



id: SI\_006

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

seeds:

04\_risiko: sårbarhet

03\_relasjon: makt

05\_handling: vent

01\_perspektiv: operasjonell

02\_tidhorisont: langsiktig

## Instans SI\_006

**Seeds:** operasjonell | langsiktig | makt | sårbarhet | vent

Analysen skifter fra tradisjonell IT-sikkerhet til kontroll over en ny type digital aktør.

### Arkitektur og Kontroll: Fra Determinisme til Probabilistisk Styring

Dagens arkitektur er bygget for instruksjoner. Agentisk KI opererer på intensjoner. Dette skaper en fundamental **sårbarhet**: gapet mellom hva en agent tror den skal gjøre, og hva den faktisk gjør i et legacy-system.

#### Operasjonelle prinsipper:

- **Policy-as-Code (PaC) fremfor RBAC:** Glem statiske roller. Du trenger en dynamisk policy-motor (f.eks. OPA - Open Policy Agent) som evaluerer hver handling: 

```
if (action == 'DELETE' && confidence < 0.95) { require_human_approval() }.
```
- **Mikro-isolasjon (Sandboxing):** Hver agent-sesjon må kjøres i en kortlevd, isolert container med strengt definerte egress-regler. Ingen agent skal ha direkte tilgang til database-strings; de skal kun kommunisere via en "Gatekeeper API".

**Makt og Systemisk Risiko:** Når agenter får tilgang til API-er som ikke er designet for autonomi, oppstår en **maktforskyvning**. Agenter kan utnytte "race conditions" i eldre systemer eller utføre "logic bombs" ved å kombinere legitime kall på uforutsette måter. Den største tekniske blindsonen er **kaskadefeil**: En agent prøver å "fikse" en datafeil, endrer en parameter, som trigger en feil i et legacy-system, som igjen fører til en massiv datakorupsjon.

**Langsiktig strategi vs. Kort sikt:** Pilotprosjektet på 6-9 måneder er en **sårbarhet** i seg selv. Hastighet tvinger frem snarveier i logging. Uten en "Immutable Audit Log"



(skrivebeskyttet loggføring som agenten ikke kan slette), mister dere evnen til etterprøvbarhet når noe feiler.

**Vent - Ikke rull ut full autonomi:** Implementer en “Human-in-the-loop” (HITL) eskalering som en hard teknisk sperre, ikke et administrativt ønske.

### Konkrete neste steg:

1. **Etabler en “Shadow API Layer”:** Bygg et mellomlag mellom agentene og legacy-systemene som fungerer som en kontrollert proxy med innebygd rate-limiting og semantisk validering.
2. **Definer “Kill-switch” protokoll:** Utvikle en teknisk mekanisme for øyeblikkelig terminering av alle aktive agent-sesjoner og automatisk rollback av de siste 5 minutters endringer.
3. **Implementer OPA (Open Policy Agent):** Start umiddelbart med å kode handlinger som logiske regler, ikke bare brukerrettigheter.