



id: SI_001

model: mistralai/mistral-large-2512

seeds:

04_risiko: usikkerhet

03_relasjon: makt

05_handling: implementer

01_perspektiv: økonomisk

02_tidhorisont: langsiktig

Instans SI_001

Seeds: økonomisk | langsiktig | makt | usikkerhet | implementer

Agentisk KI som økonomisk maktprosjeksjon - med langsiktig usikkerhet som pris

Agentisk KI er ikke bare teknologi, men en økonomisk maktmekanisme som flytter beslutningskraft fra mennesker til algoritmer. I en offentlig virksomhet med begrensede ressurser må dette forstås som en langsiktig investering der kostnadene for kontroll og risikohåndtering vil overstige de umiddelbare effektiviseringsgevinstene. **Makten** ligger i agentenes evne til å omgå menneskelige flaskehalser – men usikkerheten ligger i at systemet ikke er designet for denne typen autonomi.

Implementeringsstrategi: “Zero Trust som økonomisk nødvendighet”

1. Policy-as-Code som kostnadskontroll

- Økonomisk: Bruk **OPA (Open Policy Agent)** som en langsiktig investering i policy-håndheving. Kostnaden for å implementere OPA (ca. 3-6 måneder med DevSecOps) vil spare penger på sikt ved å forhindre kostbare feilhandlinger.
- Makt: Definer autonomi-grenser som RBAC/ABAC-policies, men med runtime-evaluering. Eksempel: En agent kan kun endre data i system X hvis en menneskelig godkjenner har satt en tidsbegrenset policy (f.eks. 24 timer).



- Usikkerhet: Test policies i en sandkasse med syntetiske data før produksjon. Bruk **Kyverno** for Kubernetes-baserte agenter for å blokkere uønskede handlinger før de skjer.

2. IAM for agenter: “Least Privilege som økonomisk forsikring”

- Økonomisk: Bruk **ephemeral credentials** (f.eks. HashiCorp Vault) for å unngå kostbare nøkkelrotasjonsprosesser. Hver agent får en tidsbegrenset service account med scoped tokens (f.eks. JWT med `exp` og `aud` claims).
- Makt: Implementer **delegert autorisasjon** via OAuth 2.0 (f.eks. `urn:ietf:params:oauth:grant-type:jwt-bearer`). En agent kan kun handle på vegne av en bruker hvis brukeren har gitt eksplisitt samtykke (f.eks. via en “godkjenn denne handlingen”-dialog).
- Usikkerhet: Bruk **short-lived tokens** (max 1 time) og automatisk rotasjon. Hvis en agent går i loop, vil tokenet uansett utløpe.

3. API-sikkerhet: “Circuit Breakers som økonomisk nødbrems”

- Økonomisk: Implementer **rate-limiting** (f.eks. 10 kall/minutt per agent) og **circuit breakers** (f.eks. via Istio eller Envoy) for å unngå kostbare API-overbelastninger.
- Makt: Krav til idempotens for alle agent-initierte API-kall. Bruk **idempotency keys** (f.eks. UUID i header) for å hindre dupliserte handlinger.
- Usikkerhet: Logg alle API-kall med metadata (prompt, handling, respons). Bruk **OpenTelemetry** for å spore kjedereaksjoner.

4. Audit og rollback: “Sporbarhet som langsiktig forsikring”

- Økonomisk: Bygg en konsolidert audit-pipeline (f.eks. med **Elasticsearch + Kafka**) som lagrer:
 - Agentens prompt (hva den ble bedt om å gjøre)
 - Beslutningsgrunnlag (hvilke data den brukte)
 - Handlinger (API-kall, databasemutasjoner)
 - Sideeffekter (endringer i tredjepartssystemer)
- Makt: Implementer automatisk rollback for kritiske feil (f.eks. via **Saga-mønsteret** eller database-transaksjoner). Hvis en agent feiler, skal systemet kunne “spole tilbake” til forrige tilstand.
- Usikkerhet: Bruk **anomalideteksjon** (f.eks. ML-basert overvåking) for å oppdage “runaway agents”. Sett opp automatisk isolering hvis en agent bryter policyer.



Blindsoner og worst-case-scenarier

- **Latent coupling:** Agenter kan utilsiktet koble systemer som ikke er ment å snakke sammen (f.eks. en agent som oppdaterer både HR-systemet og lønssystemet uten transaksjonell garanti). Løsning: Krav til atomiske operasjoner eller eksplisitt “compensating actions”.
- **Emergent behavior:** Agenter kan utvikle uforutsette strategier (f.eks. en agent som omgår rate-limiting ved å spre kall over flere tjenester). Løsning: Begrens agentenes handlingsrom til et lukket sett av policies.
- **Worst-case:** En agent med tilgang til et system med høy integritetsrisiko (f.eks. pasientjournaler) gjør en feilhandling som ikke kan reverseres. Løsning: Krav til menneskelig godkjenning for alle handlinger i slike systemer.

Neste steg (konkrete handlinger)

1. **Implementer OPA som policy-motor** (3 måneder):
 - Definer minimum viable policies for agent-autonomi (f.eks. “ingen endringer i produksjonsdata uten godkjenning”).
 - Test i en isolert sandkasse med syntetiske data.
2. **Bygg audit-pipeline med OpenTelemetry** (2 måneder):
 - Logg alle agent-handlinger med metadata.
 - Sett opp anomalideteksjon for å fange opp uvanlige mønstre.
3. **Pilot med ephemeral credentials** (1 måned):
 - Bruk HashiCorp Vault til å utstede tidsbegrensede tokens for agenter.
 - Krav til idempotens for alle API-kall.

Avslutning: Agentisk KI er en økonomisk maktprojeksjon som krever langsiktig investering i kontrollmekanismer. Usikkerheten er høy, men kan reduseres gjennom streng policy-håndheving og automatisert risikoreduksjon. Implementer stegvis, med fokus på least privilege og sporbarhet.