: bygg 90 dagers revisjonsdata som grunnlag for å bevise tillitsverdighet før autonomi vurderes.

Kjernebudskap: Muligheten agentisk KI representerer er reell — men den realiseres kun gjennom en tillitsinfrastruktur bygget før autonomien slippes løs.