



id: SI_003

model: anthropic/claude-opus-4.6

seeds:

04_risiko: usikkerhet

03_relasjon: makt

05_handling: implementer

01_perspektiv: operasjonell

02_tidhorisont: langsiktig

Instans SI_003

Seeds: operasjonell | langsiktig | makt | usikkerhet | implementer

Kjerneanalyse: Makt uten brems er den reelle risikoen

La meg være direkte: Dere er i ferd med å innføre en **ny maktaktør** i systemlandskapet. En agent som kan endre data, trigge prosesser og kalle eksterne API-er er ikke et verktøy — den er en operatør med delegert myndighet. Spørsmålet er ikke om noe går galt, men **hvor lenge det tar før en feilkjede eskalerer forbi menneskelig reaksjonstid**.

Dissens mot mainstream-tilnærmingen

Konsensussvaret er “implementer OPA, Zero Trust og audit-pipeline.” Mitt perspektiv: **med 6–9 måneder, begrenset DevSecOps og fragmentert logging har dere ikke kapasitet til å bygge kontrollregimet som agentisk KI krever**. Å implementere agenter før kontrollplanet er modent er å gi fra seg operasjonell makt under maksimal usikkerhet.

Hva usikkerheten faktisk betyr

- **Legacy-systemer uten idempotente API-er:** Én agent-retry kan opprette doble vedtak, doble utbetalinger, doble registreringer. Kompenserende handlinger i monolitter uten transaksjonsstøtte er ofte manuell brannslukking.



- **Emergent behavior:** Agent A kaller API X som trigger event Y som agent B plukker opp. Denne **latente koblingen** er usynlig i arkitekturdokumentasjon og oppstår først i produksjon.
- **Maktkonsentrasjon i service accounts:** Én kompromittert ephemeral credential gir en angriper en autonom aktør som allerede har tillitskjeden innebygd.

Langsiktig riktig rekkefølge

Implementer **kontrollplanet først, agentene etterpå:**

1. **Fase 0 (måned 1-4):** Konsolidert logging med immutable audit trail (f.eks. OpenTelemetry → objektlager med WORM). Ingen agent uten sporbarhet.
2. **Fase 1 (måned 3-6):** Policy-motor (OPA/Cedar) med eksplisitte allow-lists per agentrolle. Agenter får **kun lesing** — ingen skriveoperasjoner.
3. **Fase 2 (måned 6-9):** Pilot med **én agent, ett avgrenset domene, human-in-the-loop på alle skriveoperasjoner**, circuit breaker med maks 50 kall/minutt, automatisk kill-switch ved anomali.
4. **Fase 3 (måned 9+):** Gradvis autonomi basert på observert adferd, ikke antatt tillit.

Operasjonell realitet

Worst case er ikke at agenten gjør én feil. Det er at den gjør **200 feil i løpet av 30 sekunder** mot et fagsystem uten rate-limiting, og at ingen oppdager det fordi logging er distribuert. Modeller dette eksplisitt: hva skjer dersom agenten looper mot Folkeregisteret, økonomisystemet, arkivsystemet?

Neste steg

1. **Gjennomfør en makt-kartlegging:** Dokumentér nøyaktig hvilke skriveoperasjoner, datasett og systemer agentene ville fått tilgang til — og vurder konsekvens av misbruk for hvert enkelt endepunkt.
2. **Bygg kill-switch og audit trail FØR første agent-pilot:** Ingen agent i produksjon uten sentralisert, uforanderlig logg og operatørknapp som dreper alle agent-sesjoner innen 5 sekunder.
3. **Definer en eksplisitt «usikkerhetsterskel»:** Hvilke systemegenskaper (manglende idempotens, ingen rollback, svak kontrakt) diskvalifiserer et system fra agenttilgang — og håndhev dette som policy.