



id: SI_009

model: openai/gpt-5.4-mini

seeds:

04_risiko: usikkerhet

03_relasjon: makt

05_handling: vent

01_perspektiv: strategisk

02_tidhorisont: kortsiktig

Instans SI_009

Seeds: strategisk | kortsiktig | makt | usikkerhet | vent

Den strategiske feilen her er å tenke “agentisk KI” som et vanlig automasjonslag. I et heterogent landskap med legacy, svake kontrakter og begrenset kontrollmåling bør dere anta **makt-ubalanse**: agenten vil raskt få større handlingsrom enn dere faktisk kan observere eller reversere. Derfor: bygg for **venting** før handling, ikke for maksimal autonomi.

Arkitekturprinsipper

- Innfør en **companion-arkitektur**: agenten foreslår, en policy-/verifikasjonsmotor godkjenner, og en execution-runner utfører. Ikke la modellen kalle systemer direkte.
- Del opp autonomi i nivåer:
 - N0: lesing/observasjon
 - N1: skrive til sandkasse
 - N2: endringer med reversering
 - N3: irreversible handlinger krever menneskelig godkjenning
- Modellér hvert agentløp som en **execution graph** med noder: intent → plan → policy check → tool call → sideeffect → audit. Alle noder får korrelasjons-ID.

Zero Trust + policy-as-code

- Bruk korte, tidsbundne tokens per oppgave; aldri “stående” service-accounts med bred rettighet.
- Minimum: OPA/Gatekeeper-lignende policy ved runtime, selv om det først er som sidecar/proxy.
- Least privilege må være **kontekstuell**: system, datasett, tidspunkt, formål, risikonivå.



- Legg agenten i isolert runtime/container med egress-kontroll; ingen fri nettverksadgang.

IAM

- Ephemeral credentials med TTL på minutter, ikke timer/dager.
- Token-scoping per API og per handlingstype; separer les/skriv/exec.
- Delegert autorisasjon må være begrenset til eksplisitt “blessede” workflows.

API- og integrasjonsrisiko

- Anta ikke-idempotens, skjulte sideeffekter og dårlig kontraktshåndheving. Sett default til **read-only** mot legacy.
- Rate-limit hardt, bruk circuit breakers, og sett per-agent kvoter for kall, feil og samtidighet.

Audit og hendelser

- Logg prompts, beslutningsgrunnlag, policyresultat, tool-calls, responser og sideeffekter i en konsolidert audit-strøm.
- Ved loop, eskalering eller uventet mønster: auto-isoler agenten, revokér tokens, stopp egress, og flagg for manuell triage.

Blindsoner

- Svak datakvalitet gjør agenten “maktfull” på feil premisser.
- Latent coupling og emergent behavior undervurderes ofte; worst case er ikke feil svar, men **feil handling i feil rekkefølge** med irreversible konsekvenser.

Neste steg

1. Definer autonomnivåer og en “no-go”-liste for irreversible handlinger.
2. Pilotér én avgrenset sandbox med OPA-lignende policy og ephemeral credentials.
3. Etabler audit-krav og kill-switch før første produksjonstest.