



id: SI\_008

model: anthropic/claude-sonnet-4.6

seeds:

04\_risiko: sårbarhet

03\_relasjon: makt

05\_handling: forhandle

01\_perspektiv: strategisk

02\_tidhorisont: kortsiktig

## Instans SI\_008

**Seeds:** strategisk | kortsiktig | makt | sårbarhet | forhandle

### Maktforholdet ingen snakker om

Den egentlige risikoen er ikke teknisk — det er at virksomheten **avgir makt** uten å ha forhandlet betingelsene. Agentisk KI er ikke et verktøy; det er en ny aktør i systemet med egne “interesser” definert av tredjeparter (LLM-leverandør, API-eiere, integrasjonspartnere). Hvert autonomt steg er en maktoverføring.

### Kortsiktig pilot = strategisk felle

6–9 måneder tvinger frem **snarveier som låser arkitekturen**. Uten OPA, uten Zero Trust, uten konsolidert logging — betyr pilot i praksis: agenter med overdimensjonerte service accounts, statiske tokens og ingen real-time circuit breakers. Dette er ikke en pilot, det er produksjon med mindre ansvar.

Konkret sårbarhet: Legacy-monolitter uten idempotente API-er + autonom agent = ukontrollerbare sideeffekter. En agent som kaller samme endepunkt to ganger ved retry kan dobbelt-booke, dobbelt-fakturere eller korrumpere tilstand. **Ingen rollback redder dette uten transaksjonelle garantier i bunn.**

### Forhandle kontrollplanen — ikke bare teknologien

Virksomheten bør forhandle eksplisitt med:

- **LLM-leverandør:** Hva logges? Hvor? Hvem eier prompt-historikk?



- **API-eiere (interne og eksterne):** Hvem har ansvar når agenten trigger uønsket hendelse?
- **Ledelsen:** Hva er akseptabelt tapsscenario — ikke som abstraksjon, men i kroner og omdømme.

Disse forhandlingene er styringsarkitekturen. Tekniske kontroller uten avklart ansvar er dekorasjon.

---

## Emergent behavior er et maktproblem

“Latent coupling” undervurderes fordi det ikke er synlig i diagrammer. En agent som koordinerer mellom tre systemer kan utilsiktet skape en fjerde prosess ingen eier. Dette er ikke en feil — det er emergent adferd. Ingen audit-pipeline fanger det hvis logging er distribuert og ukorrelert.

**Worst-case:** Agent med skrivetilgang til fagsystem + ekstern API-integrasjon + retry-logikk = eksfiltrering av sensitiv data uten at noen varsler innen 72 timer.

---

## Neste steg

1. **Definer maktgrenser før teknologi:** Skriv et “Agent Charter” — hvilke systemer er off-limits, uansett teknisk mulighet. Gjør dette juridisk bindende internt.
2. **Forhandle ned scope:** Kutt pilot til ett system, ett datadomene, null eksterne API-er. Tving frem reelle transaksjonelle garantier der.
3. **Kjøp OPA/Styra nå** — ikke etter pilot. Policy-as-code er forutsetningen, ikke etterarbeidet.