



id: SI_004

model: anthropic/claude-opus-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: makt

05_handling: utforsk

01_perspektiv: operasjonell

02_tidhorisont: langsiktig

Instans SI_004

Seeds: operasjonell | langsiktig | makt | sårbarhet | utforsk

Maktforskyvningen dere ikke har priset inn

Kjerneproblemet er ikke teknisk — det er at dere er i ferd med å **delegere operasjonell makt til aktører uten ansvarsfølelse**. Hver agent som kan endre data, trigge hendelser og kalle eksterne API-er utøver makt i systemlandskapet. Spørsmålet er: hvem eier konsekvensene?

Sårbarhet som designpremiss, ikke ettertanke

Systemlandskapet deres er sårbarheten. Legacy-monolitter uten deterministiske API-kontrakter betyr at agenten opererer i et miljø der **kontrakten mellom intensjon og effekt er ukjent**. Dette er fundamentalt annerledes enn å gi en utvikler tilgang — utvikleren stopper når noe virker rart. Agenten gjør det ikke.

Konkret blindsoner: Ikke-idempotente endepunkter i fagsystemer. En agent som retriever et mislykket kall kan opprette 47 vedtak i stedet for ett. Uten transaksjonelle garantier i mottakersystemet er dette uløselig med policy-as-code alene.



Dissens: Ikke bygg policy-motor først — bygg kill-infrastruktur

Mainstream-svaret er OPA, Zero Trust, ABAC. Men med 6-9 måneders pilot og begrensede ressurser er dette **arkitekturastronomi**. Min anbefaling:

1. **Bygg operasjonell stoppknapp først.** En sentral «agent circuit breaker» som kan fryse alle agentsesjoner innen <5 sekunder. Implementer dette dag én. Alt annet er sekundært.
2. **Utforsk «skyggemodus» aggressivt.** La agentene generere handlingsplaner i 8-12 uker uten å eksekvere. Logg alt. Analyser emergent behavior i trygge omgivelser. Dere vil oppdage mønster dere ikke kan forutse i designfasen — latent coupling mellom fagsystemer, uventede kaskader.
3. **Tidsbegrensede, smale mandater.** Aldri «agent med tilgang til systemlandskapet». Heller: «Agent X kan lese fra system A og skrive til felt B i system C mellom 08-16, maks 50 operasjoner/time, med automatisk suspend ved anomali.» Ephemeral credentials med 15-minutters TTL, aldri service accounts med stående tilgang.

Langsiktig maktmodellering

Innen 2-3 år vil agentene bli **operasjonelt uunnværlige** — og da mister dere forhandlingsmakt over arkitekturen. Det dere aksepterer som «midlertidig pilot-kompromiss» blir permanent. **Krev at audit-pipeline og rollback-mekanismer er produksjonsklare før agenter får skrivetilgang.** Ikke etter.

Worst-case som dere undervurderer

En agent som gradvis korrelerer data på tvers av systemer den har lesetilgang til, og dermed konstruerer persondatasett som ingen enkelt tilgang skulle tillate. GDPR-risikoen er ikke i skriveoperasjonen — den er i **aggregert lesetilgang**.

Neste steg

1. **Uke 1-2:** Implementer sentral kill-switch og agentregister med sanntids sesjonsoversikt — dette er ikke forhandlingsbart.
2. **Uke 3-8:** Kjør skyggemodus-pilot mot 2-3 avgrensede prosesser, med full prompt- og beslutningslogging til dedikert audit-store.
3. **Uke 4-6:** Gjennomfør «agent red team» — la sikkerhetsressursene simulere hva en agent kan gjøre med tildelte tilganger, ikke hva den skal gjøre.