



id: SI_005

model: anthropic/claude-opus-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: konkurranse

05_handling: vent

01_perspektiv: operasjonell

02_tidhorisont: historisk

Instans SI_005

Seeds: operasjonell | historisk | konkurranse | sårbarhet | vent

Hovedperspektiv: Vent — men vent strategisk, ikke passivt

Historisk lærdom: Hver gang offentlig sektor har adoptert autonome systemer under tidspress, har det gått galt. NAVs Hubro-prosjekt, britiske Post Office Horizon-skandalen, Australias Robodebt — alle deler én fellesnevner: autonome prosesser fikk endre data og initiere vedtak i systemer designet for menneskelig saksbehandling, uten at sporbarhet og rollback var på plass først. Dere beskriver nøyaktig samme utgangspunkt.

Min dissens: 6-9 måneder er ikke en pilot — det er en felle.

Med fragmentert logging, RBAC uten finmasket policy-styring, og eldre fagsystemer uten integrasjonsstøtte mangler dere tre av fire bærebjelker for ansvarlig agentisk KI. Å slippe autonome agenter inn i dette er som å gi noen nøkler til alle rom i bygget og håpe at de bare åpner de riktige dørene.

Sårbarhetsanalyse — de reelle blindsonene

- Implisitt tillitsarv.** Agenter som autentiserer via tjeneste-kontoer arver ofte bredere rettigheter enn tiltenkt i RBAC-systemer. Uten attributtbasert (ABAC) eller policy-basert kontroll (OPA/Cedar) vil agenten operasjonelt ha saksbehandler-nivå tilgang — kontinuerlig, uten pauser, 24/7.
- Eldre fagsystemer mangler idempotens.** Når en agent kaller et legacy-API som ikke støtter idempotente operasjoner, blir rollback teknisk umulig. Én feilaktig dobbel-posting i et økonomisystem er nok.



3. **Prompt injection via eksterne API-er.** Agenter som samhandler med eksterne tjenester eksponerer et angrepsvektor der manipulert respons-data kan endre agentens beslutningslogikk. Dette er ikke teoretisk — OWASP rangerer det som #1 LLM-risiko.
4. **Konkurransesargumentet er forgiftet.** Presset om å “ikke falle bak” driver beslutningen, men offentlig sektor konkurrerer ikke om markedsandel. Dere konkurrerer om tillit. Én alvorlig hendelse setter digitalisering tilbake flere år.

Hva “vent” betyr operasjonelt

Ikke vent med alt — vent med **autonomi**. Bruk 6–9 måneder på:

- Deploy agenter i **read-only/foreslå-modus** med menneske-i-løkken for alle skriveoperasjoner
- Bygg **hendelsesloggen først**: sentralisert, uforanderlig (append-only), med korrelerings-ID per agentkjede
- Implementer **policy-motor** (Open Policy Agent) som gateway foran alle systemkall — dette er forutsetningen, ikke tillegget

Konkrete neste steg

1. **Uke 1-4:** Kartlegg alle systemers idempotens-egenskaper og eksisterende tjenestekontoers faktiske rettighetsomfang. Dere vil bli overrasket over gapet mellom antatt og reell tilgang.
2. **Uke 5-12:** Implementer OPA som sentralisert policy-gateway med eksplisitt allow-list per agent-handling. Ingen agent får kalle noe som ikke er eksplisitt hvitelistet — default deny.
3. **Måned 4-9:** Kjør agenter i skyggemodus — de foreslår handlinger, mennesker utfører. Logg alt. Bruk dataene til å kalibrere policy-sandkasser før autonomi vurderes i fase 2.

Oppsummert: Den største sårbarheten her er ikke teknisk — det er tidspresset som gjør at dere hopper over fundamentet. Bygg kontrollplanet først. Autonomien kan komme senere.