



Debrief — anthropic/claude-sonnet-4.6

- Instanser: 9
 - Tokens inn: 17,240
 - Tokens ut: 8,556
 - Kostnad: \$0.1801
 - Kjørt: 2026-04-16 00:46:45.236863+00:00
-

Sverm-debrief: Agentisk KI - styring og sikkerhet

Konsensus

1. **Gateway-lag er ikke-forhandlingsbart.** Alle 9 instanser peker på at agenter aldri bør ha direkte systemtilgang – alt må gå gjennom ett kontrollert proxy-lag med OPA/Cedar som policy-motor.
 2. **Logging må etableres før pilot, ikke parallelt.** Sentralisert, immutabel audit-log (append-only/WORM) er en forutsetning, ikke et tillegg.
 3. **RBAC er utilstrekkelig.** Eksisterende IAM gir agenter effektiv overautorisasjon. Least-privilege per handling – ikke per rolle – er nødvendig.
 4. **Eldre fagsystemer er den akutte blindsonen.** De mangler rollback-API, audit-trail for maskin-til-maskin-kall, og idempotens-garantier. Agenter bør ikke røre disse uten gateway-isolasjon.
 5. **Ekstern API-lekkasje undervurderes systematisk.** Prompt-kontekst, metadata og interne identifikatorer kan eksfiltreres via utgående kall uten eksplisitt egress-filtrering.
-

Dissens

Pilot nå vs. vent strategisk: SI_001, SI_003 og SI_005 argumenterer for å utsette eller redesigne piloten radikalt (read-only, ingen skrivetilgang). SI_006, SI_007 og SI_008 argumenterer for å implementere nå, men med observerbar arkitektur fremfor perfekt kontroll. Kjernespenningen: er 6–9 måneder nok til å bygge tilstrekkelig tillit, eller skaper tidspress uakseptabel teknisk gjeld?



Sandkasse vs. reversibilitet: SI_006 dissentierer eksplisitt – sandkasser vil bli utvidet under politisk press. Design for reversibilitet er mer robust enn isolasjon.

Blindsoner avdekket

- **Ansvarsdiffusjon hos brukerne** (SI_002): Automatiserings-bias gjør at ansatte slutter å forstå – og eie – agentavgjørelser etter 12–18 måneder. Ingen teknisk kontroll løser dette.
- **Maktkonsentrasjon hos systemeierne** (SI_006, SI_008): Den enheten som eier API-nøkler og policy-definisjoner, eier i praksis prosessene. Dette er et organisatorisk maktspørsmål, ikke et teknisk.
- **Kaskadeeffekter mellom agenter** (SI_004): Agent A trigger agent B – feil propagerer eksponentielt. Undervurdert i alle standard risikovurderinger.
- **Nøkkelrotasjon under aktiv sesjon** (SI_002): Agent holder sesjon åpen lenger enn token-TTL → stille autentiseringsbrudd uten feilmelding.

Anbefalinger

1. **Frys agentens handlingsrom til maks 3 domener** – skriv en formell capability boundary-spesifikasjon før noe annet besluttet (uke 1–2).
2. **Implementer OPA-gateway som dag 1-infrastruktur** foran én avgrenset integrasjon – ikke som ettertanke i måned 5 (uke 3–4).
3. **Gjennomfør blast radius-analyse** på de 3 mest kritiske fagsystemene og kart alle irreversible operasjoner – disse er eksplisitt forbudt sone i pilot (uke 3–4).
4. **Redesign piloten som read-only beslutningsstøtte** med skriveadgang først etter fullstendig audit-log er verifisert og red team-øvelse er gjennomført.
5. **Avklar kill-switch-autoritet skriftlig** – hvem har myndighet til å stoppe agenten? Dette er organisatorisk, ikke teknisk, og må besluttet før pilot starter.