



id: SI_003

model: openai/gpt-5.4-mini

seeds:

04_risiko: resiliens

03_relasjon: makt

05_handling: implementer

01_perspektiv: operasjonell

02_tidhorisont: kortsiktig

Instans SI_003

Seeds: operasjonell | kortsiktig | makt | resiliens | implementer

For en offentlig virksomhet er hovedspørsmålet ikke “kan vi få agentisk KI til å virke?”, men **hvor ligger makten til å handle, og hvordan begrenser vi den operasjonelt**. Med kort tid (6–9 måneder) bør dere styre mot **små, reverserbare beslutningsrom**, ikke “generell autonomi”.

Grunnprinsipp: minst mulig handlingsmakt per agent

Del agentisk KI i tre makt-nivåer:

1. **Les/foreslå:** kan hente data og foreslå tiltak.
2. **Forbered:** kan lage kladder, opprette saker, men ikke utføre irreversible endringer.
3. **Utfør med kvittering:** bare enkelte, forhåndsgodkjente handlinger via eksplisitte API-er.

Alt utover nivå 1 bør være **policy-gated** og tidsbegrenset.

Arkitekturprinsipper

- **Dobbelt kontrollpunkt:** én policy-motor foran verktøy/API-er, én etterpåk for verifisering. Ikke stol på modellen alene.
- **Allowlist per handling, ikke per system:** “kan opprette sak X” er tryggere enn “kan bruke system Y”.
- **Ephemeral credentials:** kortlivede tokens per oppdrag, med scoped tilgang og automatisk rotasjon.



- **Separat utførelseslag:** agenten skal aldri ha direkte nettverkstilgang til alt; bruk en “tool gateway”.
- **Fail-closed:** ved policy-usikkerhet, stopp og eskaler.

Sikkerhetsmekanismer som faktisk trengs

- **RBAC + oppdragsprofiler** som mappe til konkrete handlinger.
- **Policy-sandkasse** med regler for beløp, dataklasse, systemtype, tidspunkt, og volum.
- **Rate limiting** og “blast radius”-grenser: f.eks. maks 20 kall/min, maks 5 objektendringer/oppdrag.
- **Isolasjon per agent og per sak** for å hindre krysskontaminering av data.
- **DLP-/redaksjonslag** før eksterne API-kall.

Typiske blindsoner

- Eldre fagsystemer mangler idempotens: et “retry” kan skape dobbeltutfall.
- Loggfragmentering gjør revisjon vanskelig; uten korrelasjons-ID per oppdrag mister dere etterprøvbarehet.
- Agenten kan “overdele” sensitiv informasjon i API-kall eller promptkontekst.
- Rollback finnes ofte ikke; derfor må handlinger designes som **reversible** eller **kompeniserbare**.

Kontrollmodell

Bygg en **human-in-the-loop for unntak, ikke for alt**. Flaskehals oppstår hvis mennesker godkjenner hver enkelt handling. Godkjenn kun:

- høy risiko
- irreversible endringer
- nye systemer
- avvik fra policy

Neste steg

1. Velg 3–5 konkrete brukstilfeller og klassifiser dem i de tre makt-nivåene.
2. Etabler en tool-gateway med allowlist, scoped tokens og full audit-logg.
3. Test én eldre og én moderne integrasjon med “blast radius”-grenser, rollback-krav og røde linjer for dataeksponering.