



id: SI_004

model: openai/gpt-5.4-mini

seeds:

04_risiko: mulighet

03_relasjon: konkurranse

05_handling: utforsk

01_perspektiv: operasjonell

02_tidhorisont: kortsiktig

Instans SI_004

Seeds: operasjonell | kortsiktig | konkurranse | mulighet | utforsk

Agentisk KI bør behandles som **en ny type produksjonsaktør**, ikke som et vanlig integrasjonsverktøy. Den viktigste operasjonelle regelen er: **ingen agent får “frie hender” i kjerne- og fagsystemer**. I et kortsiktig pilotvindu (6–9 mnd) er målet ikke full autonomi, men **kontrollert handlingsrom med eksplisitte grenser**.

Anbefalt kontrollmodell

1. Policy-first, ikke model-first

Legg beslutningsgrensene utenfor modellen i en ekstern policy-motor eller orkestreringstjeneste. Agenten foreslår, men en kontrolltjeneste avgjør hva som kan utføres.

2. Minst mulig rettigheter per oppgave

Bruk separate servicekontoer per use case, tidsbegrensede tokens, og RBAC-strengt. Når finmasket policy mangler, må du kompensere med:

- korte token-liv
- scoped API-nøkler
- dedikerte “write-limiter”-kontoer
- sperrer mot produksjonsdata der det er mulig

3. Sandkasse med irreversible-handlinger forbudt

Definer 3 soner:

- **Lese-sone:** kun hent data
- **Foreslå-sone:** generer forslag, ingen endringer



- **Utføre-sone:** kun godkjente, reversible operasjoner
Alt som er irreversibelt (sletting, masseoppdatering, betalingsflyt, tilgangsendring) krever menneskelig godkjenning.

Blindsoner som ofte undervurderes

- Eldre fagsystemer kan mangle transaksjonsstøtte → rollback blir ofte “best effort”, ikke faktisk tilbakeføring.
- Agenten kan lekke sensitiv info via eksterne API-er, selv når den “bare oppsummerer”.
- Fragmentert logging gjør at hendelser ser trygge ut lokalt, men uforståelige på tvers.
- Autonome prosesser skaper nye angrepsflater: prompt-injeksjon, overprivilegerte servicekontoer, og kjedefeil mellom systemer.

Sikkerhetsmekanismer som bør prioriteres

- **Policy engine** foran alle handlinger
- **Rate limiting** og kvoter per agent og per system
- **Isolasjon** via dedikerte arbeidsrom/miljøer
- **Audit trail** med korrelasjons-ID på tvers av systemer
- **Secrets management + nøkkelrotasjon** automatisk, helst daglig/ukentlig for pilot
- **Hendelsesmodell:** hver agenthandling må ha input, begrunnelse, policybeslutning, effekt og ansvarlig eier

Konkurrans- og mulighetsperspektiv

De som lykkes først, får ikke “smartest agent”, men **mest kontrollerbar agent**. Det er en operasjonell fordel: lavere risiko, raskere godkjenning, og høyere skalerbarhet.

Neste steg

1. Velg **2 brukstilfeller** med lav risiko og reversible handlinger.
2. Etabler en **policy-gateway** og et minimum av audit logging før pilot.
3. Lag en **tabell over tillatte/forbudte handlinger** per system før første modell kobles på.