



id: SI_001

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

seeds:

04_risiko: mulighet

03_relasjon: makt

05_handling: implementer

01_perspektiv: strategisk

02_tidhorisont: langsiktig

Instans SI_001

Seeds: strategisk | langsiktig | makt | mulighet | implementer

Analysen av agentisk KI i deres kontekst handler ikke om teknisk integrasjon, men om en fundamental forskyvning av **makt** fra deterministisk kontroll til probabilistisk autonomi. Å implementere agenter i et fragmentert landskap uten å redesigne kontrollaget er en **strategisk** risiko som kan føre til systemisk kollaps.

Arkitektonisk rammeverk: “The Guardrail Supremacy”

For å sikre **langsiktig** stabilitet må dere slutte å se på agenter som “brukere” og begynne å se på dem som uforutsigbare aktører.

- Policy-as-Code (PaC) fremfor RBAC:** Dagens RBAC er utilstrekkelig. Dere må implementere en sentralisert policy-motor (f.eks. Open Policy Agent - OPA). Agenter må operere under “Attribute-Based Access Control” (ABAC), der handlinger valideres mot kontekst (tid, datavolum, sensitivitet) i sanntid, ikke bare identitet.
- Intermediary Proxy Layer (The Sandbox):** Ikke la agenter snakke direkte med eldre fagsystemer. Implementer en “Agent Gateway”. Dette er en kontrollert sandkasse som fungerer som en protokoll-konverter og en regulator. Her innføres Rate-limiting (for å hindre loop-induserte DoS-angrep) og Semantic Filtering (for å hindre data-eksfiltrering).
- Immutable Audit Trails:** Siden logging er fragmentert, må dere tvinge alle agent-handlinger gjennom en dedikert, uforanderlig hendelsesstrøm (f.eks. en kryptert Kafka-logg) som kobler agentens resonnement (LLM-logg) med dens handling (API-kall).



Tekniske blindsoner og risiko

Den største **muligheten** ligger i automatisering, men den største blindsonen er **“Cascading Failure via Autonomy”**. Et system som ikke er designet for ikke-menneskelig interaksjon, vil tolke en agents feilaktige, men syntaktisk korrekte, kommandoer som legitime instruksjoner. Dette kan føre til irreversible datakorupsjon i eldre systemer før en menneskelig operatør oppdager avviket.

Strategisk kontrollmodell

Design en **“Human-in-the-loop-by-exception”** modell. Agenter har full autonomi i lav-risiko soner, men ved kritiske terskelverdier (f.eks. endring av over \$X\$ kr, eller endring av \$>Y\$ personopplysninger) kreves en asynkron godkjenning.

Neste steg:

1. **Etabler en “Agent Gateway” prototype:** Bygg en proxy mellom en agent og ett enkelt eldre system for å teste policy-styrte begrensninger.
2. **Definer “Blast Radius”:** Kartlegg hvilke systemer som kan forårsake irreversibel skade og isoler disse med strengere sandboxing.
3. **Implementer OPA:** Start implementering av Policy-as-Code for å erstatte statisk RBAC i piloten.