



id: SI_002

model: anthropic/claude-sonnet-4.6

seeds:

04_risiko: sårbarhet

03_relasjon: konkurranse

05_handling: vent

01_perspektiv: bruker

02_tidhorisont: langsiktig

Instans SI_002

Seeds: bruker | langsiktig | konkurranse | sårbarhet | vent

Hovudperspektiv: Brukaren er den gløymde sårbarheita

Mainstream-diskusjonen om agentisk KI-sikkerheit dreier seg om teknologi. Min analyse peikar ein annan veg: **den langsiktige risikoen ligg i at brukaren vert fråkopla ansvaret** — og dette skapar ei sårbarheit ingen sandbox kan fikse.

Den undervurderte risikoen: Ansvarsdiffusjon

Når agenten handlar “utan direkte menneskeleg styring”, skjer det psykologisk noko farleg. Brukarar sluttar å forstå kva systemet gjer på deira vegne. Etter 12–18 månader vil du ha tilsette som signerer på agentavgjerder dei ikkje kan forklare. Dette er ikkje spekulasjon — det er dokumentert i studiar av automatiserings-bias (Parasuraman & Manzey, 2010).

Konkret konsekvens: Ein agent med RBAC-rolle “saksbehandler_auto” kan endre 400 datasett før ein menneske oppdagar feilen. Rollback finst teknisk, men ingen eig avgjerda om når rollback skal skje.

Vent — det er faktisk eit arkitekturprinsipp

6–9 månaders pilot er for kort til å etablere langsiktig kontroll. Konkurransen om å vise KI-resultat raskt pressar fram teknisk gjeld i sikkerheitsarkitekturen. Mi tilråding, som avvik frå normen: **bygg inn obligatoriske “ventepoeng”** — ikkje berre for irreversible handlingar, men for alle handlingar som kryssar systemgrenser. Dette er ikkje flaskehalsar; det er sårbarheitskartlegging i sanntid.



Tre konkrete blindsoner

1. **Eldre fagsystem utan API-logging** — agenten handlar, ingenting vert skrive. Sjansen for å oppdage misbruk: nær null.
 2. **Nøkkelrotasjon under agentsesjonen** — agenten held ein sesjon open i 40 minutt; nøkkelen roterast etter 30. Autentiseringsbrot utan feilmelding til brukar.
 3. **Eksterne API-kall med implisitt datalekkasje** — prompt-konteksten til agenten kan innehalde persondata som vert sendt til LLM-leverandør utan databehandlaravtale.
-

Neste steg

1. **Gjennomfør ein “brukarfråøving-analyse”**: kartlegg kva avgjerder tilsette i dag tek som agenten vil overta — og kven som eig ansvaret etter overtak.
2. **Krev synkron logg på tvers av alle systemer før pilot startar** — ikkje etterpå. Ingen agenthandling utan sporbarheit.
3. **Utset ekstern API-integrasjon til månad 4**, ikkje månad 1. Bygg innafor murane fyrst, lær sårbarheitsmønsteret, ekspander deretter.