



Liikenne- ja viestintävirasto

Jussi Eronen

Jussi Eronen

Varjo-AI
SaaS
BYOD
Varjo-IT

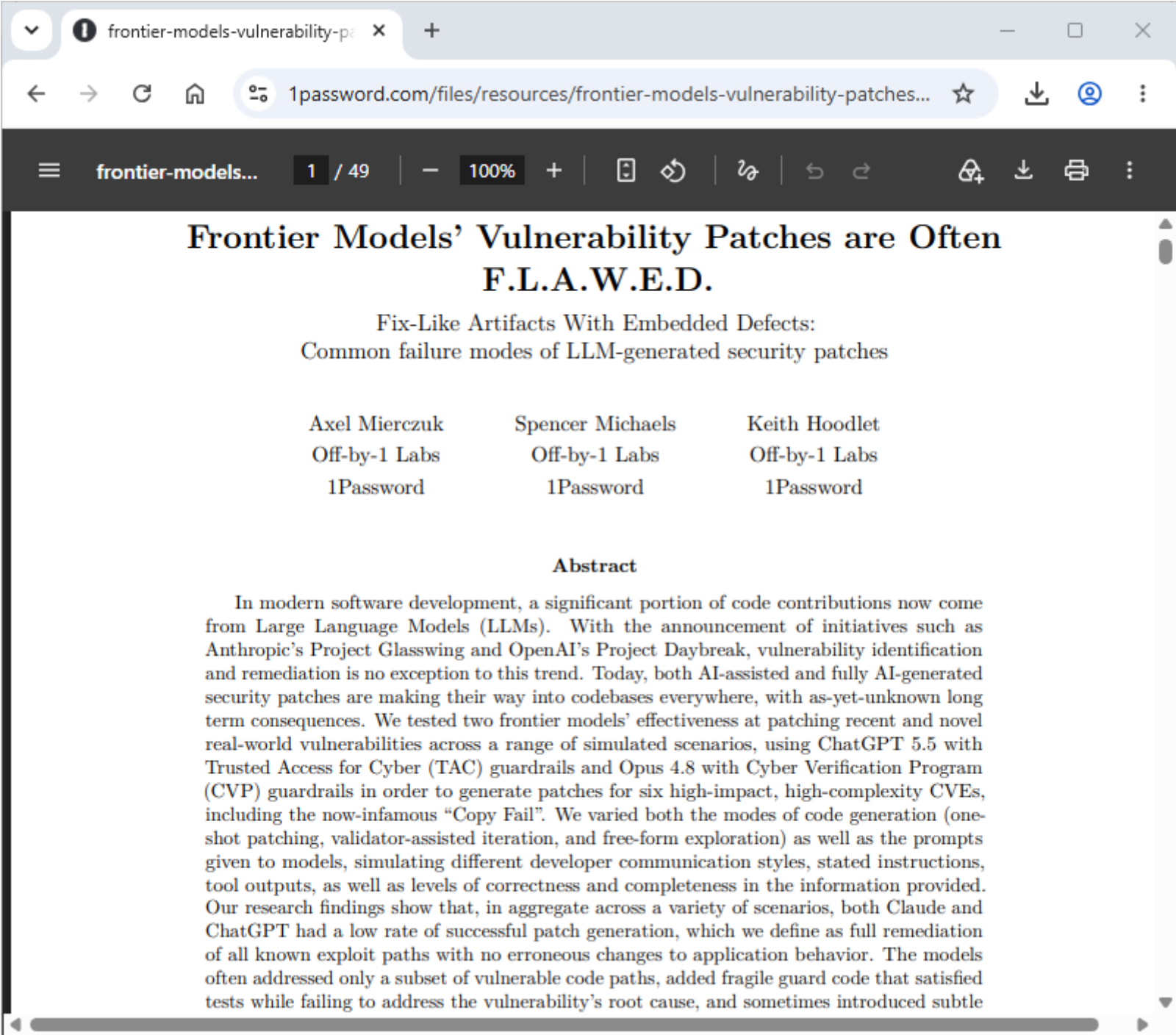


Code is a liability (not an asset)

AI psychosis, tech boss edition.

