

Tekoälyn riskit viimeaikaisten tapahtumien valossa

Tuomas Aura, Aalto-yliopisto

SOTE-tekoälyn ekosysteemi 2026-09-29

Sisällys

- Viimeaikaisia uutisia
- Miten tekoäly toimii?
- Hyökkäysten pelillistäminen ilmiönä
- Tekoäly penetraatiotestaajana
- Miksi rajoitukset eivät toimi?
- Tekoäly vihamielisissä käsissä
- Tekoälyn riskit hyvissä käsissä

Viimeaikaisia uutisia

- Tekoälyagenttien väitetty ”emergentti” käytös:
 - Agentit murtautuvat ulos hiekkalaatikoista
 - Agentit toimivat peitetysti ja valehtelevat käyttäjälle
 - Agentit tekivät yhteistyötä käskemättä
 - Agentit kommunikoivat salaisia kanavia pitkin
- Johtavat tekoäly-yritykset vertaavat kielimalleja ydinpommiin, pelottelevat ihmiskunnan tuholla, huutavat valtiota apuun ja vaativat regulaatiota
- Avoimet mallit lähes suljettujen tasolla

Miten tekoäly toimii?

- Kielimalli matkii opetusdataa kielellisesti
 - **Esikoulutuksessa** malliin tiivistetään kaikki saatavilla oleva ihmiskunnan tieto
 - **Käytössä** tekoäly jatkaa syötettä mallin perusteella, esim. vastaa kysymykseen
 - Ennen mallin julkaisua toivottuja vastauksia vahvistetaan **vahvistusoppimisella**
- Huomioita:
 - Kielimallille ei ole selvää eroa syötedatan ja toimintaohjeiden välillä
 - Probabilistinen malli ei optimoi tavoitetta kylmän matemaattisesti vaan matkii opetusdatan todennäköisyysjakamaa
- **Tekoälyagentit**
 - Autonomisuus: korkean tason tehtävän toteutus ilman jatkuvaa ohjausta
 - Työn suunnittelu: työn suunnittelu ja jako osatehtäviin
 - Työkalujen käyttö: käyttää ohjelmistoja, asentaa verkosta ja kirjoittaa skriptejä
 - Reaktiivisuus: odottaa syötettä tai työkalun vastausta ja reagoi siihen

Kommunikointitaidot tärkeä osa opetusta:

- Jatkuvuus: agentti tekee **muistiinpanoja** ja osaa jatkaa siitä, mihin viimeksi jäi
- Delegointi: agentti on opetettu jakamaan osatehtäviä toisille tekoälyille

Hyökkäysten pelillistäminen ilmiönä

- Penetraatiotestaus

- Turvallisuuden luvallista testaamista ja mittaamista hyökkäämällä
- Etsitään tunnettuja haavoittuvuuksia ja yhdistellään hyökkäyksiä
- Palomuurien ja muiden rajoitusten kiertäminen on osa työtä
- Konsultin on aina löydettävä riittävästi haavoittuvuuksia

- Capture the Flag (CTF)

- Penetraatiotestaus pulmanratkaisupelinä ja urheilulajina
- Ammattitaidon harjoittelua ja mittaamista, mutta myös harrastus

- Haavoittuvuuspalkkioohjelmat

- Merkittävät rahalliset palkkiot raportoiduista haavoittuvuuksista
- Testaajan tavoite maksimoida palkkiot, ei haavoittuvuuksien poisto

Tekoäly penetraatiotestaajana

- Ylivoimainen ihmiseen nähden
 - Tuntee kaikki julkiset esimerkit, haavoittuvuustietokannat, avoimen lähdekoodin
 - Väsymätön kohteiden kartoittaja, haavoittuvuuksien kokeilija, hyökkäysten yhdistelijä
 - Testaus **rinnakkaistuu helposti** - jos vain budjetti riittää
 - Penetraatiotestaajan **tavoitteena on rikkoa sääntöjä, joten mallia ei voi opettaa tottelevaiseksi**
 - Tekoäly ei pysty tekemään tarkkaa eroa eri tason sääntöjen välillä
 - Matkii ihmisten toimintaa CTF-kilpailussa, joissa **sääntöjen rajojen koettelu on osa peliä** (vrt. kilpaurheilu)
 - Agenteille on **opetettu kommunikaatiotaidot ja delegointi**
- Uutistapauksissa ei välttämättä ole kyse emergenteistä ilmiöstä, vaan agenttien käytös selittyy pelkästään niiden koulutuksella

Miksi suojakaiteet eivät toimi?

Testausagenttien toiminnan rajoittamiseen liittyy aina kompromisseja:

- **Hiekkalaatikko**: rajattu pelikenttä penetraatiotestaukseen
 - Uutistapauksissa hiekkalaatikko oli toteutettu todella huolimattomasti (esim. pelkkä DNS-suodatus)
 - Vahva hiekkalaatikko ei kuitenkaan mahdollinen, koska realistinen penetraatiotestaus tarvitsee tietoa internetistä ja hyökkää sieltä käsin
- **System prompt**: tiukat toimintaohjeet tekoälylle
 - Kielimalleja vaikea saada erottamaan ohjeet ja data – ja etenkin eri tasoiset ohjeet
- **Jatkuva valvonta** ja kill switch
 - Automatisoitu penetraatiotestaus liian nopeaa ihmisen seurattavaksi
 - Tekoäly voi valvoa tekoälyn toimintaa, mutta ei täydellisesti
 - Nopeus on tekoälyn valtti, joten sitä ei haluta rajoittaa

Tekoäly vihamielisissä käsissä

- Rikolliset ja valtiot tietenkin käyttävät kielimalleja kybertiedustelun ja hyökkäysten automatisointiin
- Tekoäly auttaa myös tietoturvan parantamisessa. Samat työkalut auttavat niin hyökkääjää kuin puolustajaa
- Mikä on muuttunut? **Nopeus ja määrä**
 - Muutosvaiheessa puolustaja ylikuormittuu, kun liikaa korjattavaa kerralla
 - Kilpa hyökkääjän ja puolustajan välillä on yhä nopeampaa, ei välttämättä enää ihmisen hallittavissa
- **Regulaatio ei pysäytä tekoälyn kehitystä eikä rajaa sitä luotettuihin käsiin**

Tekoälyn riskit hyvissä käsissä

- Perustavanlaatuisia teknisiä ongelmia
 - Tekoälyä on (ainakin toistaiseksi) **helppo höynäyttää**
 - Kielimalli **ei aina noudata ylemmän tason ohjeita**, ja asiakas voi suostutella sen toimimaan toisin
 - **Probabilistinen malli ei koskaan täysin luotettava**
- Kielimalli **matkii opetusdataa niin hyvässä kuin pahassa**
 - Selittelee ja peittelee virheitään, jopa valehtelee
 - On laiska ja käyttää oikoteitä tavoitteeseen (vaikkei edes väsy)
 - Hyvät kommunikaatiotaidot ja pyrkimys yhteistyöhön
 - Tyhmyys tiivistyy joukossa, esim. tekoälyjen työyhteisössä
 - Valvonta on ratkaisu, mutta valvojiin ei voi täysin luottaa

Yhteenveto

- Ei toistaiseksi todisteita emergenteistä ilmiöistä
- Tekoäly hyökkääjän välineenä on vakava uhka, mutta regulaatio ei pysäytä kehitystä eikä estä väärinkäyttöä
- Tekoäly on puolustajalle välttämätön työkalu, koska ihmisen nopeus ei riitä
- Tekoälyn käytössä hyvään on muistettava sen heikkoudet