



id: SI_006

model: anthropic/claude-sonnet-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: samarbeid

05_handling: implementer

01_perspektiv: økonomisk

02_tidhorisont: langsiktig

Instans SI_006

Seeds: økonomisk | langsiktig | samarbeid | sårbarhet | implementer

Perspektiv: Den langsiktige sårbarhetskostnaden undervurderes systematisk

Mainstream-rådet er å bygge teknisk kontroll lag for lag. Min analyse peker et annet sted: **den største risikoen er ikke teknisk - det er at dere implementerer for raskt uten å prise inn langsiktig sårbarhetskostnad.**

Økonomisk realisme om pilotvinduet

6–9 måneder er for kort til å implementere agentisk KI trygt i et heterogent landskap uten Zero Trust, uten sentralisert policy-motor, uten konsolidert logging. Estimat: et minimalt forsvarlig kontrollregime (OPA, ephemeral credentials, audit-pipeline) krever 3–4 månedsverk sikkerhetsarkitektur før agentene slippes løs. Det har dere ikke budsjettet. En enkelt feilhandling mot et fagsystem uten transaksjonelle garantier kan koste mer enn hele piloten er verdt.

Dissens: Ikke start med agenter som endrer data. Start med agenter som kun leser og rapporterer. Det er 80% av verdien til 20% av risikoen.

Sårbarhet i systemkoblingene, ikke agenten

Latent coupling er underestimert. Eldre monolitter har ofte udokumenterte sideeffekter ved API-kall – agenten trigger API X, som internt kaller Y, som skriver til Z uten logg. Ingen audit-pipeline fanger dette. Emergent behavior oppstår ikke i agenten – det oppstår i koblingene mellom systemer som aldri var designet for å snakke sammen autonomt.



Konkret tiltak: Kartlegg alle API-er agenten kan nå. Merk dem: idempotent/ikke-idempotent, reversibel/irreversibel. Blokkér alle irreversible kall i pilot-fase.

Samarbeid som kontrollmekanisme

IAM-modellen bør ikke bygges alene. Etabler et tverrfaglig “agent governance board” med juridisk, sikkerhet og systemansvarlige. Dette er ikke byråkrati – det er den billigste formen for risikoreduksjon. Felles eierskap av policy-as-code betyr at noen faktisk oppdaterer den når systemer endres.

Neste steg

1. **Gjennomfør API-risikoklassifisering** (irreversibel/idempotent) for alle systemer agenten kan nå – ferdig innen uke 3 av piloten.
2. **Begrens pilot til read-only agenter** med strukturert output til menneskelig beslutningsstøtte – implementer først, autonomi senere.
3. **Pris inn worst-case-kostnad** for én feilhandling per integrasjonspunkt og presenter dette som go/no-go-kriterium for ledelsen.