



id: SI_004

model: mistralai/mistral-large-2512

seeds:

04_risiko: mulighet

03_relasjon: konkurranse

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: langsiktig

Instans SI_004

Seeds: strategisk (dominerende), langsiktig (rammeverk), konkurranse (offentlig vs privat), mulighet (førstemann), implementer (pragmatisk)

Strategisk rammeverk for agentisk KI: Fra kontroll til konkurransefortrinn

Offentlige aktører som ikke mestrer agentisk KI innen 24 måneder, vil tape konkurransekraft mot private aktører (eks: banker, helseplattformer) som utnytter autonomi for å kutte kostnader og øke hastighet. Deres pilot er en mulighet til å bygge en **langsiktig kontrollmodell** som kan skaleres til andre offentlige virksomheter – men kun hvis de unngår tre vanlige fallgruver:

- Falsk trygghet i RBAC** RBAC er ubrukelig for autonome agenter fordi det ikke håndterer kontekst. Eksempel: En agent med “lesetilgang” til pasientdata kan likevel ekstrapolere sensitive mønstre gjennom API-kall til eksterne tjenester (f.eks. Google Maps for å korrelere adresse + diagnose). **Løsning:** Implementer en policy-motor (f.eks. Open Policy Agent) som evaluerer handlinger dynamisk basert på:
 - **Tidsvindu** (kun tillatt mellom 08-16)
 - **Datakontekst** (ikke tillatt å kombinere API A + API B)
 - **Historikk** (agenten har allerede gjort 3 lignende kall i dag)
- Fragmentert logging = blindsoner** Eldre systemer logger ikke intensjon, bare handlinger. En agent som endrer en faktura kan logges som “oppdatering”, uten



å avsløre hvorfor (f.eks. påvirket av en ekstern vær-API). **Løsning:** Bygg en hendelsesstrøm (f.eks. Kafka) som kobler:

- Agentens input (prompt, data)
- API-kall (med headers, payload)
- Systemrespons (feilmeldinger, endringer)
- Menneskelig godkjenning (hvis eskalering skjer) **Eksempel:** Helseplattformen Epic bruker en lignende modell for å spore beslutninger tatt av deres "AI Clinician".

3. **Sandboxing er ikke nok - trenger "reversible handlinger"** En agent som sletter en fil i et eldre system kan være umulig å rulle tilbake. **Løsning:** Designe atomære operasjoner med:

- **Tidsbegrensede transaksjoner** (f.eks. endringer låses i 48 timer før permanent lagring)
- **Skrivebeskyttede kopier** av kritiske data (f.eks. versjonert i Git-lignende struktur)
- **Rate-limiting per agent** (maks 10 handlinger/time for nye agenter)

Tre tekniske blindsoner som undervurderes

1. **Eskalering av feilhandlinger** En agent som misforstår en regel kan eskalere en liten feil til en systemkrise (eks: en "refusjonsagent" som godkjenner 1000 falske krav fordi den tolker et API-svar feil). **Løsning:** Bygg en feilmodus-database som trener agenter på historiske hendelser (f.eks. "Hva skjedde sist en agent endret X?").
 2. **Eksterne API-er som trojanere** Agenter som bruker eksterne API-er (f.eks. Google Translate for dokumenter) kan utilsiktet sende sensitive data. **Løsning:** Implementer en proxy (f.eks. Kong) som:
 - Stripper metadata fra forespørsler
 - Logger alle utgående kall
 - Blokkerer API-er som ikke er på en godkjent liste
 3. **Kulturell motstand mot autonomi** Ansatte vil sabotere agenter hvis de føler seg overkjørt. **Løsning:** Kjør en parallell pilot hvor agenter kun foreslår handlinger (men krever menneskelig godkjenning), og mål effekten på produktivitet.
-



Neste steg (prioritert)

1. Bygg en “Policy Sandbox” på 4 uker

- Velg én kritisk prosess (f.eks. fakturahåndtering) og definer 5 regler for agenten (f.eks. “maks 5 endringer/time”, “kun tillatt å bruke API X + Y sammen”).
- Bruk Open Policy Agent til å validere handlinger i sanntid.
- **Mål:** Reduser feilhandlinger med 80% før piloten.

2. Implementer “Reversible Actions”-arkitektur

- Lag en midlertidig lagring (f.eks. PostgreSQL med tidsstempler) for alle endringer gjort av agenter.
- Bygg en rollback-API som kan gjenopprette data innen 24 timer.
- **Test:** Simuler 10 feilhandlinger og mål gjennomsnittlig gjenopprettingstid.

3. Kjør en “Red Team”-øvelse

- Engasjer en ekstern aktør (f.eks. mnemonic) til å prøve å lure agentene til å:
 - Eksportere data via eksterne API-er
 - Utføre handlinger utenfor policy
 - Eskalere feil til systemkollaps
- **Mål:** Identifisere minst 3 kritiske sårbarheter før piloten.

Strategisk utfall: Ved å løse disse problemene før piloten, posisjonerer dere dere som en førstebeveger i offentlig sektor – med en kontrollmodell som kan lisensieres til andre virksomheter. Konkurrentene deres vil bruke 12–18 måneder på å kopiere dette.