



# Debrief — anthropic/claude-sonnet-4.6

---

- Instanser: 9
  - Tokens inn: 23,118
  - Tokens ut: 8,637
  - Kostnad: \$0.1989
  - Kjørt: 2026-04-15 17:23:07.112094+00:00
- 

## Sverm-debrief: Agentisk KI i offentlig sektor

---

### Konsensus

1. **Latent coupling er den farligste enkeltrisikoen.** Alle ni instanser peker på udokumenterte avhengigheter mellom legacy-systemer som den mest undervurderte trusselen – ikke agenten selv, men koblingene mellom systemer som aldri var designet for autonom interaksjon.
  2. **Ephemeral credentials er ikke-forhandlbart fra dag én.** Maks 15 minutters TTL via Vault eller Azure Managed Identity er konsensus – langlivede service accounts er uakseptabelt.
  3. **Start read-only.** Ingen skrivetilgang, ingen eksterne API-kall i pilotfasen. Dette er 80% av verdien til 20% av risikoen.
  4. **Idempotens-klassifisering av alle API-er må gjøres før piloten starter.** Ikke-idempotente endepunkter blokkeres automatisk fra agentens scope.
  5. **Audit-pipeline er forutsetning, ikke etterarbeid.** Strukturert logging med korrelasjon-ID per agent-kjøring til én sentral sink må være på plass før agenten handler.
- 

### Dissens

**Tempo vs. forsiktighet:** SI\_001/SI\_003 anbefaler iterativ åpning basert på bevist adferd. SI\_004/SI\_006 argumenterer for at 6–9 måneder er for kort til å gjøre dette forsvarlig overhodet – piloten bør snevres drastisk inn eller utsettes. Dette er en reell motsetning uten enkel løsning.



**Logging vs. grenser først:** SI\_002 dissenter eksplisitt fra mainstream: implementer harde operasjonelle grenser før logging, ikke omvendt. Observasjon uten kontroll gir falsk trygghet.

**OPA nå vs. OPA senere:** SI\_008 krever OPA som forutsetning. SI\_001/SI\_003 mener én enkel gateway-sjekk er tilstrekkelig for pilot. Ressursbegrensningen gjør dette til et reelt valg, ikke et akademisk spørsmål.

---

## Blindsoner avdekket

**Juridisk ansvarsvakuum (SI\_009, kun delvis berørt av andre):** Ingen teknisk kontroll løser spørsmålet om hvem som bærer strafferettslig ansvar når agenten feilberegner en ytelse. Forvaltningsloven §2b krever hjemmel for automatisering av enkeltvedtak – dette er et showstopper-spørsmål som må avklares før arkitekturvalg, ikke etter.

**Maktakkumulering over tid (SI\_002/SI\_007/SI\_008):** Agenten bygger de facto systemkontroll gjennom legitim bruk – API-mønstre, datatilganger og prosesskoblinger som er vanskelige å reversere etter 6–9 måneder. Dette er ikke en sikkerhetsrisiko, det er en styringsrisiko ingen enkelt instans ville fanget alene.

---

## Anbefalinger

1. **Juridisk risikoanalyse denne uken** – kartlegg hvilke agenthandlinger berører enkeltvedtak eller lovpålagte prosesser. Dette er go/no-go-kriterium, ikke etterarbeid.
2. **Idempotens-audit av topp-10 API-endepunkter** innen uke 2 – test eksplisitt for dobbelt-kall. Blokkér alle ikke-idempotente endepunkter fra pilotscope.
3. **Definer én navngitt person** (ikke rolle) med mandat til å isolere agenten innen 15 minutter. Skriv dette inn i pilotmandatet.
4. **Shadow-modus i 90 dager** – agenten logger og foreslår, eksekverer ikke. Emergent behavior kartlegges før reell makt gis.
5. **Skriv “Agent Charter”** – hvilke systemer er off-limits uansett teknisk mulighet. Juridisk bindende internt, signert av governance-laget før pilotstart.