



Debrief — anthropic/claude-opus-4.6

- Instanser: 9
 - Tokens inn: 25,049
 - Tokens ut: 10,654
 - Kostnad: \$0.3916
 - Kjørt: 2026-04-15 17:24:09.528688+00:00
-

Sverm-debrief

Konsensus

1. **Fundamentet mangler.** Alle ni instanser er enige: kombinasjonen av fraværende policy-motor, fragmentert logging, ufullstendig IAM og legacy-systemer uten deterministiske kontrakter gjør at kontrollregimet ikke tåler autonome agenter i dag.
 2. **Kill-switch og audit trail er ufravikelige forutsetninger.** Ingen agent i produksjon uten sentralisert, uforanderlig hendelseslogg og operatørknapp som terminerer alle agentsesjoner innen sekunder.
 3. **Start i skyggemodus.** Agenter som foreslår handlinger uten å eksekvere dem gir læringsdata om emergent behavior, latente koblinger og reell agentadferd — uten produksjonsrisiko.
 4. **Ephemeral credentials, aldri statiske service accounts.** Tidsbegrensede, scoped tokens med kort TTL (15 min) og automatisk revokering ved anomali er minimumskrav.
 5. **Latent coupling er den farligste blindsonen.** Agent endrer felt X → trigger batch-jobb Y → korrupperer system Z uker senere. Denne kausalkjeden er usynlig uten end-to-end correlation og aktiv kartlegging.
-

Dissens

Vent vs. implementer nå. Seks instanser (SI_001-003, 006-008) sier vent — bygg fundament først. To instanser (SI_005, SI_009) sier implementer kontrollert med minimal kontrollamme, fordi venting dreper momentum og læringsvindu. SI_004 lander midt i mellom: bygg kill-infrastruktur først, utforsk i skyggemodus.



Startpunkt: Juridisk eller teknisk? SI_007 og SI_008 krever juridisk sonekartlegging (forvaltningsloven, EU AI Act) før teknisk arkitektur. Øvrige starter med tekniske kontroller. Denne spenningen er uløst og kritisk for offentlig sektor.

Ambisjonsnivå for kontrollplanet. SI_009 argumenterer for at 500k og en PostgreSQL-tabell er nok til å starte. SI_002 og SI_003 krever OPA, ABAC-oppggradering og full event-sourcing som forutsetning. Ressursrealismen spriker.

Blindsoner avdekket

- **Aggregert lesetilgang som GDPR-risiko** (SI_004): Agenten trenger ikke skrivelesetilgang for å bryte personvern — korrelering av data på tvers av systemer med kun lesetilgang kan konstruere ulovlige persondatasett.
- **Irreversibelt kompetansetap** (SI_007): Når agenter håndterer legacy-kompleksitet, mister organisasjonen systemkunnskapen. Innen 3 år eier dere ikke forståelsen av egne systemer.
- **Juridisk vakuum for myndighetsutøvelse** (SI_007/008): Ingen norsk lovhjemmel regulerer eksplisitt autonome agents rett til å iverksette beslutninger med rettsvirkning. Dette er ikke en teknisk risiko — det er en institusjonell.

Anbefalinger

1. **Uke 1-4:** Gjennomfør juridisk sonekartlegging (grønn/gul/rød) og makt-kartlegging av alle planlagte agenthandlinger. Ingen teknisk arkitektur starter før dette foreligger.
2. **Uke 1-6:** Implementer kill-switch og minimal append-only audit-pipeline parallelt — dette har selvstendig verdi uavhengig av agentprosjektet.
3. **Uke 5-16:** Kjør skyggemodus-pilot med én agent, ett avgrenset domene, kun leseoperasjoner, full prompt- og beslutningslogging. Human-in-the-loop på 100 % av foreslåtte handlinger.
4. **Uke 8-12:** Gjennomfør pre-mortem og red team — kartlegg latente koblinger, skjulte batch-jobber og simuler hva agenten kan gjøre med tildelte tilganger.
5. **Måned 6+:** Evaluer skrivelesetilgang kun for systemer som scorer «grønt» på idempotens, transaksjonsstøtte og rollback-kapasitet. Alt annet forblir sperret.