



id: SI_007

model: openai/gpt-5.4-mini

seeds:

04_risiko: mulighet

03_relasjon: makt

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: kortsiktig

Instans SI_007

Seeds: strategisk | kortsiktig | makt | mulighet | implementer

Agentisk KI må behandles som **makt med begrenset mandat**, ikke som “smart automatisering”. Den strategiske feilen er å optimalisere for kapabilitet først; i en offentlig virksomhet er riktig rekkefølge **kontroll → sporbarhet → gradvis autonomi**. Med 6–9 måneder og svak sikkerhetskapasitet bør dere velge en **kortsiktig, defensiv arkitektur** som maksimerer mulighet uten å åpne for irreversible handlinger.

Arkitekturprinsipper

1. Least privilege per handling, ikke per agent

Agenten får ikke “systemtilgang”; den får tilgang til en forhåndsdefinert handlingskatalog: lese, foreslå, klargjøre, utføre med godkjenning, utføre automatisk (kun for lavrisiko).

Et godt startmål: **80% av handlinger = foreslå/klargjøre, 15% = menneskelig godkjenning, 5% = automatisk.**

2. Policy-sandkasse som transaksjonsfilter

Legg en kontrolltjeneste foran alle API-kall: validerer formål, datakilde, mottaker, beløpsgrenser, tidsrom, og om handlingen er reversibel.

For eldre fagsystemer uten integrasjon: bruk **proxy-/wrapper-mønster** eller RPA-lag med hardt begrensede operasjoner.

3. To-delt autorisasjon: identitet + kontekst

IAM/RBAC alene er utilstrekkelig. Bruk kortlevde service tokens (f.eks. 15 min), separat identitet per agent/oppgave, og nøkkelrotasjon minst hver **24. time** for pilot, helst per sesjon for høyrisiko-kall.



4. **Observability som sikkerhetskontroll**

All agentaktivitet må logges som hendelsesrekker: input, plan, policy-sjekk, kall, respons, effekt. Fragmentert logging er en blindsoner. 1.

Krav: sentral loggsone, korrelasjons-ID per agent, og immutable audit trail.

Typiske blindsoner

- Agenten kan **eskalere via sekundære API-er** som ikke er direkte eksponert.
- Den kan lekke sensitiv info gjennom eksterne tjenestekall, prompt-injeksjon eller “harmless” metadata.
- Legacy-systemer har ofte **ukjente sideeffekter**; et enkelt kall kan trigge irreversible batchprosesser.
- Feilhandlinger blir ofte undervurdert: rollback finnes ikke, bare manuell reparasjon.

Kontrollmodell

Bruk et **tretrinns løp**:

1. **Planlegging i sandkasse**
2. **Policy-godkjenning**
3. **Utførelse via styrt gateway**

Legg inn rate-limit, tidsvinduer, og “kill switch” for alle agentklasser.

Neste steg

1. Definer en **handlingsmatrise** med 20–30 konkrete operasjoner og klassifiser dem som lav/middels/høy risiko.
2. Etabler en **policy-gateway** foran 2–3 pilotintegrasjoner, inkludert logging og kortlevde credentials.
3. Velg ett **legacy-system** og test rollback/feilhåndtering før bred pilot.