



id: SI_006

model: openai/gpt-5.4-mini

seeds:

04_risiko: blindsone

03_relasjon: makt

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: syklisk

Instans SI_006

Seeds: strategisk | syklisk | makt | blindsone | implementer

Agentisk KI må behandles som en **ny maktaktør i arkitekturen**, ikke som en vanlig integrasjon. Strategisk betyr dette: begrensn først **hva agenten kan påvirke**, ikke bare hva den kan lese. I en offentlig virksomhet bør første pilot være **“read-mostly + lav-risiko write”** med tydelig grense mot irreversible handlinger (utbetaling, journalendring, sletting, tilgangsvedtak, eksterne utsendelser).

Arkitekturprinsipper

1. **Minst privilegium per handling, ikke per agent:** del opp agentens kapasitet i små “abilities” med egne tjenestekontoer, scopes og tidsbegrensede tokens (f.eks. 15–30 min TTL).
2. **To-lags kontroll:** agenten foreslår og forbereder; en policy-motor godkjenner; systemadapter utfører. For høyrisiko-aksjoner kreves menneskelig godkjenning.
3. **Handlinger må være idempotente og reversible** der det er mulig. Hvis ikke: bruk “hold-to-execute” med eksplisitt eskalering.
4. **Kontraktsbasert integrasjon** mot eldre systemer: adaptere som oversetter agentens intensjon til smale, verifiserte API-kall. Ikke gi agent direkte tilgang til legacy-grensesnitt.

Nødvendige sikkerhetsmekanismer

- **RBAC + service accounts per domene;** unngå delt superbruker.
- **Policy-sandbox** med allowlist for systemer, datatyper, volum, tid på døgnet og kommandoer.
- **Rate limiting og kvoter** per agent/oppgave for å stoppe loops og kost/DoS.
- **Sekretisasjon:** ingen rå nøkler i prompt; bruk vault + kortlevde delegasjonstokens.



- **Output-filtrering/DLP** for eksterne API-kall og e-post/tekstgenerering.

Blindsone du må regne med

- Legacy-systemer har ofte **sideeffekter og skjult status**; agenten kan “lykkes” logisk men bryte forretningsregler.
- Logging blir ofte **syklisk fragmentert**: hver komponent logger seg selv, men ingen kan rekonstruere kjeden. Bygg en sentral hendelsesstrøm med korrelasjons-ID fra start.
- Autonome agenter skaper **maktforskyvning**: den som kontrollerer policy-laget kontrollerer reell virksomhetsmakt. Forankre eierskap i både sikkerhet og fag.

Kontrollmodell

- Lav risiko: full autonomi.
- Medium: forslag + policy-godkjenning.
- Høy: menneskelig godkjenning + tidsbegrenset tilgang.
- Kritisk: aldri autonomt.

Neste steg

1. Kartlegg 10 konkrete agent-aksjoner og klassifiser dem i lav/medium/høy/kritisk risiko.
2. Etabler en pilot med én policy-motor, én sentral audit-logg og to isolerte service accounts.
3. Definer “stopp-regler” og rollback for de tre mest irreversible handlingene før pilotstart.