



id: SI_007

model: google/gemma-4-26b-a4b-it

seeds:

04_risiko: sårbarhet

03_relasjon: tillit

05_handling: forhandle

01_perspektiv: økonomisk

02_tidhorisont: langsiktig

Instans SI_007

Seeds: økonomisk | langsiktig | tillit | sårbarhet | forhandle

Analysen av agentisk KI i en offentlig kontekst må skifte fra “funksjonalitetsfokus” til en modell basert på **kontrollert eksponering**.

Arkitektoniske prinsipper og sikkerhetsmekanismer

For å sikre **langsiktig** stabilitet må vi gå bort fra tanken om at agenten er en “bruker”. Agenten er en risiko-enhet.

- Policy-as-Code (PaC) som forhandlingsgrunnlag:** Siden IAM-plattformen deres er for grov, må dere bygge et mellomlag (f.eks. ved bruk av Open Policy Agent - OPA). Dette fungerer som en digital kontrakt; agenten får kun utføre handlinger som er eksplisitt godkjent i en regelmotor, uavhengig av RBAC-nivå.
- Økonomisk isolasjon (Sandboxing):** Implementer “Transaction Limits”. En agent skal ikke ha ubegrenset tilgang til å endre data. Hver handling må ha en “verdi” eller et risikonivå. Overskrides en terskel (f.eks. > 5 endringer på samme datasett i minuttet), kreves menneskelig intervensjon. Dette er en **økonomisk** sikring mot “runaway loops”.
- Observability som Tillit:** Fragmentert logging er en fatal **sårbarhet**. Dere må etablere en sentralisert “Agent Traceability Log” som korrelerer agentens resonnement (hvorfor den gjorde det) med systemets handling (hva som skjedde). Uten dette brytes den offentlige **tilliten** ved første uforklarlige feil.



Tekniske blindsoner og undervurderte risikoer

- **Legacy-sårbarhet:** Eldre systemer forstår ikke “intensjon”. En agent som prøver å “rydde opp” i data kan trigge uforutsette kaskadefeil i systemer uten rollback-kapasitet. Dette er en kritisk **sårbarhet** i arkitekturen.
- **Kontekst-lekkasje:** Den største risikoen er ikke nødvendigvis direkte angrep, men at agenten gjennom eksterne API-kall (f.eks. til en LLM) “leker” sensitiv kontekst som en del av sin beslutningsprosess.

Kontrollmodell: “The Human-in-the-Loop Negotiator”

Ikke design for full autonomi, design for **forhandlet autonomi**. Agenten opererer i lavrisiko-soner autonomt, men ved usikkerhet eller høyverdi-transaksjoner må den “forhandle” med et menneske via et eskalering-grensesnitt. Dette balanserer effektivitet mot sikkerhet.

Neste steg:

1. **Pilot-avgrensning:** Velg én isolert prosess i et moderne API-miljø (ikke legacy) for 6-måneders piloten.
2. **Implementer OPA:** Sett opp en prototype på Open Policy Agent for å skape et finmasket kontrollag over eksisterende RBAC.
3. **Definer “Kill-switch” protokoll:** Utarbeid en teknisk prosedyre for umiddelbar terminering av alle aktive agent-sesjoner.