



id: SI_009

model: mistralai/mistral-large-2512

seeds:

04_risiko: sårbarhet

03_relasjon: konkurranse

05_handling: forhandle

01_perspektiv: juridisk

02_tidhorisont: kortsiktig

Instans SI_009

Kortsiktig juridisk sårbarhet i en konkurransepreget autonomi

Dere står foran en **juridisk tidsinnstilt bombe**. 6-9 måneder er en urealistisk kort tidsramme for å etablere robust kontroll over autonome agenter i et fragmentert systemlandskap – spesielt når **konkurransen** om å være først ute med KI-drevne tjenester presser beslutninger. Den største **sårbarheten** ligger ikke i teknologien, men i **forhandlingsrommet** mellom innovasjon og compliance.

Tre kritiske blindsoner med juridisk eksponering:

1. **RBACs “alt-eller-intet”-felle:** Nåværende IAM støtter ikke **finmaskede policyer** for autonome agenter. Løsningen? Et **kortsiktig forhandlingslag** – en proxy-tjeneste som oversetter agenthandling til RBAC-kompatible kall, med manuell godkjenning for handlinger utenfor predefinerte “sikre korridorer”.
Eksempel: Agenten kan foreslå en fakturabetaling, men må vente på menneskelig sign-off før utførelse.
2. **Logging som juridisk skjold:** Fragmentert sporbarhet gjør det umulig å bevise hva agenten visste når den handlet. **Krav:** Alle agentbeslutninger logges med (1) input-data (hash), (2) beslutningsmodell (versjon), (3) ekstern API-respons. Bruk **juridiske metadata** – lagre loggene i et uforanderlig format (f.eks. blockchain-lignende struktur) for å møte dokumentasjonskrav i personvernforordningen.
3. **Eksterne APIers “svart boks”-risiko:** Agentens interaksjon med eksterne tjenester skaper **juridisk ansvarsvakuum**. Løsning: **Forhandle inn klausuler** i API-kontrakter som krever (a) at leverandører logger alle kall fra deres side, (b) at de aksepterer delt ansvar for feilhandling, og (c) at de gir dere tilgang til revisjonslogger ved tvist.



Arkitekturprinsipper for kortsiktig kontroll:

- **Sandboxing via “juridiske skott”:** Isoler agentene i domener basert på juridisk risiko (f.eks. “personopplysninger”, “økonomiske transaksjoner”). Bruk **policy-motorer** som automatisk blokkerer handlinger som krysser skottgrensene uten menneskelig godkjenning.
- **Rate-limiting som sårbarhetsbuffer:** Begrens agentenes handlinger til X kall/time per system – ikke av tekniske årsaker, men for å **forhandle ned risikoen** for eskalering. Eksempel: Maks 5 fakturaendringer/time, med varsling til økonomisjef ved grenseoverskridelse.
- **Rollback som juridisk nødbrems:** Design alle agenthandlingene som **reversible** innen 72 timer. For uopprettelige handlinger (f.eks. sletting av data), krev **dobbelt sign-off** via to separate IAM-roller.

Neste steg:

1. **Juridisk risiko-workshop:** Kartlegg hvilke agenthandlingene som kan utløse erstatningsansvar eller brudd på personvernregler. Prioriter disse i piloten.
2. **Forhandlingsplan for API-leverandører:** Utarbeid en mal for kontraktstillegget som adresserer logging, ansvarsdeling og revisjonsrettigheter.
3. **Proxy-arkitektur:** Bygg en minimal proxy-tjeneste som kan håndtere RBAC-oversettelse og logging innen 3 måneder – før agentene settes i produksjon.