Cory Doctorow [Follow](#) 12 min read · Jan 6, 2026





Tekoälyagenttien kyberturvallisuus



Tekoälyn soveltamisen kyberturvallisuus ja riskienhallinta

The security threat of
AI-enabled cyberattacks



Applying artificial intelligence in
Cybersecurity



Tekoälyagenttien kyberturvallisuus

Ohjeistus tekoälyagenttien turvalliseen suunnitteluun ja toteutukseen

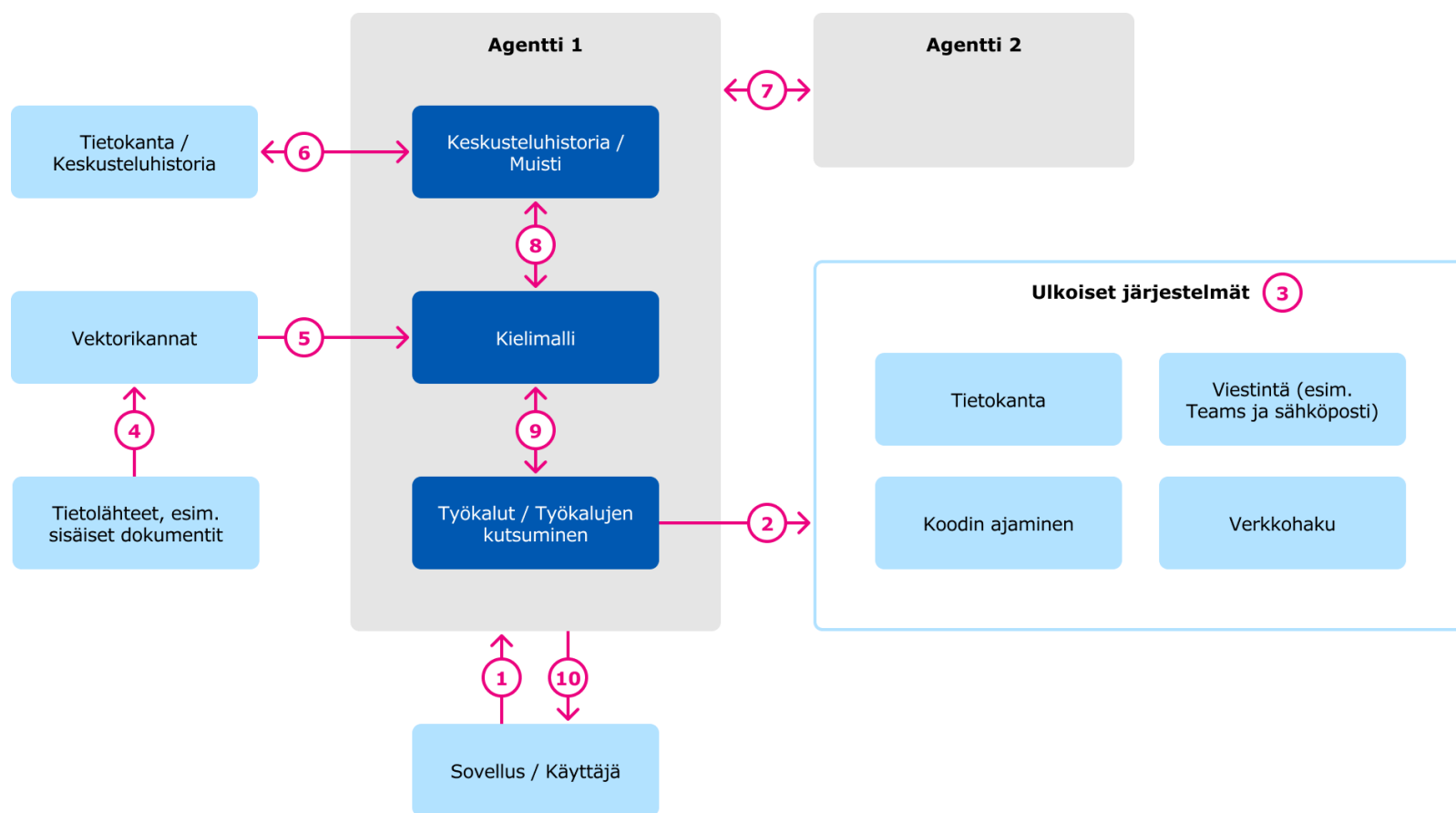
