



Debrief — anthropic/claude-opus-4.6

- Instanser: 9
 - Tokens inn: 19,605
 - Tokens ut: 11,064
 - Kostnad: \$0.3746
 - Kjørt: 2026-04-16 00:47:46.915423+00:00
-

Sverm-debrief

Konsensus

1. **Policy-motor (OPA/Cedar) er ufravikelig grunnmur.** Alle ni instanser peker på at eksisterende RBAC er utilstrekkelig for autonome agenter og anbefaler en dedikert policy-motor som mellomlag — implementerbar uten å erstatte IAM-plattformen.
 2. **Graduert tillitsmodell, ikke binær autonomi.** Bred enighet om at agenter må starte i read-only/foreslå-modus og opparbeide rettigheter over tid basert på observert atferd. Nivåinndelingen (observe → propose → execute) går igjen i samtlige analyser.
 3. **Sentralisert, uforanderlig hendelseslogg fra dag 1.** Fragmentert logging identifiseres som den farligste eksisterende svakheten. Alle anbefaler append-only audit-log med korrelerings-ID per agentkjede, uavhengig av agentene selv.
 4. **Eldre fagsystemer er aktive sårbarhetsvektorer.** Manglende idempotens, rollback og semantisk validering i legacy-systemer krever et proxy-/gateway-lag foran hvert integrasjonspunkt — agenter skal aldri kommunisere direkte med fagsystemer.
 5. **Juridisk handlingsrom må avklares før teknisk implementering.** Særlig GDPR art. 22, forvaltningsloven § 40a og legalitetsprinsippet setter harde rammer som direkte dikterer hva agentene kan gjøre.
-

Dissens

Tempo: Implementer nå vs. vent. SI_004 argumenterer for MVP innen uker for å forme kommende standarder. SI_005 og SI_008 mener 6–9 måneder bør brukes



utelukkende på kontrollplanet, uten autonome agenter i produksjon. SI_001 tar mellomposisjon: minimal agent i produksjonslikt miljø innen uke 6 for å avdekke operasjonelle sårbarheter.

Startpunkt: Juridisk vs. teknisk vs. operasjonell. SI_002/SI_003 insisterer på juridisk forankring først. SI_001/SI_009 mener operasjonell erfaring må drive arkitekturen. SI_006/SI_007 prioriterer tillitsinfrastruktur som strategisk fundament.

Blindsoner avdekket

- **Prompt injection via eksterne API-responser** — manipulert tredjepartsdata kan endre agentens beslutningslogikk. Ingen enkeltanalyse ville vektet dette like tungt uten at flere instanser uavhengig flagget det.
- **Kumulativ drift** — små feilhandlinger som individuelt ikke utløser alarm, men akkumulerer til alvorlig datatilstandsavvik over uker. Krever semantisk logging, ikke bare teknisk.
- **Agent-til-agent privilege escalation** — én agent eskalerer til en annen med høyere rettigheter, en vektor tradisjonell IAM ikke håndterer.
- **Identitetskollaps** — RBAC skiller ikke mellom menneske og agent som handler på vegne av menneske, noe som undergraver hele autorisasjonsmodellen.

Anbefalinger

1. **Uke 1-2:** Gjennomfør juridisk-teknisk workshop — kartlegg kandidatprosesser mot hjemmelsgrunnlag og definer maskinlesbare policyregler. Ingen agent deploys uten signert DPIA.
2. **Uke 3-6:** Implementer OPA som policy-gateway med default-deny foran de 3 mest kritiske fagsystemene, og deploy sentralisert audit-pipeline (OpenTelemetry → SIEM) med korrelerings-ID.
3. **Uke 7-12:** Kjør agenter i propose-modus (foreslår, menneske utfører) med proxy-lag og write-ahead-log foran legacy-systemer. Mål MTTD for simulerte feilhandlinger.
4. **Måned 4-6:** Gjennomfør red team-øvelse fokusert på prompt injection, dataeksfiltrering og kaskadesvikt. Resultatene avgjør om skrivetilgang åpnes.
5. **Løpende:** Etabler tverrfaglig syklisk review-team (sikkerhet + jus + fag + KI) med ukentlig kadense — kontrollmodellen må itereres, ikke bare implementeres.