



Problemstilling: Agentisk KI - teknisk styring, kontroll og ...

Problemstilling: Agentisk KI – teknisk styring, kontroll og risiko for arkitektur og sikkerhet

Vi er en mellomstor offentlig virksomhet som vurderer å ta i bruk agentisk KI som kan utføre handlinger på tvers av interne systemer, API-er og datakilder. Agentene vil kunne initiere prosesser, endre data, foreslå beslutninger og samhandle med eksterne tjenester uten direkte menneskelig styring.

Tekniske begrensninger og kontekst:

- Infrastruktur består av en blanding av moderne API-tjenester og eldre fagsystemer uten fullverdig integrasjonsstøtte.
- IAM-plattformen støtter RBAC, men ikke finmasket policy-styring for autonome prosesser.
- Logging og sporbarhet er fragmentert på tvers av systemer.
- Sikkerhetsarkitekturen er designet for deterministiske systemer, ikke autonome agenter.
- Tidslinjen for pilot er kort (6–9 måneder), og sikkerhetsressurser er begrenset.

Uavklarte tekniske beslutninger:

- Hvordan definere og begrense agentenes handlingsrom på tvers av systemer.
- Hvordan etablere “policy-sandkasser” som hindrer uønskede eller irreversible handlinger.
- Hvordan sikre robust autentisering, autorisasjon og nøkkelrotasjon for KI-drevne prosesser.
- Hvordan designe logging, revisjonsspor og hendelseshåndtering når agentene tar autonome valg.
- Hvordan håndtere feilhandling, eskalering og rollback-mekanismer.
- Hvordan sikre at agentene ikke eksponerer sensitiv informasjon gjennom eksterne API-kall.

Hvilke arkitekturprinsipper bør ligge til grunn for å kontrollere autonome KI-agenter i komplekse systemlandskap.

- Hvilke sikkerhetsmekanismer (IAM-modeller, policy-motorer, sandboxing, rate-limiting, isolasjon) er nødvendige for å hindre uønskede handlinger.
- Hvilke tekniske blindsoner oppstår når agenter får tilgang til systemer som ikke er designet for autonom interaksjon.
- Hvilke risikoer knyttet til integrasjoner, dataflyt, eskalering og feilhandling undervurderes typisk.



- Hvordan bør vi designe en kontrollmodell som kombinerer autonomi, sikkerhet og etterprøvbarhet uten å skape flaskehalser.