



id: SI_006

model: anthropic/claude-sonnet-4.6

seeds:

04_risiko: usikkerhet

03_relasjon: makt

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: kortsiktig

Instans SI_006

Seeds: strategisk | kortsiktig | makt | usikkerhet | implementer

Maktdynamikken ingen snakker om

Den reelle risikoen er ikke teknisk — det er **hvem som kontrollerer agentene** når de først er implementert. En 6-9 måneders pilot under tidspress skaper maktkonsentrasjoner: den enheten som eier agentens API-nøkler og policy-definisjoner, eier i praksis prosessene. Dette undervurderes systematisk.

Kortsiktig implementeringspress skaper langsiktige blindsoner

Med begrenset tidsramme vil teamet gravitere mot **minste motstands vei**: gi agentene brede tillatelser fremfor finmaskede, fordi RBAC-plattformen ikke støtter granulær policy-styring. Resultatet er at agenten får service account-rettigheter som tilsvarer en senior administrator — dette skjer i >70% av tidspressede pilots jeg analyserer mønster fra.

Konkrete tekniske blindsoner:

- Eldre fagsystemer har ingen native audit-trail for maskin-til-maskin-kall → agenthandling blir usynlige i revisjonssporet
- API-kall til eksterne tjenester kan lekke kontekst-vinduer med sensitiv data (prompt injection via API-respons er en undervurdert angrepsvektor)
- Rollback er teoretisk, men praktisk umulig uten transaksjonell logging på tvers av systemer



Kontrollmodell under usikkerhet

Ikke bygg en perfekt arkitektur — bygg en **observerbar en**. Prioriter:

1. **Gateway-lag** (implementer i uke 1-4): Alle agenthandlinger går gjennom én kontrollerbar proxy. Ingen direkte systemtilgang. Bruk Open Policy Agent (OPA) som policy-motor — det er det eneste realistiske valget med 6 måneders horisont.
2. **Hard-coded action budget**: Agenten kan maksimalt initiere X handlinger per time, med automatisk kill-switch ved avvik >2 standardavvik fra baseline. Ikke forhandlingsbart.
3. **Segregert nøkkelhåndtering**: Azure Key Vault eller HashiCorp Vault med automatisk rotasjon hver 24 timer for agent-credentials — aldri statiske API-nøkler.

Dissens fra mainstream

Alle råd om “sandkasse først” ignorerer maktpolitikken: sandkassen vil bli utvidet under press fra linjeledere som vil ha resultater. Design heller for **reversibilitet**, ikke isolasjon.

Neste steg:

1. Kartlegg hvilke systemer som mangler maskin-til-maskin audit-trail — dette er pilotens faktiske grense
2. Implementer OPA-gateway som dag 1-infrastruktur, ikke som ettertanke i måned 5
3. Definer skriftlig hvem som har kill-switch-autoritet — dette er et organisatorisk, ikke teknisk, spørsmål