



id: SI_008

model: mistralai/mistral-large-2512

seeds:

04_risiko: mulighet

03_relasjon: tillit

05_handling: aksepter

01_perspektiv: strategisk

02_tidhorisont: langsiktig

Instans SI_008

Agentisk KI som langsiktig tillitskontrakt - ikke teknisk quick-fix

Agentisk KI er ikke et teknisk prosjekt, men en **strategisk tillitsavtale** med fremtidens infrastruktur. De fleste feiler ved å fokusere på hva agentene kan gjøre - vi må starte med hva vi aksepterer at de aldri gjør. Dette krever en **mulighetsramme** som er like mye juridisk og organisatorisk som teknisk.

Tre arkitekturprinsipper for langsiktig kontroll

1. Policy-sandkasser som dynamiske tillitsgrenser

- Bygg "tillitsdomener" der agentene kun kan operere innenfor forhåndsdefinerte handlingskorridorer (eks: "kun lese i system A, kun skrive til system B med maks 3 transaksjoner/time").
- Bruk **Open Policy Agent (OPA)** til å definere regler som automatisk strammes når agentene nærmer seg grenser (eks: "hvis agenten har utført 80% av sin ukentlige kvote, send varsel til menneskelig godkjenner").
- Eksempel: En agent som håndterer refusjoner kan kun godkjenne beløp < 10 000 NOK uten menneskelig sign-off.

2. Autonomi innenfor aksepterte feilmarginer

- Design **escalation-as-a-service**: Agentene må ha klare regler for når de skal avbryte og eskalere (eks: "hvis svaret fra ekstern API avviker >20% fra forventet, stopp og be om menneskelig vurdering").
- Implementer **reversible handler** som standard: Alle endringer må kunne rulles tilbake innen 24 timer (eks: lagre tidligere tilstand i en "time-machine"-database).



- Blindsoner: Eldre systemer uten transaksjonsstøtte vil kreve manuelle rollback-prosedyrer – dokumenter disse før piloten.

3. Logging som tillitsbyggende mekanisme

- Alle agenthandlinger må logges i et **uforanderlig, sentralisert system** (eks: **OpenSearch** med WORM-lagring) med følgende metadata:
 - Hvem (agent-ID + menneskelig eier)
 - Hva (handling + dataendringer i JSON Patch-format)
 - Hvorfor (begrunnelse fra agentens resonnering, f.eks. “valgte X fordi Y var utenfor policy”)
 - Konsekvens (forventet vs. faktisk utfall)
- Eksempel: Hvis en agent endrer en pasientjournal, må loggen vise både den tekniske handlingen og den kliniske begrunnelsen (f.eks. “oppdaterte medisiner basert på labresultat ID-1234”).

Tre undervurderte risikoer

1. **Integrasjonsgjeld:** Agentene vil avsløre hvor dårlig systemene dine faktisk snakker sammen. Forvent at 30% av pilotens tid går med til å fikse eksisterende API-mangler.
2. **Policy-drift:** Regler som virker trygge i dag (eks: “maks 10 API-kall/minutt”) kan bli farlige når agentene optimaliserer seg selv. Planlegg for **kvartalsvise policy-revisjoner**.
3. **Tillitserosjon:** Hvis agentene gjør én feil som ikke kan spores eller reverseres, vil organisasjonen miste troen på hele konseptet. Sikkerhet er ikke teknisk, men psykologisk.

Konkrete neste steg

1. **Kartlegg “tillitsgrenser” i en workshop** med både teknisk og juridisk/faglig deltakelse. Spørsmål: Hvilke handlinger er så kritiske at de aldri skal kunne automatiseres? (Eks: “slette pasientdata”, “godkjenne >500 000 NOK”).
2. **Bygg en “policy-prototyp”** med OPA som tester regler mot reelle historiske hendelser (eks: “Hva om agenten hadde håndtert saken fra 2023-11-05?”). Bruk dette til å kalibrere tillitsnivået.
3. **Design en “tillitsdashboard”** som viser sanntidsstatus for alle agenter (eks: “Agent-7 har utført 42/50 tillatte handlinger i dag, neste eskalering om 3 timer”). Gjør dette synlig for ledelsen – tillit bygges gjennom transparens.



Husk: Agentisk KI handler ikke om å gi slipp på kontroll, men om å flytte kontrollen fra ad-hoc menneskelige beslutninger til forutsigbare, etterprøvbare systemer. Det er en langsiktig investering i tillit – ikke en teknisk snarvei.