



id: SI_003

model: anthropic/claude-opus-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: samarbeid

05_handling: utforsk

01_perspektiv: juridisk

02_tidhorisont: langsiktig

Instans SI_003

Seeds: juridisk | langsiktig | samarbeid | sårbarhet | utforsk

Dissens: Ikke start med teknisk arkitektur — start med juridisk handlingsrom

De fleste vil angripe dette som et arkitekturproblem. Mitt perspektiv: den farligste blindsonen er at dere bygger en autonom agent uten å ha juridisk forankret hva agenten har lov til å gjøre. I offentlig sektor er legalitetsprinsippet ikke valgfritt — enhver handling som endrer data eller initierer vedtak krever hjemmel. En KI-agent som autonomt endrer saksdata i et fagsystem utfører potensielt myndighetsutøvelse uten lovhjemmel.

Sårbarhet som designprinsipp

Eldre fagsystemer uten fullverdig integrasjonsstøtte er ikke bare teknisk gjeld — de er **aktive sårbarhetsvektorer**. Når en autonom agent får skrive tilgang via umoderne grensesnitt (SOAP, filbasert integrasjon, screen-scraping), kollapser hele sikkerhetskjeden:

- **Ingen semantisk validering:** Agenten kan sende syntaktisk korrekte, men logisk destruktive kall
- **Manglende tilbakekallsmekanismer:** Fagsystemer fra 2005–2015 har sjelden transaksjonell rollback via API
- **Identitetskollaps:** RBAC-systemet deres skiller ikke mellom “ansatt bruker X” og “agent som handler på vegne av X” — dette er en fundamental svakhet som OPA/Cedar-basert policystyring må løse før pilot



Langsiktig kontrollmodell gjennom samarbeid

Kontrollmodellen må designes for **10 år, ikke 9 måneder**. Konkret:

1. **Juridisk-teknisk kontrakt per agent:** Hvert agentscenario får et maskinlesbart policydokument (f.eks. Cedar/Rego) som speiler en juridisk vurdering. Ingen agent deployes uten signert DPIA og juridisk notat.
2. **Felles sårbarhetskatalog:** Etabler tverretattlig samarbeid (f.eks. med Digitaliseringsdirektoratet, NSM) om en delt trussel-/sårbarhetskatalog for agentisk KI i offentlig sektor. Dere er ikke alene om dette problemet.
3. **Utforskende sandkasse med juridiske grenser:** Policy-sandkassen bør ikke bare begrense tekniske handlinger — den bør implementere juridiske invarianter. Eksempel: agenten kan aldri endre vedtaksstatus i fagsystem Y uten at en menneskelig saksbehandler har godkjent innenfor siste 300 sekunder.

Undervurderte risikoer

- **Informasjonslekkasje via prompt-innhold:** Agenter som kaller eksterne API-er kan eksponere taushetsbelagt informasjon i request-payloads — dette er et brudd på forvaltningsloven § 13
- **Kumulativ drift:** Små feilhandlinger akkumulerer over tid uten at noen enkelthandling utløser alarm. Uten semantisk logging (ikke bare teknisk) oppdages dette for sent
- **Ansvarsfraskrivelse:** Når ingen person “trykket på knappen”, hvem eier feilen juridisk? Dette må avklares før pilot, ikke etter tilsynsklage

Konkrete neste steg

1. **Uke 1-3:** Gjennomfør juridisk-teknisk workshop der jurist og arkitekt sammen kartlegger hvert planlagt agentscenario mot hjemmelsgrunnlag og lager maskinlesbare policyregler (Cedar/Rego)
2. **Uke 4-6:** Implementer en identity-bifurcation-løsning i IAM — agenter får egne identiteter med separat autorisasjonsflyt, aldri gjenbruk av brukeridentiteter. Test mot de to mest sårbare fagsystemene først.
3. **Parallelt:** Inviter NSM og Digdir til et utforskende samarbeidsforum om agentisk KI i offentlig sektor — dere bygger presedens, og presedens bør bygges åpent.