# Tekoälyn eettiset riskit ja vastuullinen suunnittelu

Tomi Kokkonen

Robophilosophy, AI Ethics, and Datafication Research Group (RADAR)

Käytännöllinen filosofia, Helsingin yliopisto

Humanismin päivän seminaari 25.11.2023: Ihminen teknistyvässä yhteiskunnassa

# Robophilosophy, AI Ethics and Datafication Research Group

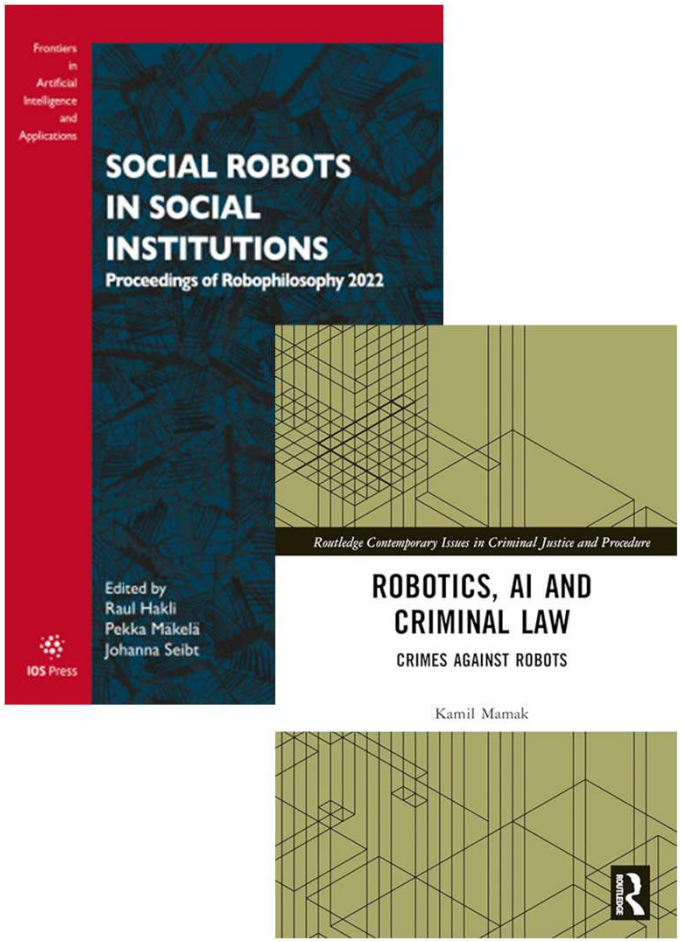
Monitieteinen tutkimusryhmä

- filosofia, tietojenkäsittelytiede, oikeustiede
- 3 vanhempaa tutkijaa / professoria
- 4 tutkijatohtoria
- 2 väitöskirjatutkijaa

Uusien teknologioiden vaikutukset yhteiskuntaan ja sosiaalisiin käytäntöihin

- robotiikka, tekoäly, koneoppiminen, automaatio, datafikaatio ja digitalisaatio
- filosofia: käsitteelliset, metafyysiset ja eettiset kysymykset

