



id: SI\_005

model: openai/gpt-5.4-mini

seeds:

04\_risiko: mulighet

03\_relasjon: konkurranse

05\_handling: vent

01\_perspektiv: strategisk

02\_tidhorisont: langsiktig

## Instans SI\_005

---

**Seeds:** strategisk | langsiktig | konkurranse | mulighet | vent

For en offentlig virksomhet er den største feilen å behandle agentisk KI som “enda en integrasjon”. Strategisk bør dere se dette som en **ny kontrollflate**: agenten blir en aktør med delvis operasjonell myndighet. Langsiktig er konkurransefortrinnet ikke å være mest autonom, men å være **mest kontrollerbar**. Her er det en tydelig mulighet: bygg en styringsmodell som andre etater kan arve.

---

### Prinsipp 1: Minste nødvendige autonomi

Gi agenten **ikke tilgang til systemer**, men til **avgrensede handlinger** via en mellomliggende policy-/orchestrator-tjeneste. Ikke la modellen kalle fagsystemer direkte. Bruk:

- **Allowlist av handlinger** per use case
- **Tidsbegrensede tokens** per oppgave
- **Maks 1-3 steg autonomi** før menneskelig kontroll for risikable prosesser
- **Separate identiteter** per agent, per formål, per miljø

---

### Prinsipp 2: Policy-sandkasse foran legacy

Siden IAM kun har RBAC, må dere kompensere med en **beslutningsgateway**:

- Policy engine (f.eks. OPA-lignende mønster) før hvert kall
- **Transaksjonsklasser**: les, foreslå, klargjør, utfør
- “Utfør” kun for lavrisiko og reverserbare handlinger



- For eldre systemer: bruk **proxy/API-adapttere** som oversetter til kontrollerte operasjoner

---

## Prinsipp 3: Sporbarhet som designkrav

Fragmentert logging er en blindsonde. Krev:

- **Korrelasjons-ID** fra prompt til systemkall
- Immutable audit log med: input, policybeslutning, kall, resultat, rollback
- Separat logg for **modellens resonnering er ikke nødvendig**; logg heller beslutningsgrunnlag og policyutfall

---

## Prinsipp 4: Feil er forventet

Den største undervurderte risikoen er ikke “hacking”, men **feilhandlinger med legitim tilgang**.

- Innfør rate limiting per agent og per datakilde
- “Two-man rule” for irreversible handlinger
- Pre-approved rollback for kritiske prosesser
- Automatisk stopp ved avvik: uvanlig volum, nye destinasjoner, sensitiv data i utgående kall

---

## Blindsoner dere må anta finnes

- Agenten kan **eskalere indirekte** ved å kombinere flere små lovlige handlinger
- Legacy-systemer mangler ofte granularitet: RBAC blir for grovt
- Eksterne API-er kan lekke sensitiv informasjon via prompt, parameter eller feilsvar

---

## Kontrollmodell

Vent med full autonomi. Start med **shadow mode** i 6–8 uker: agenten foreslår, menneske utfører. Deretter begrenset autonomi i **ikke-kritiske, reverserbare** prosesser.

### Neste steg

1. Velg 2 use cases med lav risiko og høy nytte, og definer eksplisitte handlingsgrenser.
2. Etabler en policy-gateway med audit log før all systemtilgang.



3. Lag en risikomatrise med tre klasser: reverserbar, delvis reverserbar, irreversibel.