



id: SI_003

model: anthropic/claude-sonnet-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: makt

05_handling: vent

01_perspektiv: juridisk

02_tidhorisont: langsiktig

Instans SI_003

Seeds: juridisk | langsiktig | makt | sårbarhet | vent

Maktforskyvningens stille risiko

Det som ikke diskuteres åpent: agentisk KI omfordeler makt internt. Når en agent kan initiere prosesser og endre data uten menneskelig godkjenning, forflyttes beslutningsmyndighet fra saksbehandler til systemarkitekt – og til leverandøren bak agenten. Dette er et juridisk og organisatorisk problem like mye som et teknisk et.

Vent er faktisk et legitimt strategisk valg her. 6–9 måneder med fragmentert logging, RBAC uten finmasket policy-styring og eldre fagsystemer er ikke en pilotbetingelse – det er en ulykkesoppskrift. EU AI Act (som gjelder fra 2026) klassifiserer sannsynligvis slike systemer i offentlig sektor som høyrisiko-KI, med krav om revisjonsspor, menneskelig oversikt og risikovurdering før deployment. Å presse en pilot gjennom nå skaper **juridisk gjeld** som er vanskelig å reversere.

Konkrete tekniske sårbarheter som undervurderes

- **Lateral privilege escalation:** En agent med legitimt begrenset tilgang kan kombinere API-kall på tvers av systemer og oppnå effektiv tilgang ingen enkeltrolle skulle hatt. RBAC alene stopper ikke dette.
- **Prompt injection via datakilder:** Hvis agenten leser ekstern data (e-post, API-svar), kan fiendtlig innhold omdirigere agentens handlinger. Dette er dokumentert i GPT-4-baserte agenter siden 2023.
- **Irreversibilitetsblindhet:** Eldre fagsystemer har sjelden rollback-API. Én feil skriveoperasjon kan kreve manuell datareparasjon over dager.



Arkitekturprinsipp som mangler i debatten

Innfør **“capability containment by default”**: agenten har ingen rettigheter initialt, og hvert handlingsrom tildeles eksplisitt med TTL (time-to-live). Dette er radikalt annerledes enn å starte bredt og innsnevre – det krever mer design, men bygger langsiktig kontroll som overlever turnover og leverandørbytte.

Logging må behandles som **juridisk bevismateriell fra dag én** – ikke som driftslogg. Det betyr uforanderlig lagring (WORM), tidsstempling og separat tilgangsstyring.

Neste steg

1. **Gjennomfør en juridisk risikovurdering** mot AI Act og Sikkerhetsloven før piloten godkjennes – ikke parallelt.
2. **Kart alle irreversible operasjoner** i de aktuelle fagsystemene og sett disse som eksplisitt forbudt sone for agenter i pilot-fase.
3. **Utsett piloten med 3 måneder** og bruk tiden på å implementere sentralisert, uforanderlig audit-logging som eneste ikke-forhandlingsbare infrastrukturkrav.