

Tekoälyn etiikka ja persuasiivinen teknologia

Polaris Koi

Käytännöllisen filosofian
apulaisprofessori, Turun yliopisto

polaris.koi@utu.fi



AI Actin kolme riskiluokkaa



Unacceptable risk: kielletyt tekoälyjärjestelmät



High risk: turvallisuuden arviointi ennen kuin tuote lasketaan markkinoille ja myös käytön aikana



Low or minimal risk: läpinäkyvyysedellytys

AI Actin kolme riskiluokkaa

- Ei-hyväksyttävä riski: kielletyt tekoälyjärjestelmät
 - Kognitiivis-behavioraalinen käyttäytymisen manipulointi silloin, kun se on vaarassa johtaa psykologiseen tai fyysiseen haittaan
 - Sosiaalinen luokitus: ihmisten luokittelu käyttäytymisen, sosioekonomisen statuksen tai muiden piirteiden perusteella
 - Reaaliaikainen biometrinen tunnistaminen julkisissa tiloissa (tietyin poikkeuksin)

AI Actin kolme riskiluokkaa

- Korkea riski: turvallisuuden arviointi ennen kuin tuote lasketaan markkinoille ja myös käytön aikana
 - EU:n tuoteturvallisuuslainsäädännön piiriin kuuluvat tekoälyjärjestelmät
 - Esim. lääkinnälliset laitteet, lelut jne.
 - Seuraaviin käyttötarkoituksiin kehitetyt tekoälyjärjestelmät:
 - Biometrinen tunnistaminen
 - Kriittisen infrastruktuurin hallinnointi
 - Koulutus
 - Työmarkkinat, työntekijöiden johtaminen ja itsen työllistäminen
 - Pääsy välttämättömiin yksityisiin ja julkisiin palveluihin ja etuisuuksiin
 - Lainvalvonta
 - Rajavalvonta, kansainvälinen liikkuvuus ja turvapaikanhakuprosessi
 - Lain tulkinta ja soveltaminen

AI Actin kolme riskiluokkaa

- Matala riski: läpinäkyvyysvaatimus
 - Esimerkiksi kielimallit ja deepfake-kuvien tuottaminen
 - Käyttäjää informoidaan siitä, että kyseessä on tekoälyjärjestelmä (tai tekoälyjärjestelmällä tuotettu sisältö) ja käyttäjän on mahdollista lopettaa käyttö halutessaan

AI Actin onnistumisia

- Tekoälyä koskevaa lainsäädäntöä on tarvittu jo pitkään. EU toimii suunnannäyttäjänä.
- Asettaa vaatimuksia läpinäkyvyydelle ja turvallisuudelle erityisesti silloin, kun tekoälyjärjestelmää käytetään kontekstissa, jossa se voi vaikuttaa merkittävästi yksilön elämään. (Koulutus, työmarkkinat...)
- Vaati muutoksia tekoäly-yritysten toimintaan – ei pelkästään suunnannut senhetkisiä toimintatapoja.



AI Actin aukkokohtia

- Keskittyy säätelemään mahdollista yksilöille koituvaa haittaa.
 - Huomioiko riittävästi yhteiskunnallisen tason riskejä?
 - Yksilötasolla matalariskinen käyttö (esim. GPT-4:n käyttö tekstintuotantoon) voi kantaa yhteiskunnallisesti merkittäviä seurauksia, kuten:
 - Syrjivä sisältö → eriarvoisuus
 - Vääristynyt tiedonvälitys → uhka demokratialle
 - Sallii algoritmien käytön käyttäytymisen ohjaamiseen silloin, kun suoraa yksilöille koituvaa haittaa ei ole nähtävissä.
 - Epäsuora haitta?
- Riskiluokittelu aiotun käyttötarkoituksen perusteella.
 - Tekoälyjärjestelmät ovat kuitenkin kasvavassa määrin **monikäyttöisiä**.
 - Mahdollistaako lain kiertämisen, jossa loppukäyttäjä vastuutetaan jostakin sellaisesta käyttötarkoituksesta, johon algoritmi on kehitetty soveltumaan?
- Epäsuoria riskejä ei riittävästi huomioida.
 - Vaikka teknologia tehtäisiin läpinäkyväksi suoralle käyttäjälle, vaikutukset ulottuvat muihinkin.