http://radar.cs.helsinki.fi



# Tekoäly, koneoppiminen ja syväoppiminen

**Tekoäly** = kone suorittaa tehtävän, jossa ihminen tarvitsisi älykkyyttä

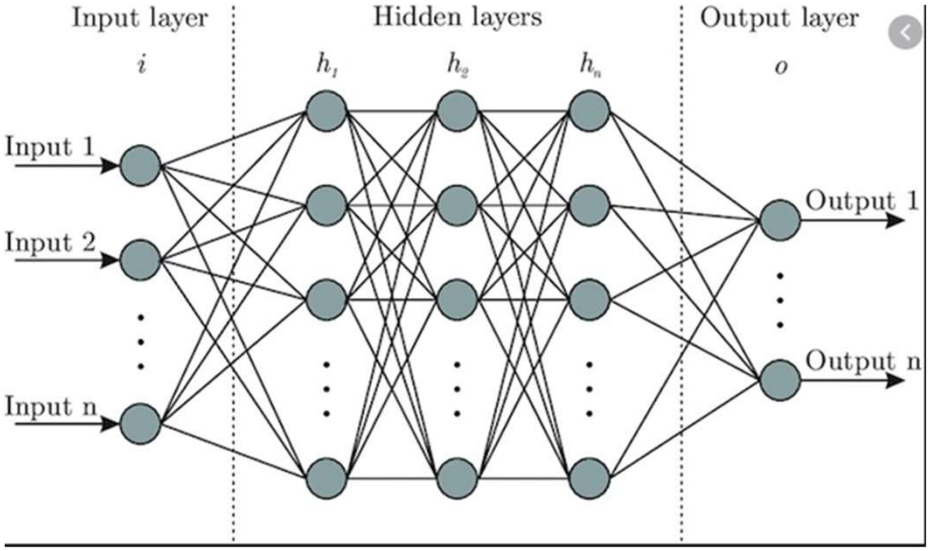
- mahdollistaa älykkyyttä vaativien toimintojen automatisoitumisen

**Koneoppiminen** = algoritmi muuttaa toimintaansa palautteen perusteella

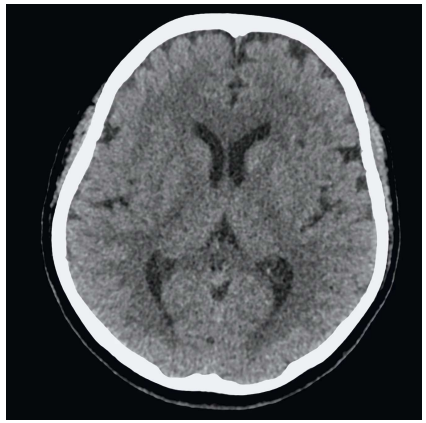
- luo assosiaatioita, ei tee päättelyjä
- behavioristista oppimista

**Syväoppiminen** = assosiaatoiden rakenteisuus

- monimutkaisempien kategorioiden oppiminen



# Analyysi, päätöksenteko ja jäljittely



Työkalu datan analyysissä

- lääketieteessä, poliisitoiminnassa, tiedustelussa, työhönotossa, virastoissa, osakekaupassa jne.

Itsenäinen päätöksenteko

- perustuu opittuihin assosiaatioihin
- valkokaulusammattien automaatio (palveluammatit, toimistotyö)
- digitaalinen filtteri: esim. mitä näyttää käyttäjälle hakukoneissa tai sosiaalisessa mediassa

Jäljittely: opittujen rakenteiden toistoa ja variaatiota

- generatiivinen tekoäly, esim. kielimallit (ChatGPT)

Fyysisiä sovelluksia

- robotit (kodinhoito, hoivarobotit), itseohjautuvat autot (taksit, rekat)
- älylaitteet, älykodit