Traficom
Huoltovarmuuskus

Traficom Research Reports
07/2024

Traficom julkaisu
5/2026

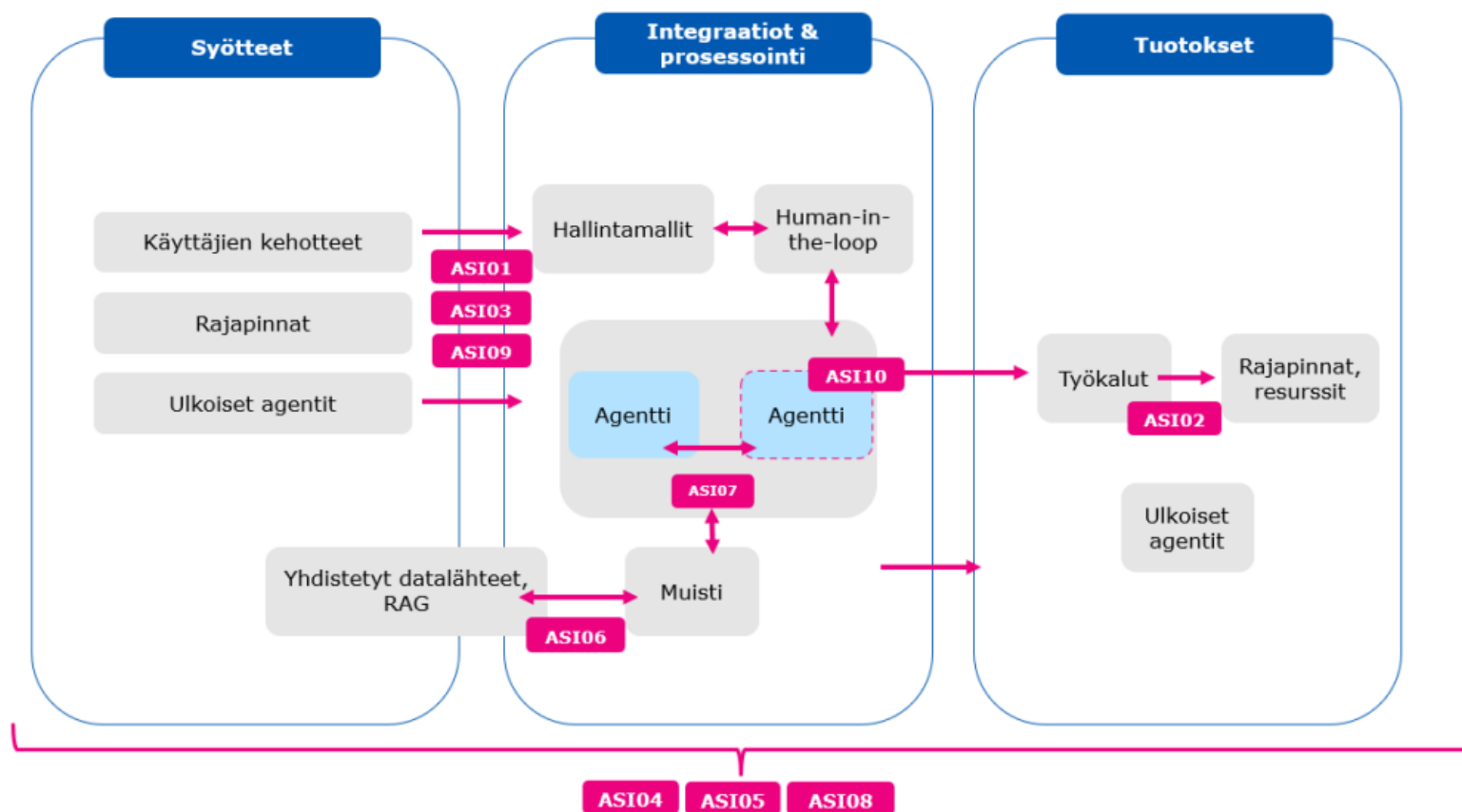
- Tekoälyagenttien kyberturvallisuus on Huoltovarmuuskuskeskuksen kanssa yhteistyössä toteutettu projekti, kohderyhmänä huoltovarmuuskriittiset yritykset.
- Jatkoa aiempiin tekoälyselvityksiin ja vastaa osaltaan tekoälyn yhteiskunnalliseen vaikutukseen ja nopeaan kehitykseen.

