



id: SI\_009

model: anthropic/claude-sonnet-4.6

seeds:

04\_risiko: sårbarhet

03\_relasjon: konkurranse

05\_handling: utforsk

01\_perspektiv: strategisk

02\_tidhorisont: langsiktig

## Instans SI\_009

**Seeds:** strategisk | langsiktig | konkurranse | sårbarhet | utforsk

### Konkurransefortrinnet ligger i kontrollarkitekturen - ikke agentene selv

Det strategisk viktigste valget dere tar nå er ikke hvilken KI-agent dere velger, men hvordan dere bygger kontrollrammeverket rundt den. Dette er et langsiktig konkurransefortrinn: virksomheter som løser tillitslaget robust tidlig, vinner i neste fase der autonomi skalerer.

### Undervurderte sårbarheter

**Lateral movement via agentidentitet.** Når en agent autentiserer mot system A og B, oppstår en implisitt tillitskjede. Uten finmasket policy-styring (som deres RBAC ikke støtter) kan en kompromittert agent traversere systemer som en menneskelig bruker aldri ville ha tilgang til kombinert. Dette er ikke hypotetisk - det er dokumentert i Microsoft-studier av Copilot-misbruk (2024).

**Irreversible handlinger i fagsystemer.** Eldre systemer mangler typisk transaksjongrenser og rollback-API. En agent som skriver til disse uten idempotensgaranti kan skape datatilstand ingen kan rette uten manuell intervensjon.

**Ekstern API-lekkasje.** Agenter som kaller eksterne tjenester kan utilsiktet inkludere interne identifikatorer, personnummer eller metadata i request-payload. Uten utgående datafiltrering er dette en åpen kanal.



---

## Arkitekturprinsipp: “Minimal Footprint, Maximum Auditability”

Utforsk en **Agent Gateway-modell**:

- Alle agenthandlinger rutes gjennom ett sentralt lag med policy-motor (OPA/Cedar)
- Ingen direkte system-til-system-tilgang – kun via gateway
- Hver handling genererer et signert audit-event til uforanderlig logg (append-only, f.eks. WORM-storage)
- Handlingsrom defineres som eksplisitte whitelister, ikke blacklister

Dette løser IAM-gapet uten å bytte plattform, og gir dere etterprøvbarehet som tåler tilsynspress.

---

## Mot mainstreamen

De fleste piloterer agenter med brede tilganger “for å teste kapabilitet”. Strategisk feil. Test med smalest mulig handlingsrom og ekspander kontrollert. Seks måneder er nok til å bevise kontrollmodellen – ikke agentens fullstendige potensial.

---

## Neste steg

1. **Uke 1-2:** Kartlegg alle systemer agenten skal nå – klassifiser etter reversibilitet og sensitivitetsnivå
2. **Uke 3-4:** Implementer OPA eller Cedar som policy-motor foran én integrasjon som proof-of-concept
3. **Uke 5-6:** Definer minimum viable audit-log format og test replay-evne for én hendelseskjede