



id: SI_001

model: anthropic/claude-sonnet-4.6

seeds:

04_risiko: mulighet

03_relasjon: tillit

05_handling: vent

01_perspektiv: operasjonell

02_tidhorisont: kortsiktig

Instans SI_001

Seeds: operasjonell | kortsiktig | tillit | mulighet | vent

Hoveddissens: Vent - men strategisk

Mainstream-rådet er å “starte smått og iterere.” Mitt perspektiv: **vent aktivt** – ikke passivt. En 6–9 måneder pilot med fragmentert logging, RBAC uten finmasket policy og eldre fagsystemer er ikke “kort tidslinje” – det er en arkitektonisk tidsbombe.

Operasjonell realitet vs. ambisjon

Det kritiske gapet er ikke teknologi, men **operasjonell tillit-infrastruktur**. Tillit til autonome agenter bygges ikke på intensjoner, men på verifiserbare kontrollpunkter. Dere mangler per nå:

- **Revisjonsbar kausalitet:** Når agent A kaller API B og endrer post C – hvem er ansvarlig? RBAC uten attributt-basert kontroll (ABAC) kan ikke svare.
- **Irreversibilitets-buffer:** Ingen rollback-mekanisme i eldre fagsystemer = én feilhandling kan koste uker med manuell retting.
- **Ekstern API-datalekkasje:** Agenter som kaller eksterne tjenester uten egress-filtering vil eksponere interne datastrukturer via request-payloads – undervurdert risiko i 90% av piloter.



Mulighetsvinduet er reelt – men smalt

Den kortsiktige muligheten er ikke å deploye agenter, men å **bygge tillit-grunnlaget** de trenger. Konkret: bruk 6 måneder på én kontrollert integrasjon (ikke tverrsystem), med full observability stack på plass først.

Teknisk minimumskrav før pilot

1. **Policy-motor** (f.eks. OPA/Open Policy Agent) foran alle agenthandlinger – ikke etterpå
2. **Sentralisert audit-log** med immutabel skrivebane (f.eks. append-only S3 + CloudTrail-mønster)
3. **Egress-proxy** med allowlist for alle eksterne API-kall fra agenter
4. **Dead man's switch**: Agent som ikke får bekreftelse innen X sekunder stopper – ikke fortsetter

Undervurdert blindsoner

Eldre fagsystemer responderer ikke deterministisk under last. En agent som tolker timeout som “ikke utført” og **gjentar handlingen** kan skape dupliserte transaksjoner. Dette er ikke edge case – det er daglig risiko.

Neste steg

1. **Gjennomfør en “blast radius”-analyse** på de 3 mest kritiske fagsystemene – hva er worst case ved en ukontrollert agenthandling?
2. **Implementer OPA som sidecar** i én avgrenset pilot-integrasjon innen uke 8 – mål policy-treff, ikke funksjonalitet
3. **Definer “tillit-terskel”**: Hvilke 5 konkrete målbare kriterier må være oppfylt før agenten får skriverettigheter – ikke leserettigheter?