



id: SI\_007

model: openai/gpt-5.4-mini

seeds:

04\_risiko: sårbarhet

03\_relasjon: makt

05\_handling: forhandle

01\_perspektiv: operasjonell

02\_tidhorisont: syklisk

## Instans SI\_007

**Seeds:** operasjonell | syklisk | makt | sårbarhet | forhandle

Agentisk KI må behandles som en **ny aktør med makt til å flytte prosesser**, ikke som et vanlig verktøy. Min linse: bygg kontroll som **operasjonelle sluser**, ikke som “forbud med håp om etterlevelse”. I et heterogent landskap er den største risikoen ikke én feil, men **syklisk feilforsterkning**: agent → API-kall → sideeffekt → ny observasjon → nytt kall. Dette må brytes med design.

### Arkitekturprinsipper

- Bruk **deny-by-default** og eksplisitte “capability envelopes”: agenten får bare gjøre 1) lesing, 2) forslag, 3) begrensede handlinger i forhåndsdefinerte domener.
- Sett en **policy-gate per handling**, ikke per sesjon. Uten sentral OPA: start med sidecar-/proxy-baserte sjekker og server-side enforcement der det er mulig.
- Del agentflyt i tre soner: **observe → decide → execute**. Kun execute-sonen kan virke på produksjon, og den skal være smal, tidsbegrenset og reverserbar.

### Execution graph

- Modellér hver agentkjøring som en **DAG med immutable event IDs**: prompt, verktøyvalg, policybeslutning, kall, respons, sideeffekt.
- Hvert steg får en **idempotency key** og en “blast radius”-tag. Alt over terskel sendes til menneskelig godkjenning.
- Legg inn **loop-deteksjon**: maks N=5 verktøykall per sak, maks 2 minutter uten progress.

### IAM

- Bruk **ephemeral credentials** via short-lived service accounts, scoped per oppgave og per system.



- Token skal være **formål-, tids- og databegrenset**; roter hvert 15.–30. minutt.
- Delegert autorisasjon må være “just-in-time”, ikke permanente agentidentiteter.

### API-risiko

- Legacy-API-er uten idempotens er høy sårbarhet: anta dupliserte kall, delvise writes og skjulte sideeffekter.
- Beskytt med **rate limits, circuit breakers, dead-man switch** og “dry-run first”.
- Krev compensating actions for alle irreversible operasjoner.

### Audit og hendelser

- Logg prompts, tool calls, policyavgjørelser og utfall i en **konsolidert audit pipeline**; uten dette får dere makt uten revisjon.
- Ved avvik: **auto-isoler agenten**, revoke tokens, og flytt saken til manuell triage.

### Blindsoner

- Worst case er ikke feil svar, men **maskinert eskalering** i integritetskritiske systemer. Modeller det som: feilhandling × re-kjøring × spredning.

### Neste steg

1. Definér 10–15 tillatte agenthandlinger og bygg en minimal policy-gate for dem.
2. Etabler audit-format og loop-/blast-radius-regler før pilot.
3. Test én avgrenset use case i sandbox med ephemeral credentials og rollback-plan.