



id: SI\_005

model: anthropic/claude-sonnet-4.6

seeds:

04\_risiko: mulighet

03\_relasjon: tillit

05\_handling: vent

01\_perspektiv: strategisk

02\_tidhorisont: kortsiktig

## Instans SI\_005

**Seeds:** strategisk | kortsiktig | tillit | mulighet | vent

### Hovudargument: Vent strategisk - bygg tillit først

Mainstream-rådet er å “starte smått og iterere”. Mitt perspektiv: 6–9 måneder er for kort til å bygge den tilliten agentisk KI krever i en offentlig kontekst. Piloten som planlegges er ikke en mulighet – den er en risiko forkledd som fremdrift.

### Tekniske blindsoner som undervurderes

**Fragmentert logging er ikke et teknisk problem - det er et tillitsproblem.** Når agenter tar autonome valg på tvers av fagsystemer uten fullstendig revisjonsspor, mister dere ikke bare sporbarhet – dere mister evnen til å bevise at systemet oppførte seg lovlig. For en offentlig virksomhet er dette eksistensielt.

Konkret risiko: En agent med RBAC-tilgang i IAM-plattformen men uten finmasket policy-styring vil operere med effektiv overautorisasjon. Den vil ha tilgang til det rollen tillater, ikke det oppgaven krever. I et legacy-system uten moderne API-støtte betyr dette ukontrollerte sideeffekter.

### Hva kortsiktig tenkning koster dere strategisk

- **Rate-limiting** på tvers av systemer er umulig uten sentralisert policy-motor (f.eks. OPA/Cedar). Uten dette: ingen throttling, ingen eskalering, ingen rollback.
- **Nøkkelrotasjon** for KI-drevne prosesser krever secrets management (Vault, AWS Secrets Manager). Sannsynligvis ikke på plass i dag.



- **Ekstern API-eksponering:** Agenter som kaller eksterne tjenester kan lekke kontekstuell informasjon selv uten å sende eksplisitte data – prompt-strukturer, timing, mønster.

---

## Den strategiske muligheten dere faktisk har

Bruk de 6–9 månedene til noe mer verdifullt enn en halvferdig pilot:

1. **Kartlegg handlingsrom eksplisitt** – definer hvilke handlinger som er reversible vs. irreversible for hvert system. Dette er arkitekturarbeid, ikke KI-arbeid.
2. **Bygg logging-infrastruktur først** – én sentralisert audit-log (f.eks. OpenTelemetry + SIEM) som dekker alle systemer agenter skal røre. Ingen agent før dette er på plass.
3. **Piloten bør være read-only** – la agenter observere og foreslå, aldri handle. Bygg tillit med ansatte og ledelse gjennom transparens, ikke autonomi.

---

## Konkrete neste steg

1. **Stopp planleggingen av handlende agenter** – redesign piloten som et beslutningsstøttesystem uten skrive tilgang (tidslinje: uke 1–2).
2. **Gjennomfør en “blast radius”-analyse** per system: hva er worst-case hvis en agent feiler? Dokumenter dette før noe annet (uke 3–4).
3. **Anskaff policy-motor** (OPA eller Cedar) og definer 10 eksplisitte “aldri gjør”-regler som teknisk håndheves – ikke bare dokumenteres.