



id: SI\_004

model: anthropic/claude-sonnet-4.6

seeds:

04\_risiko: usikkerhet

03\_relasjon: tillit

05\_handling: utforsk

01\_perspektiv: økonomisk

02\_tidhorisont: historisk

## Instans SI\_004

**Seeds:** økonomisk | historisk | tillit | usikkerhet | utforsk

### Historisk mønster vi gjentar - og bør unngå

Offentlig sektor har gjentatte ganger innført teknologi som overgår eksisterende kontrollrammer – fra tidlig internetttilgang på 90-tallet til ukontrollert skytransisjon på 2010-tallet. Resultatet: reaktiv sikkerhet, kostbar opprydding, tapt tillit. Agentisk KI er samme mønster, men med høyere fart og irreversibilitet.

### Tillit som arkitekturprinsipp, ikke etterpåklokskap

Det kritiske spørsmålet er ikke hva agentene kan gjøre, men hvem og hva som stoler på hvem. Anbefalt tilnærming: **Zero-Trust for agenter** – ingen agent får implisitt tillit basert på identitet alene. Konkret:

- Hver agent tildeles et **korttidslivslangt token** (maks 15 min TTL) via en dedikert agent-identitetsprovider, atskilt fra menneskelige IAM-kontoer
- Handlingsrom defineres som eksplisitte **allow-lists per systemdomene**, ikke generelle RBAC-roller
- Eldre fagsystemer isoleres bak et **proxy-lag** (f.eks. Kong Gateway eller AWS API Gateway) som oversetter og logger alle agentforespørsler



---

## Undervurderte risikoer - den økonomiske blindsonen

Virksomheter kalkulerer pilotkosten (6-9 måneder), men ikke **katastrofekostnaden**: én agent som sletter eller korrupperer data i et fagsystem uten rollback kan koste 10-50x pilotkostnaden i gjenoppretting og omdømmetap. Spesielt undervurdert:

- **Kaskadeeffekter**: agent A kaller API som trigger agent B – feil propagerer eksponentielt
- **Data-eksfiltrering via prompt injection** i LLM-baserte agenter mot eksterne tjenester
- **Manglende revisjonsspor** i eldre systemer betyr at ansvar ikke kan plasseres – juridisk og regulatorisk risiko

---

## Kontrollmodell: Utforsk innenfor grenser

Design etter prinsippet **“sandbox-first, expand by evidence”**:

1. Start med **read-only agenter** i pilot – ingen skrivetilgang før revisjonslogging er fullstendig
2. Implementer **circuit breakers**: agent stoppes automatisk ved >3 uventede API-svar innen 60 sekunder
3. All ekstern kommunikasjon via **egress-proxy** med statisk allowlist – ingen dynamisk DNS-oppslag

---

## Konkrete neste steg

1. **Uke 1-2**: Kartlegg alle systemer agenten skal nå – klassifiser hvert som “kontrollert” (API + logging) vs. “ukontrollert” (eldre system). Ukontrollerte systemer er out-of-scope for pilot.
2. **Uke 3-4**: Etabler dedikert agent-IAM atskilt fra eksisterende bruker-IAM, med automatisk nøkkelrotasjon via HashiCorp Vault eller tilsvarende.
3. **Måned 2**: Gjennomfør **red team-øvelse** spesifikt rettet mot agentens handlingsrom – ikke generell penetrasjonstest.