



id: SI_002

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

seeds:

04_risiko: mulighet

03_relasjon: makt

05_handling: vent

01_perspektiv: juridisk

02_tidhorisont: langsiktig

Instans SI_002

Seeds: juridisk | langsiktig | makt | mulighet | vent

Analysen av agentisk KI i deres kontekst avdekker en fundamental mismatch mellom deterministisk arkitektur og probabilistisk autonomi.

Arkitekturprinsipper og maktbalanse Dere må slutte å se på sikkerhet som en “mur” og begynne å se på det som en **kontraktuell styring**. Siden deres IAM støtter RBAC, men mangler finmasket kontroll, vil agentene i praksis utøve uregulert **makt** hvis de tildeles eksisterende tjenestekontoer. Prinsipp: **“Policy-as-Code Enforcement”**. Implementer en sentralisert Policy Decision Point (PDP) (f.eks. Open Policy Agent - OPA) mellom agenten og API-ene. Agenten må ikke ha direkte tilgang, men må be om en “handlings-token” som valideres mot kontekstuelle regler.

Sikkerhetsmekanismer og de tekniske blindsonene Den største blindsonen er **“Semantisk Eskalering”**. En agent kan bruke legitime API-kall til å utføre illegitime handlinger ved å manipulere dataflyten (f.eks. endre en utbetalingskonto via tre separate, tilsynelatende ufarlige kall). Mekanisme: Innfør **“Human-in-the-loop” (HITL) for irreversible handlinger** (f.eks. sletting, store transaksjoner). Bruk Rate-limiting ikke bare på antall kall, men på endringsfrekvens i kritiske datafelt.

Juridisk risiko og langsiktig ansvar Fra et **juridisk** perspektiv er fragmentert logging en katastrofe. Hvis en agent tar en autonom beslutning som bryter forvaltningsloven, er det ikke agenten, men virksomheten som står ansvarlig. Uten et uforanderlig (immutable) revisjonsspor som knytter agentens “tankeprosess” (reasoning logs) til den faktiske systemendringen, har dere ingen rettslig etterrettelighet.

Strategisk dissens: Vent Mainstream-rådet vil si: “Gå i gang med pilot nå”. Mitt råd er: **Vent**. Deres infrastruktur er ikke moden. Å rulle ut agenter på eldre fagsystemer uten integrasjonsstøtte er å bygge et høytårn på kvikksand. Risikoen for “silent



corruption” – hvor data endres gradvis uten at det trigger tradisjonelle alarmer – er for høy.

Neste steg:

1. **Kartlegg kritiske beslutningspunkter:** Identifiser de 5 mest irreversible handlingene i systemet og definer krav til manuell godkjenning (HITL).
2. **Prototype en “Intermediary Proxy”:** Bygg en enkel kontroll-gateway (f.eks. med OPA) som loggfører agentens intensjon før API-kallet utføres.
3. **Etabler “Observability-first”:** Ikke implementer agenter før dere har en sentralisert logg-strøm som kan korrelere agent-ID med systemendringer i sanntid.