



id: SI\_008

model: anthropic/claude-opus-4.6

seeds:

04\_risiko: mulighet

03\_relasjon: makt

05\_handling: vent

01\_perspektiv: strategisk

02\_tidhorisont: kortsiktig

## Instans SI\_008

**Seeds:** strategisk | kortsiktig | makt | mulighet | vent

### Kjerneanalyse: Makt uten kontroll er ikke en mulighet — det er en risiko

Her er mitt dissens-standpunkt: **Ikke rull ut agentisk KI om 6-9 måneder. Vent.**

Ikke fordi teknologien er umoden, men fordi maktstrukturen i arkitekturen deres ikke er klar til å absorbere den. Dere har RBAC uten finmasket policy-styring, fragmentert logging, og en sikkerhetsarkitektur bygget for deterministiske flyter. Å slippe autonome agenter inn i dette er som å gi nøkkelkortet til hele bygget til en svært kompetent, men uforutsigbar vikar — uten vaktlogg.

### Hva “vent” faktisk betyr strategisk

Vent  $\neq$  passivitet. Vent = **bruk de 6-9 månedene på å bygge kontrollplanet først, ikke agentene.**

#### Konkret prioritering:

- Policy-motor før agent-motor.** Implementer en dedikert policy-beslutningsmotor (f.eks. OPA/Styra eller Cedar fra AWS) som et separat lag mellom agenter og fagsystemer. Estimat: 3-4 måneder. Dette gir finmasket, deklarativ kontroll — ikke den binære RBAC-en dere har i dag.
- Handlingsrom som kontrakt, ikke konfigurasjon.** Definer agentenes action envelope som maskinlesbare kontrakter (JSON Schema/OpenAPI-basert). Hvert API-kall en agent kan gjøre skal ha eksplisitt allow-list, rate-limit (maks 50



skriveoperasjoner/time per agent), og irreversibilitetsflagg som trigger human-in-the-loop.

3. **Unified audit log nå.** Fragmentert logging er den farligste blindsonen. Uten korrelert sporbarhet på tvers av systemer kan dere ikke gjøre forensics når (ikke hvis) en agent feiler. Sentraliser med OpenTelemetry-basert pipeline til én SIEM-instans. Estimat: 2–3 måneder parallelt.

---

## Blindsoner dere undervurderer

- **Eldre fagsystemer uten idempotente API-er:** En agent som retry-er et kall mot et legacy-system kan duplisere transaksjoner. Rollback er ofte umulig. Denne risikoen er akutt.
- **Maktkonsentrasjon i servicekontoer:** Agenter vil arve tilgangene til sine servicekontoer. Uten just-in-time privilege escalation (à la CyberArk/HashiCorp Vault) får dere *standing privileges* som er attraktive angrepsflater.
- **Eksfiltrering via prompt/kontekst:** Agenter som kaller eksterne API-er kan lekke sensitive data i request-body uten at DLP-løsninger fanger det, fordi trafikken ser “legitim” ut.

---

## Muligheten som faktisk finnes

Den strategisk riktige kortsiktige gevinsten er å pilotere **read-only agenter** med observatør-rolle — agenter som analyserer, foreslår og logger, men aldri skriver. Dette gir dere reell erfaring med agent-atferd i produksjonsmiljøet uten irreversibel risiko.

---

## Neste steg

1. **Uke 1-2:** Kartlegg alle skriveoperasjoner agentene ville trenge, klassifiser som reversibel/irreversibel, og definer action envelopes.
2. **Måned 1-4:** Implementer policy-motor (OPA) og sentralisert audit-pipeline parallelt, med read-only agent-pilot som testcase.
3. **Måned 5-6:** Evaluer om kontrollplanet er modent nok til å åpne for skrive-agenter i én avgrenset sandkasse — med kill-switch og 100% human-in-the-loop på irreversible handlinger.

**Makten til å handle autonomt er en mulighet. Men den krever at kontrollen finnes før autonomien slippes løs.**