



# Debrief — mistralai/mistral-large-2512

---

- Instanser: 9
  - Tokens inn: 20,957
  - Tokens ut: 13,666
  - Kostnad: \$0.2679
  - Kjørt: 2026-04-16 00:50:36.352003+00:00
- 

## Sverm-debrief: Agentisk KI i offentlig sektor

---

### Konsensus

1. **Maktfordeling er kjerneproblemet** – Agentisk KI flytter beslutningskraft fra mennesker til kode, og kontrollmekanismer må designes før piloten, ikke etter.
  2. **Policy-sandkasser er nødvendige** – Alle instanser anbefaler å begrense agentenes handlingsrom med dynamiske regler (f.eks. Open Policy Agent) og “negative lister” (hva agenten aldri får gjøre).
  3. **Reversible handlinger og logging er kritisk** – Alle endringer må kunne rulles tilbake, og agenthandlinger må logges uforanderlig med kontekst (input, beslutningsgrunnlag, output).
  4. **Eldre systemer er en blindsoner** – Legacy-integrasjoner må isoleres eller pakkes inn i proxy-lag for å hindre utilsiktede eskaleringer.
  5. **Menneskelig oversikt må bevares** – Selv autonome agenter trenger “break-glass”-mekanismer og eskalering til menneskelig godkjenning for høyrisiko-handlinger.
- 

### Dissens

1. **Kortsiktig vs. langsiktig kontroll**
  - Kortsiktige instanser (SI\_003, SI\_005) prioriterer raske løsninger med eksisterende IAM og proxy-tjenester.
  - Langsiktige instanser (SI\_001, SI\_006) krever ny arkitektur (f.eks. SPIFFE/SPIRE, immutabel logging) for å sikre skalerbarhet.



## 2. Tillit vs. teknisk kontroll

- Noen (SI\_008) mener tillit bygges gjennom transparens og juridiske rammer.
- Andre (SI\_007) advarer mot at tillit er illusorisk uten fysiske kontrollmekanismer (f.eks. "kill switches").

## 3. Autonomi-nivåer

- Optimistene (SI\_004) ser agentisk KI som et konkurransefortrinn og anbefaler å teste høy autonomi tidlig.
- Forsiktige (SI\_002) advarer mot å gi agenter reell makt før "fiasko-simuleringer" er gjennomført.

---

## Blindsoner avdekket

1. **Juridisk ansvarsvakuum** – Ingen instanser hadde fullstendig oversikt over hvem som er ansvarlig hvis en agent bryter personvernregler eller forårsaker økonomisk skade.
2. **Policy-drift** – Agenter vil tilpasse seg og bryte regler på uforutsette måter (f.eks. ved å kombinere tillatte handlinger for å oppnå forbudte resultater).
3. **Eksterne API-er som trojanere** – Agenter kan utilsiktet sende sensitive data til tredjeparter (f.eks. via oversettelsestjenester eller vær-APIer).
4. **Kulturell motstand** – Ansatte kan sabotere agenter hvis de føler seg overkjørt, men dette ble bare nevnt av én instans (SI\_004).
5. **Eskalering av feilhandlinger** – En liten feil (f.eks. en misforstått regel) kan eskalere til systemkollaps hvis agenten gjentar handlingen automatisk.

---

## Anbefalinger

### 1. Etabler en "Agent Governance Board" innen 2 uker

- Med representanter fra IT-sikkerhet, juridisk, fagsystemeiere og ledelse.
- Oppgave: Definere negative lister (hva agentene aldri får gjøre) og godkjenne alle nye agenthandlinger før implementering.

### 2. Kjør en "Chaos Engineering"-pilot på 4 uker

- Test en dummy-agent i et isolert miljø med syntetiske data.
- Mål: Identifisere uforutsette dataflyter, policy-brudd og eskaleringer.
- Bruk funnene til å justere policy-sandkasser (f.eks. Open Policy Agent).



### 3. Implementer “reversible autonomy” for én kritisk prosess

- Velg ett fagsystem (f.eks. saksbehandling) og bygg:
  - En proxy-tjeneste for eldre systemer (med rate-limiting og logging).
  - En immutable logg (f.eks. immudb) for alle agenthandlinger.
  - En rollback-mekanisme for dataendringer (innen 24 timer).

### 4. Forhandle inn juridiske klausuler i API-kontrakter

- Krev at eksterne leverandører:
  - Logger alle kall fra agentene.
  - Aksepterer delt ansvar for feilhandlinger.
  - Gir tilgang til revisjonslogger ved tvist.

### 5. Bygg en “makt-dashboard” for ledelsen

- Vis sanntidsdata om:
  - Agentenes handlingsfrihet vs. policy-brudd.
  - Eskaleringsrater (hvor ofte mennesker må gripe inn).
  - Reverseringsfrekvens (hvor ofte handlinger må rulles tilbake).
- Mål: Sikre transparens og tillit på tvers av organisasjonen.

---

**Prioritering:** Start med punkt 1 og 2 umiddelbart – de krever minst teknisk innsats, men gir størst risikoreduksjon. Punkt 3–5 kan fases inn under piloten.