Tekoälyjärjestelmien haavoittuvuudet



- 1** Kehoteinjektio, rajoittamaton kulutus
- 2** Liiallinen toimivalta
- 3** Toimitusketjun haavoittuvuus
- 4** Vektorikannan saastuttaminen varmentamattomilla dokumenteilla
- 5** Kielimallin käytöksen muutos
- 6** Muistin saastuttaminen
- 7** Väärän tiedon jakaminen agenttien kesken, kapinoiva tekoälyagentti
- 8** Muistin saastuttaminen, tietosuojaa
- 9** Väärän työkalun kutsuminen
- 10** Hallusinaatio, virheellinen tieto, järjestelmäkehötteen vuotaminen, arkaluontoisen tiedon paljastuminen, virheellinen tuotoksen käsittely

Agenttisten järjestelmiin erityisesti kohdistuvat haavoittuvuudet



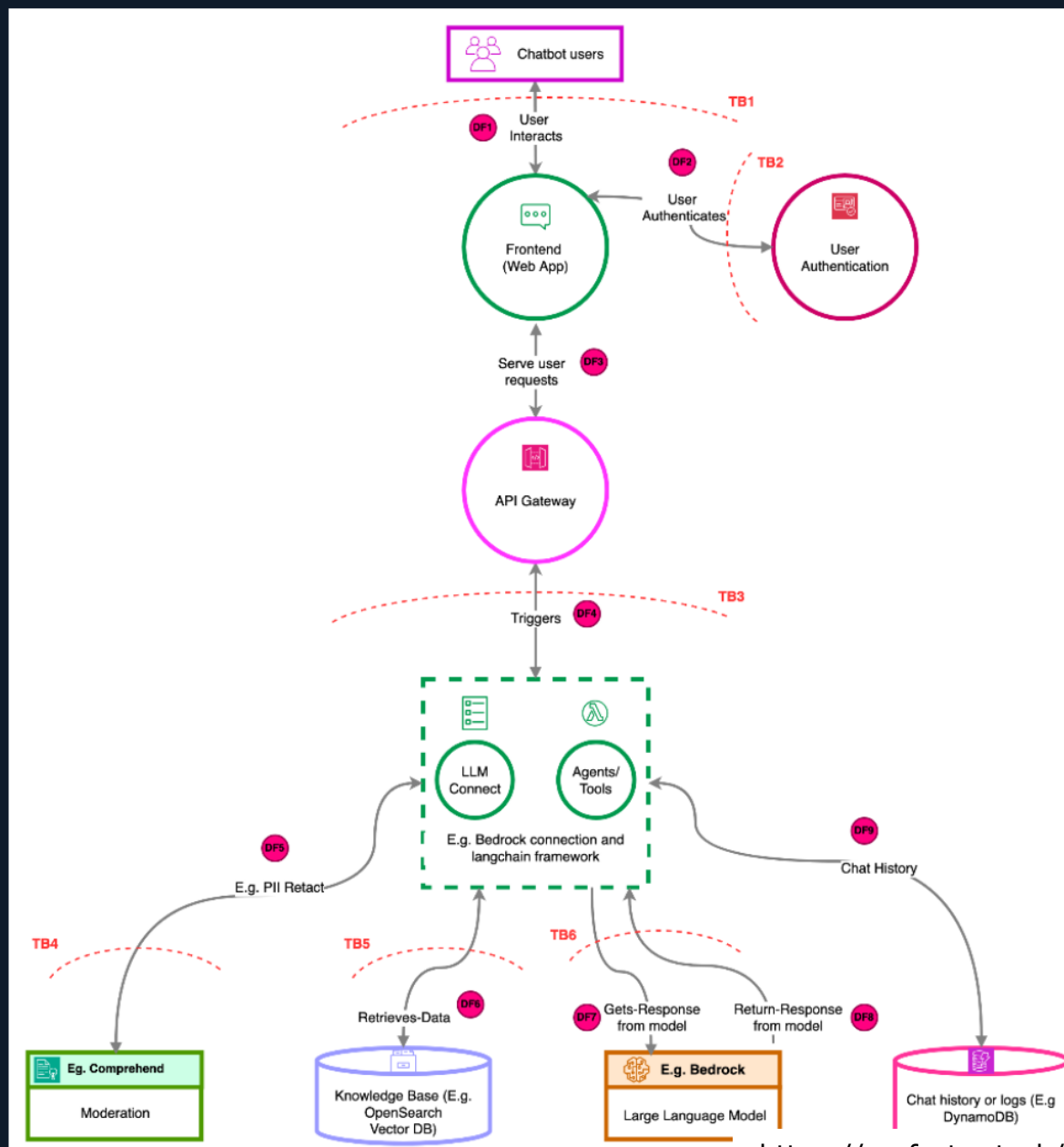
- **ASI01:** Agentin tavoitteen manipulointi
- **ASI02:** Työkalujen väärinkäyttö
- **ASI03:** Identiteettien ja oikeuksien väärinkäyttö
- **ASI04:** Toimitusketjuriskit
- **ASI05:** Koodin etäsuorittaminen
- **ASI06:** Muistin ja kontekstin manipulointi
- **ASI07:** Turvaton agenttien välinen viestintä
- **ASI08:** Ketjureaktiot
- **ASI09:** Ihmisen ja agentin välisen luottamuksen hyväksikäyttö
- **ASI10:** Villiintyneet agentit