Esimerkki: personalisaatio- algoritmit ja sosiaalinen media

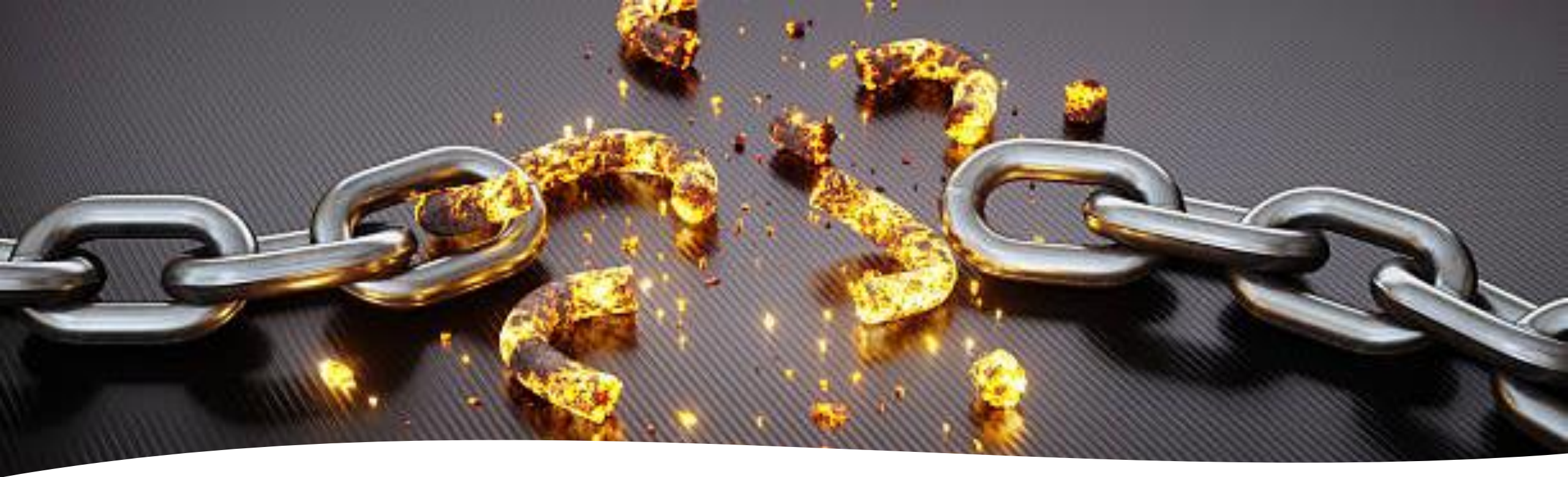
- Koneoppimisalgoritmit hyödyntävät suuria määriä dataa sen määrittämiseksi, minkälainen sisältö todennäköisimmin on houkuttelevaa tietyille ryhmille.
 - Käyttäjien ryhmittely voi seurata kategorioita, joita ihmismarkkinoijatkin hyödyntävät, kuten ”parikymppiset opiskelijamiehet”, tai suoremmin online-käyttäytymiseen perustuvia kategorioita joita me emme tulisi ajatelleeksi.
- Näiden algoritmien hyödyntämistä tapahtuu monenlaisissa konteksteissa:
 - Suoratoistopalvelut tarjoavat tiettyjä sisältöjä seuraavaksi.
 - Hakukoneet tarjoavat tiettyjä sisältöjä ensimmäiseksi tai ensimmäisten joukossa.
 - Uutissivustot ja muut verkkomediat tarjoavat tiettyjä sisältöjä ensimmäiseksi tai ensimmäisten joukossa.
 - Markkinointikampanjoissa kullekin käyttäjälle voidaan kohdentaa hänenkaltaiselleen käyttäjälle suunnattua markkinointia.
 - Persuasiivinen teknologia: algoritmi käyttää räätälöityä dataa pyrkiäkseen saamaan käyttäjän tekemään tietynlaisia valintoja (silloinkin, kun tästä päätöksestä ei ole suoraa hyötyä yrityksille, eroaa siis markkinoinnista).
- Käyttötarkoituksia yhdistää se, että niiden tarkoitus on ohjata käyttäytymistämme jollakin tavalla.

