



id: SI\_008

model: anthropic/claude-sonnet-4.6

seeds:

04\_risiko: sårbarhet

03\_relasjon: makt

05\_handling: implementer

01\_perspektiv: operasjonell

02\_tidhorisont: langsiktig

## Instans SI\_008

**Seeds:** operasjonell | langsiktig | makt | sårbarhet | implementer

### Maktforskyvningens blindsoner

Det mainstream-perspektivet overser: agentisk KI handler ikke primært om teknisk risiko — det handler om **maktforskyvning**. Når en agent kan initiere prosesser og endre data autonomt, forflyttes beslutningskraft fra identifiserbare mennesker til systemkonfigurasjoner. Dette er en langsiktig, irreversibel endring i virksomhetens maktstruktur som 6-9 måneder er altfor kort til å absorbere.

### Operasjonelle sårbarheter ingen snakker høyt om

**IAM-gapet er akutt, ikke kronisk.** RBAC uten finmasket policy-styring betyr at agenten arver bredeste mulige tilgangsprofil for en rolle. I praksis: én kompromittert agent-sesjon kan eskalere til full lese/skrive-tilgang på fagsystemer. Implementer umiddelbart **least-privilege per agent-handling**, ikke per agent-rolle.

**Fragmentert logging er ikke et rapporteringsproblem — det er et rollback-problem.** Uten atomisk sporbarhet på tvers av systemer kan du ikke rekonstruere hva agenten faktisk gjorde ved en feilhandling. Implementer en dedikert agent-audit-bus (f.eks. OpenTelemetry + immutable log-sink) før første pilot-handling, ikke etterpå.

**Eldre fagsystemer uten integrasjonsstøtte er langsiktige giftkapsler.** Disse systemene har ingen konsept om “ugyldig avsender” — de stoler på at kallet er legitimt. En agent som kaller dem ser ut som et internt system. Isoler dem bak et gateway-lag med eksplisitt whitelist-semantikk: ingen agent-kall passerer uten godkjent handlingsmal.



---

## Kontrollmodell uten flaskehalser

Implementer **tre-lags autorisasjon**:

1. Hva agenten kan forsøke — policy-motor (OPA/Cedar) med handlingsromsdefinisjon
2. Hva agenten faktisk gjør — sandkasse med dry-run-modus for alle irreversible operasjoner
3. Hva som godkjennes — async human-in-the-loop for handlinger over definert risikoscore

Rate-limiting på agent-nivå (ikke bruker-nivå): maks 50 API-kall/minutt per agent-instans, hard stopp ved avvik >2 standardavvik fra baseline.

---

## Neste steg

1. **Denne uken:** Kartlegg alle fagsystemer uten integrasjonsstøtte og frys agent-tilgang til disse inntil gateway-lag er på plass — dette er ikke valgfritt.
2. **Innen 30 dager:** Implementer agent-audit-bus med immutable logging og definer rollback-prosedyre for de 10 mest risikable handlingstypene.
3. **Innen 60 dager:** Gjennomfør red team-øvelse der agenten forsøker privilege escalation via legacy-systemer — bruk funnene til å kalibrere policy-motoren før pilot går live.