



id: SI_005

model: openai/gpt-5.4-mini

seeds:

04_risiko: sårbarhet

03_relasjon: konkurranse

05_handling: vent

01_perspektiv: strategisk

02_tidhorisont: langsiktig

Instans SI_005

Seeds: strategisk | langsiktig | konkurranse | sårbarhet | vent

Den strategiske hovedregelen er: **ikke gi agentisk KI direkte makt over produksjonssystemer før dere har en kontrollsløyfe som er sterkere enn agentens handlingsrom**. I offentlige miljøer er den reelle sårbarheten ofte ikke modellen, men **koblingen mellom heterogene systemer, uklare API-kontrakter og svak sporbarhet**. Min litt mindre mainstream vurdering: start med **semi-autonomi** og bygg konkurransefordel gjennom sikker styring, ikke gjennom maksimal frihet.

Arkitekturprinsipper

- **Capability-first:** hver agent får et smalt, eksplisitt handlingssett per oppgave, ikke "tilgang til systemet".
- **Two-person rule for høy risiko:** alt som påvirker integritet, økonomi eller persondata krever enten godkjenning eller forhåndsdefinerte guardrails.
- **Policy-as-code før handling:** innfør OPA/Gatekeeper/Kyverno-lignende evaluering som hard stop i runtime. Uten sentral policy-motor bør dere bruke en **API-gateway + policy proxy** som midlertidig kontrollpunkt.
- **Execution graph:** modeller hvert agentløp som en DAG med noder for plan, beslutning, kall, respons, sideeffekt og kompensasjon. Hver node får korrelasjons-ID, tidsstempel og policy-resultat.

IAM og tilgang

- Bruk **ephemeral credentials** med TTL på 5–15 minutter.
- Scoped service accounts per agent og per use case.



- Ingen permanente nøkler; rotasjon automatisk ved hver deploy og ved avvik.
- Autorisasjon bør være **kontekstuell**: system, datasett, tidspunkt, risikonivå, miljø.

API-sikkerhet

Legacy-API-er uten idempotens er høy risiko. Beskytt med:

- rate limiting per agent og per mål-API
- circuit breakers ved feilrate > 5 % eller avvik i responsmønster
- “dry-run”/simulate først der det er mulig
- kompensasjonshandlinger definert før produksjon

Logging og hendelser

Bygg en **append-only audit-pipeline**: prompts, tool calls, beslutningsdata, policy-avslag, diff av endringer og sideeffekter. Distribuert logging uten sentralisert revisjonsløp er en blind sone; dette er et av de viktigste tiltakene på 6–9 måneders sikt.

Worst case

Modellér scenarier der agenten:

1. gjentar handlinger i loop,
2. eskalerer privilegier indirekte,
3. forårsaker latent coupling mellom systemer,
4. skaper inkonsistente data som ikke kan rulles tilbake.

Neste steg

1. Velg 1–2 laveksponeringsbrukstilfeller og definer eksplisitte autonomigrenser.
2. Etabler en policy-proxy og audit-strøm før første pilot.
3. Lag en risiko-matrise med “stop conditions” og kompensasjon per API/system.