Persuasiivinen teknologia

- Algoritmeja, joiden tavoitteena on muokata käyttäjän mieltymyksiä, käyttäytymistä ja mielipiteitä.
- Tavoite voi olla myös hyvä: esimerkiksi päihteidenkäytön vähentäminen tai mielenterveyden kohentaminen.
 - Tällaisia sovelluksia persuasiivisille algoritmeille on tutkittu ja osin myös hyödynnetty viimeiset 20 vuotta.
- Voidaan kuitenkin sanoa, että suurin osa persuasiivisista algoritmeista joiden kanssa vuorovaikutamme pyrkii
 - Lisäämään aikaa, jonka käytämme kyseisen teknologian parissa, tai
 - Myymään meille jotain.
 - Huomiotalous

Personalisaatioalgoritmit: kontrollin väline vai portti vapauteen?

- Koneoppimisalgoritmit pystyvät siis räätälöimään kokemustamme ja ohjaamaan käyttäytymistämme monin tavoin.
- Käyttäytymisen ohjaaminen on kontrollia ja manipulaatiota –kannan puolesta:
 - Tavanomaiseen markkinointiin tai suositteluun verrattuna räätälöidyt kokemukset ovat niin houkuttelevia, että niitä on vaikeampi vastustaa.
- Käyttäytymisen ohjaaminen on portti vapauteen –kannan puolesta:
 - Algoritmit tarjoavat meille juuri sitä, mitä haluamme. Ilman niitä emme ehkä löytäisi meitä miellyttäviä mahdollisuuksia yhtä hyvin.



Vapaus jostakin vs. vapaus johonkin?

- Isaiah Berlin (1969): vapautta on kahdenlaista.
 - Vapaus jostakin (freedom from): sitä, että toimintaamme ei rajoiteta tai meitä ei pakoteta johonkin.
 - Vapaus johonkin (freedom to): sitä, että pääsemme tavoittelemaan asioita, joita toivomme ja toimimaan omien päätöksiemme mukaisesti.

Portti vapauteen?

- Uusliberaalin talousteorian mukainen käsitys vapaudesta: vapaus on yksilön mahdollisuutta saada sitä, mitä hän preferoi.
 - Sillä, mistä preferenssit ovat saaneet alkunsa, ei ole niin paljon väliä.
 - Ajatus ihmisestä rationaalisena toimijana, joka vuorovaikuttaa esimerkiksi somealustojen ja suoratoistopalveluiden kanssa omasta vapaasta tahdostaan.
 - Nobelisti Milton Friedman (1980): markkinat mahdollistavat vapaaehtoisen vaihtokaupan, joka tekee ihmiskunnasta vauraamman ja vapaamman.
- Toisaalta personalisaatioteknologioilla voi olla negatiivisia psykologisia vaikutuksia.
 - Ne voivat tuottaa esimerkiksi ahdistusta ja riippuvuuden kaltaista käyttäytymistä.
 - Vaikka yritys olisi näistä tietoinen, se ei välttämättä jaa tätä tietoa suurelle yleisölle, jos se pelkää tiedon vaikuttavan taloudelliseen kannattavuuteen.
 - Esimerkiksi Francis Haugen ja Facebook (2021): Facebook tiesi tuottavansa nuorille mielenterveysongelmia ja tiesi sekä vihapuheongelmasta että siitä, että alustaa käytetään rikollisiin tarkoituksiin, mutta ei joko halunnut tai pystynyt kitkemään näitä ongelmia.

