



id: SI\_009

model: openai/gpt-5.4-mini

seeds:

04\_risiko: sårbarhet

03\_relasjon: samarbeid

05\_handling: utforsk

01\_perspektiv: strategisk

02\_tidhorisont: historisk

## Instans SI\_009

**Seeds:** strategisk | historisk | samarbeid | sårbarhet | utforsk

Agentisk KI bør ikke behandles som “enda en integrasjon”, men som en ny aktørklasse med **begrenset fullmakt**. Strategisk betyr det at dere må designe for **minste privilegium + minste autonomi** fra start: agenten skal kunne foreslå, klargjøre og forberede handlinger, men bare utføre et smalt sett av forhåndsgodkjente operasjoner. Historisk ser vi at virksomheter som først bygger “full frihet og så kontroll” ofte ender med retrofitting av sikkerhet, og det blir dyrt og ufullstendig.

## Anbefalt kontrollmodell

### 1. Policy-sandkasse per handlingsdomene

- Del agentene i soner: lesing, klargjøring, simulering, godkjent utførelse.
- Tillat kun handlinger via en **policy-gateway** som verifiserer: system, objekt, operasjon, tidspunkt, kontekst.
- Bruk “deny by default” og tidsbegrensede tokens.

### 2. IAM med delegert identitet

- Ikke gi agenten menneskeidentitet. Bruk **egen maskinidentitet per agent og per oppdrag**.
- Knytt token-levetid til handlingstype: f.eks. 5–15 min for skriveoperasjoner, 1–24 t for lesing.
- Krev step-up for irreversible handlinger.

### 3. Transaksjonell sikkerhet og rollback

- For alt som kan endre data: innfør **2-fase mønster**: foreslå → valider → utfør.



- Der rollback ikke er mulig, må handlingen være “quarantine-first” eller blokkert.
- Lag en “kill switch” som stopper alle agentkall innen 60 sekunder.

---

## Blindsoner dere lett undervurderer

- **Eldre fagsystemer:** De mangler ofte granular audit og tåler ikke idempotente retry-mønstre. Agenten kan skape dobbeltbokføring eller skjulte sideeffekter.
- **Dataeksfiltrasjon via API:** Agenten kan lekkesensitivitet indirekte gjennom parametere, feilmeldinger eller oppsummeringer.
- **Autonom eskalering:** En agent kan “løse oppgaven” ved å be om mer tilgang i stedet for å be om hjelp.
- **Observabilitetsgap:** Fragmentert logging gjør årsakskjede nesten umulig å rekonstruere.

---

## Sikkerhetsmekanismer som bør være obligatoriske

- Policy-motor før alle skrivekall
- Rate limiting per agent og per system
- Output-filtrering for sensitiv info
- Sentral audit-logg med korrelasjons-ID
- Simulering/testmodus mot “mock”-endepunkter før produksjon
- Hendelsesregler for anomalier: volum, feilrate, uvanlige mål

---

## Samarbeid og sårbarhet

Samarbeid mellom sikkerhet, arkitektur, drift og fagdomene må være operativt, ikke bare styringsmessig. Agentisk KI eksponerer sårbarheten i dagens arkitektur: det er et spill som viser hvor lite deterministisk virksomheten egentlig er.

---

## Neste steg

1. Definer 3 agent-arbeidsflyter og klassifiser dem som lesing, endringsforslag eller begrenset utførelse.
2. Etabler en policy-gateway og sentral audit-logg for én pilotintegrasjon.
3. Gjennomfør en “red team”-øvelse mot dataeksfiltrasjon og irreversible handlinger før pilotstart.