Riskitason arviointi suunnitteluvaiheessa

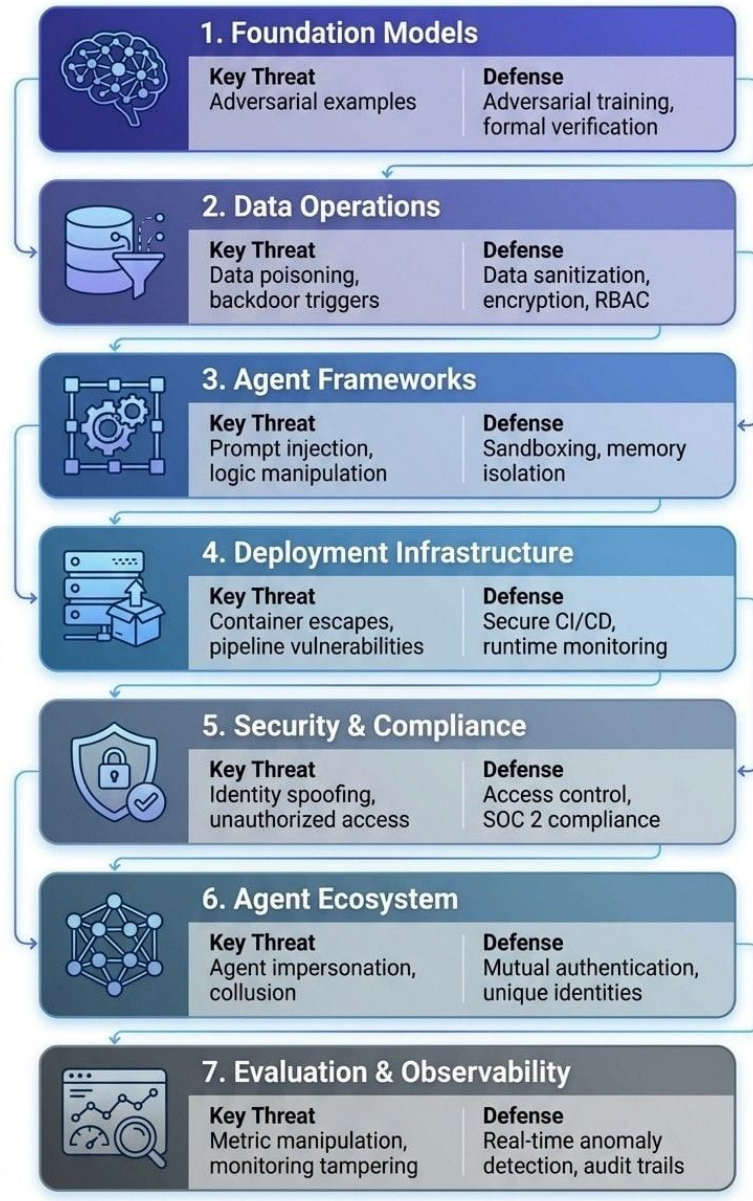
	Matala riskitaso	Kohtalainen riskitaso	Korkea riskitaso	Erittäin korkea riskitaso
Järjestelmän vaikutus ydinliiketoimintaan	Järjestelmä ei suorita ydinliiketoimintaan liittyviä tehtäviä.	Järjestelmä suorittaa helposti hallittavia ja rajoitettuja osia työkuluista, jotka ovat osa ydinliiketoimintaa.	Järjestelmä suorittaa osia organisaation ydinliiketoiminnasta.	Järjestelmä suorittaa keskeisiä osia organisaation ydinliiketoiminnasta.
Järjestelmän autonomia ja monimutkaisuus	Järjestelmä ei ole integroitu muihin järjestelmiin tai toisiin agentteihin.	Järjestelmä on integroitu vähäiseen määrään ulkoisia järjestelmiä, jotka eivät ole ydinliiketoiminnan kannalta kriittisiä.	Järjestelmä on integroitu useisiin ulkoisiin työkaluihin ja/tai on moniagenttijärjestelmä. Osa integraatioista kohdistuu ydinliiketoiminnan kannalta kriittisiin järjestelmiin.	Moniagenttijärjestelmä, jossa on useita rajapintoja ja integraatioita ydinliiketoiminnan kannalta kriittisiin järjestelmiin.
Järjestelmän käyttämän datan arkaluontoisuus	Järjestelmä käsittelee vain julkista tietoa.	Järjestelmä käsittelee organisaation sisäistä tietoa.	Järjestelmä käsittelee luottamuksellista tietoa.	Järjestelmä käsittelee salaista ja arkaluontoista tietoa.

<https://aws.amazon.com/blogs/security/threat-modeling-your-generative-ai-workload-to-evaluate-security-risk/>

Dataflow Diagram



MAESTRO Framework: 7-Layer Architecture for AI Agent Security



<https://prefactor.tech/blog/maestro-framework-threat-modeling-for-ai-agents>

TRAFICOM

- Kriittisen infrastruktuurin turvallisuussäntely
- Hoitotyön eettiset periaatteet ja lääkinnällisten laitteiden säntely
- Demokraattisilta prosesseilta vaadittu läpinäkyvyys ja lainmukaisuus (ml. hallintolain vaatimukset automaattisesta päätöksenteosta)

KISS
Tilannekuva riskeistä
Vastuu
Osaaminen
Osto-osaaminen

WALT DISNEY

AKU ANKKA

Ilman tokeneita



Kiitos