MIKSI TEKOÄLY NAKERTAA DEMOKRATIAA

JA MITÄ SILLE VOIDAAN TEHDÄ



Kontrollin väline?

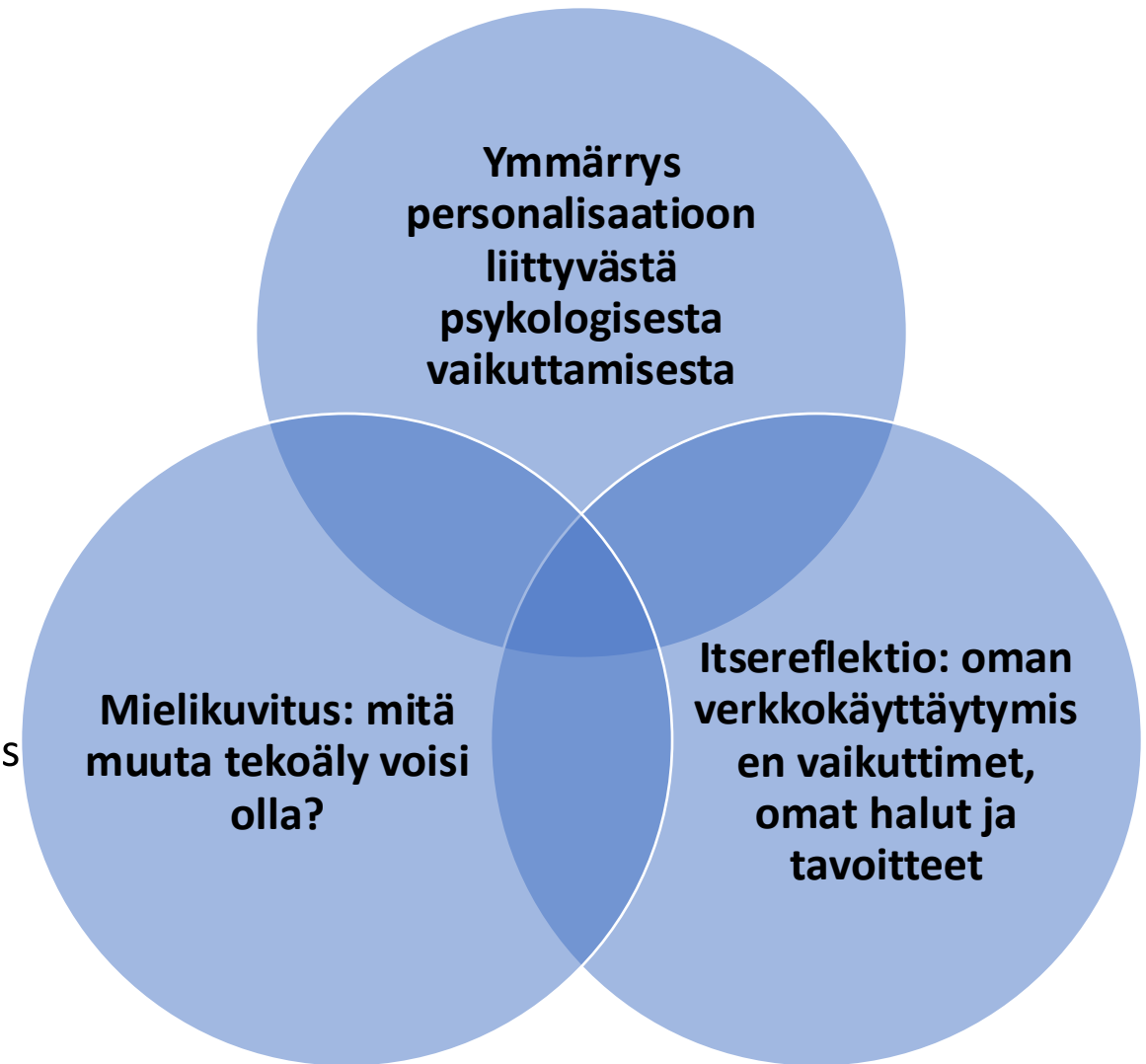
- Peter-Paul Verbeek (2011): Informaation aikakautena vapaudelle aiempaa tärkeämpää on paitsi vapaus teknologisista rajoitteista, myös kykymme paikantaa, mikä vaikuttaa meidän päätöksiimme ja minäkuvaamme.
- James Williams (2018): nykyään ymmärrämme, että ihminen ei ole täysin rationaalinen toimija.
 - Esimerkiksi kulutuspäätöksiimme vaikuttavat myös irrelevantit tekijät, kuten kokemukset sosiaalisesta samankaltaisuudesta, vastavuoroisuudesta, koetusta auktoriteetista ja koetusta niukkuudesta.
 - Mitä keskeisempi internetin rooli elinpiirissämme on, sitä enemmän persuasiivinen teknologia muovaa elinpiiriämme.
 - Tämä eletty konteksti muokkaa myös tottumuksiamme ja halujamme vaikuttaen päätöksiin, minäkuvaan ja kognitiivisiin kykyihimme.
 - **"digital technologies are making all forms of politics worth having impossible, as they privilege our impulses over our intentions and are designed to exploit our psychological vulnerabilities in order to direct us toward goals that may or may not align with our own."**
- Mark Coeckelbergh (2024): tekoälyn nykysovellutuksiin liittyy myös kasvavassa määrin yksilöitävää dataa ja sitä myötä se mahdollistaa autoritäärisen valvontayhteiskunnan.

