



id: SI\_006

model: anthropic/claude-opus-4.6

seeds:

04\_risiko: sårbarhet

03\_relasjon: tillit

05\_handling: implementer

01\_perspektiv: strategisk

02\_tidhorisont: historisk

## Instans SI\_006

**Seeds:** strategisk | historisk | tillit | sårbarhet | implementer

### Tillit er ikke en konfigurasjon — det er et historisk regime du bygger

Historisk har offentlig sektor gjentatte ganger innført teknologi med autonomi før kontrollregimet var modent. NAVs Hades-system på 90-tallet, Storbritannias Horizon-skandalen (Post Office), og Boeing 737 MAX-autopiloten viser alle samme mønster: **systemer fikk handlekraft før tilliten var forankret i verifiserbare strukturer.** Dere står ved nøyaktig samme veiskille.

### Strategisk kjerneargument: Ikke automatiser tillit — konstruer mistillit

Mainstream-rådet er å bygge guardrails rundt agenter. Mitt perspektiv er skarpere: **behandle enhver KI-agent som en upålitelig intern aktør fra dag én.** Ikke fordi teknologien er ondsinnet, men fordi sårbarhetsflatene i deres landskap — legacy-monolitter uten deterministiske API-kontrakter, distribuert logging, fraværende policy-motor — gjør at tillitsantagelser vil kollapse under press.

### Konkrete arkitekturbeslutninger

**1. Implementer «Mistillitsbudsjettet»:** Hver agent får et kvantifisert handlingsrom — maks 15 skriveoperasjoner/time, maks 3 systemer per kjede, 0 eksterne API-kall uten



human-in-the-loop. Bruk OPA som runtime-policy-motor, men start med en hardkodet allow-list, ikke deny-list. Deny-lists er sårbare for det ukjente.

**2. Execution graph som førsteklases audit-objekt:** Hver agenthandling genererer en uforanderlig DAG (directed acyclic graph) med prompt, mellomsteg, API-kall, respons og sideeffekter. Lagre i append-only event store (EventStoreDB eller enkel Kafka-topic med compaction=off). Dette er ikke logging — det er **bevisføring**.

**3. IAM: Ephemeral, scoped, revocable.** Agenter får aldri service accounts med statiske credentials. Implementer SPIFFE/SPIRE for workload-identitet, 15-minutters token-TTL, og automatisk revokering ved anomali. ABAC-policyer evalueres per kall, ikke per sesjon.

**4. Circuit breaker med «sårbarhetsscore»:** Hvert målsystem klassifiseres etter integritetsrisiko (1-5). Agenter som treffer systemer med score  $\geq 3$  utløser automatisk human-approval-gate. Legacy-monolitter starter på score 4.

---

## Blindsone dere undervurderer

**Latent coupling via data:** Agenten endrer felt X i system A. System B leser X via nattlig batch. Tre uker senere feiler System C. Denne kausalkjeden er usynlig i sanntid. Historisk er dette kilden til de verste feilene — ikke den akutte feilen, men den forsinkede.

---

## Neste steg

- Uke 1-2:** Kartlegg alle skrive-APIer agenten kan nå, klassifiser integritetsrisiko 1-5, og identifiser minst 10 latente koblinger mellom systemer.
- Uke 3-6:** Implementer OPA med allow-list-policyer og append-only execution graph for én pilot-agent i sandkasse — ingen produksjonsdata.
- Uke 7-9:** Gjennomfør «red team»-øvelse der et internt team forsøker å få agenten til å utløse kjedereaksjoner. Dokumenter funnene som grunnlag for tillitsmodellen.