



id: SI_006

model: openai/gpt-5.4-mini

seeds:

04_risiko: mulighet

03_relasjon: makt

05_handling: vent

01_perspektiv: strategisk

02_tidhorisont: kortsiktig

Instans SI_006

Seeds: strategisk | kortsiktig | makt | mulighet | vent

Min linse: **ikke bygg “fri” agentisk KI først; bygg kontroll over maktutøvelsen først**. I et heterogent offentlig landskap er hovedrisikoen ikke modellfeil alene, men at agenten får **operasjonell makt** raskere enn dere får **styring, revisjon og reversering**. Derfor bør dere starte med en **begrenset mulighets-arkitektur**: små, tydelige handlingsrom, sterke stopp-mekanismer, og et bevisst “vent”-prinsipp for alt som ikke er reversibelt.

Arkitektur og kontroll

- Del agenten i 4 soner: **planlegger, policy-gate, executor, observer**. Agenten kan foreslå; kun policy-gaten kan slippe handlinger ut.
- Bruk **deny-by-default** med eksplisitte handlingsprofiler per use case: system, datasett, API, tidsrom, endringsnivå.
- Runtime-kontroll må ligge nær handlingen: **OPA/Gatekeeper/Kyverno** for container-/klyngehandling, og en tilsvarende **API-policy gateway** for legacy og eksterne tjenester.
- Modellér **execution graph** som en kjede av noder: prompt → beslutning → policy-sjekk → kall → sideeffekt → verifikasjon. Hver node får korrelasjons-ID og “reason code”.

IAM

- Bruk **ephemeral credentials** (5–15 min), scoped til én oppgave, én ressursklasse, én risikokategori.



- Service accounts per agent-rolle, ikke per modell. Nøkkelrotasjon automatisk, minst hver 24. time.
- Delegert autorisasjon bør være **tids- og kontekstavhengig**: geografi, tidspunkt, dataklassifisering, oppgave-ID.

API og risiko

- Anta at legacy API-er er **ikke-idempotente og delvis fiendtlige**. Sett rate limits, idempotency keys der mulig, og circuit breakers med hard stop etter f.eks. 3 feil eller 30 sekunder avvik.
- For eksterne kall: data-minimering, masking, egress-kontroll og eksplisitt godkjenningsliste.

Logging og hendelser

- Bygg en enkel audit-pipeline nå: beslutning, prompt-hash, policyresultat, kall, respons, sideeffekt, rollback. Ikke vent på perfekt event-sourcing.
- Ved runaway/loop: **auto-isolér agenten**, trekk tilbake tokens, freeze context, varsle menneskelig operatør.

Blindsoner

- Typisk undervurdert: latent coupling mellom systemer, “falsk suksess” i delsteg, og irreversible dataendringer med dårlig datakvalitet.
- Worst case er ikke bare feil output, men **sekvensielle feil**: én liten endring utløser flere automasjoner i gamle systemer.

Neste steg

1. Definer 3 pilot-oppgaver med lav risiko og full reverserbarhet.
2. Innfør en minimal policy-gate og ephemeral IAM før første test.
3. Lag en “kill switch + rollback”-øvelse på tvers av 2–3 systemer før pilot.