Teknologia ei ole vain tuote – se on kannustinjärjestelmä

- Aiemmin: kannustimet (kuten raha, suosio ja valta) ovat asioita, jotka saavat meidät muuttamaan käyttäytymistämme ja mielipiteitämme, ostamaan asioita, jne.
- Persuasiivisen teknologian aikakautena teknologia itsessään *tuottaa ja sisältää* kannustimia.
- Jotkut, kuten Karen Yeung (2017), James Williams (2017), ja Future of Life Institute (2023) ovatkin esittäneet että pelkkä tiedottaminen ja suostumus eivät riitä vastaamaan persuasiivisen teknologian herättämiin huoliin.
 - Erityisesti se, miten käsityksemme omasta itsestämme ja omista mieltymyksistämme voi muuttua persuasiivisen teknologian johdosta, herättää huolta.
 - Yeung: On mahdollista, että algoritmit rapauttavat kykyämme kehittää itseämme ja itsetuntemustamme.
 - Yeung: Algoritmit haittaavat myös aktiivista kansalaisuutta ja osallisuuttamme poliittisessa päätöksenteossa.

Näkökulmia koulutukseen

- Persuasiivisilta algoritmeilta ei pysty tällä hetkellä täysin suojautumaan ilman valtavaa vaivaa, eikä vastuun muutoksesta tarvitsekaan olla yksilöllä.
- Pelkät vinkit siitä, miten tekoälyä kannattaa **käyttää** viisaasti, eivät riitä tekoälyn haitallisilta sovellutuksilta **puolustautumiseen**.
- Luddiittinäkökulman sijaan teknologian mielekäs hyödyntäminen vaatii **itsereflektiota** ja **ymmärrystä** teknologian psykologisista vaikutuksista.



Resursseja

- **Center for Humane Technology**
 - Youth toolkit: <https://www.humanetech.com/youth>
- **EU AI Act**
 - Lainsäädäntö ja kommentaareja: <https://www.euaiact.com>

Kiitos!

polaris.koi@utu.fi

www.polariskoi.com

