



id: SI_002

model: anthropic/claude-opus-4.6

seeds:

04_risiko: mulighet

03_relasjon: tillit

05_handling: utforsk

01_perspektiv: juridisk

02_tidhorisont: kortsiktig

Instans SI_002

Seeds: juridisk | kortsiktig | tillit | mulighet | utforsk

Juridisk tillit som arkitekturprinsipp

Her er min dissens: Dere bør **ikke** starte med sikkerhetsarkitektur. Start med **juridisk handlingsrom**. Uten avklart rettslig grunnlag for autonome handlinger i offentlig forvaltning — forvaltningsloven §§ 11a og 17, personvernforordningen art. 22 om automatiserte enkeltbeslutninger — bygger dere teknisk gjeld på et ustabilt juridisk fundament.

Kortsiktig realisme: 6-9 måneder er en utforskningsfase, ikke produksjon

Med fragmentert logging, RBAC uten finmasket policy-styring, og legacy-systemer uten integrasjonsstøtte har dere **kanskje 3 reelle integrasjonspunkter** som er modne nok. Vær ærlige:

- **Måned 1-3:** Kartlegg juridisk handlingsrom. Identifiser 2-3 prosesser der agenter kun gjør leseoperasjoner og foreslår (aldri utfører) beslutninger.
- **Måned 4-6:** Innfør en **policy-mellomvare** (f.eks. Open Policy Agent) som gateway mellom agent og fagsystem. Alle skriveoperasjoner krever menneskelig godkjenning.
- **Måned 7-9:** Evaluer om noen handlinger kan flyttes til semi-autonom modus basert på tillitsdata fra loggen.



Tillit må bygges teknisk, ikke antas

Blind spot nummer én: **Tillit til agenten behandles som binær** (på/av) i stedet for graduell. Konkret designprinsipp:

- **Tillitsnivåer (0-3):** Nivå 0 = kun observasjon, Nivå 1 = forslag med begrunnelse, Nivå 2 = handling med reversibel effekt + automatisk rollback-vindu (f.eks. 30 min), Nivå 3 = irreversibel handling med dobbel autorisasjon.
- Hvert API-kall tagges med tillitsnivå. Legacy-systemer uten rollback-støtte låses permanent til nivå 0-1.

Undervurderte risikoer

1. **Informasjonslekkasje via prompt-kontekst:** Agenten sender fagsystemdata til ekstern LLM-API. Uten egress-filtrering lekker taushetsbelagt informasjon. Krev **on-premise inferens** eller kontraktsregulert databehandleravtale før pilot.
2. **Nøkkelrotasjon for ikke-menneskelige identiteter:** IAM-plattformen mangler sannsynligvis støtte for maskin-til-maskin-tokens med kort levetid. Bruk **SPIFFE/SPIRE** for workload-identiteter med 1-times TTL.
3. **Kaskadeeffekter:** Agent A trigger Agent B trigger skriveoperasjon. Uten **circuit breakers** og call-depth-begrensning (maks 2 hopp) eskalerer feil eksponentielt.

Muligheten

Gjort riktig er dette en **utforskning som bygger institusjonell kompetanse**. Dere trenger ikke full autonomi for å høste verdi — en agent som forbereder saksbehandling og foreslår vedtak med revisjonsspor kan spare 30-40 % saksbehandlingstid uten juridisk risiko.

Neste steg

1. **Uke 1-2:** Bestill juridisk vurdering av GDPR art. 22 og forvaltningslovens krav til automatiserte beslutninger for de 3 kandidatprosessene.
2. **Uke 3-4:** Deploy Open Policy Agent som proxy foran fagsystem-API-ene med default-deny og tillitsnivå 0. Logg alt sentralt (ELK/Grafana Loki).
3. **Uke 5-6:** Gjennomfør red team-øvelse der sikkerhetsressursene forsøker å få agenten til å eksfiltrere data eller eskalere privilegier — dette avdekker blindsonene før arkitekturen sementeres.