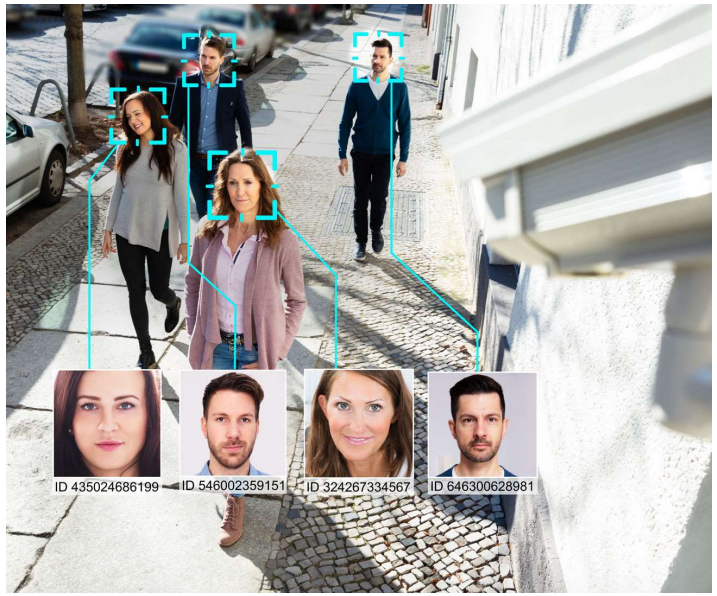
# Eettisiä riskejä

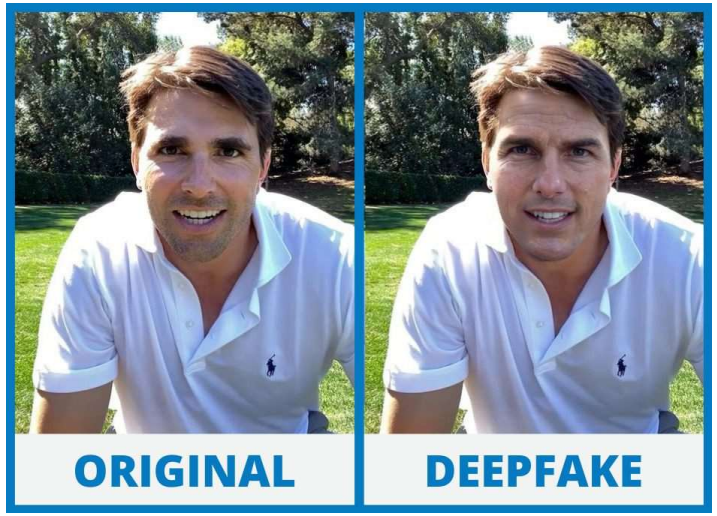
## Yksityisyys ja valvonta

- digitalisaatio ja datafikaatio → massiivinen datan keruu ja varastointi; tekoäly → tehokas analyysi
- Facebook: tiedämme pariskuntien eroavan ennen kuin he tietävät sen itse
- automatisoitu tarkkailu (digitaalinen ja fyysinen)
- eettinen kysymys: yksilön oikeus kontrolloida kuka tietää hänestä ja mitä

## Manipulaatio

- filtterien ohjaava vaikutus (voi olla ei-aiottua)
- täsmämarkkinointi, syväväärennökset



# Eettisiä riskejä

Läpinäkyvyys ja luotettavuus

- miten algoritmi päätyy tulokseensa? ei perusteluja, ei selitettävyyttä
- irrelevantit yhteydet, "hallusinoiva" tekoäly:
  (esim. ChatGPT ei sovellu tiedonhakuun)
- prosessien varmistus ja korjattavuus, vastuukysymykset
- miten olla eri mieltä tekoälyn kanssa?

Vinoumat päätöksenteossa

- opetusdatan vinoumat periytyvät algoritmin toimintaan
- irrelevantit rakenteelliset yhteydet voivat vahvistaa vinoumia ja tuottaa epäoikeudenmukaisia tuloksia
- osa ihmisen kognitiivisista vinoumista *korjaa* datan vähyyttä (opitut intuitiot); missä määrin näitä tulisi jäljitellä?

# Eettisiä riskejä

## Ihmisten ja robottien sosiaalinen vuorovaikutus

- sosiomorfismi (Johanna Seibt)
- ihmisen tekojen vaikutukset ihmiseen itseensä
- vuorovaikutuksen muotojen siirtyminen ihmisvuorovaikutukseen (esim. seksirobotit)
- antropomorfismin “petollisuus”
- ihmisarvokysymykset (esim. hoivarobotit)
- onko roboteilla oikeuksia?

## Automaatio, työllisyys ja taloudellisten rakenteiden muutokset

- uusin teknologinen vallankumous – jopa 40 % nykyisistä työpaikoista katoaa?
- uusia tilalle, mutta valtava yhteiskunnallinen muutos – oikeudenmukainen siirtymä?
- teknojätit, alustatalous, kulutuksen keskittyminen: kapitalismista teknofeodalismiin? (Yannis Varoufakis)
- kuluttajat algoritmin kouluttajina ilman korvausta
- vaikutukset demokratiaan?

# Eettisiä riskejä

Ilmastoeettiset ongelmat

- sekä tekoälyn kehittäminen että käyttö tarvitsevat paljon laskentatehoa – ja energiaa

Immateriaalioikeudet

- loukkaako harjoittelumateriaalin käyttö immateriaalioikeuksia? (esim. kuvien generointi)

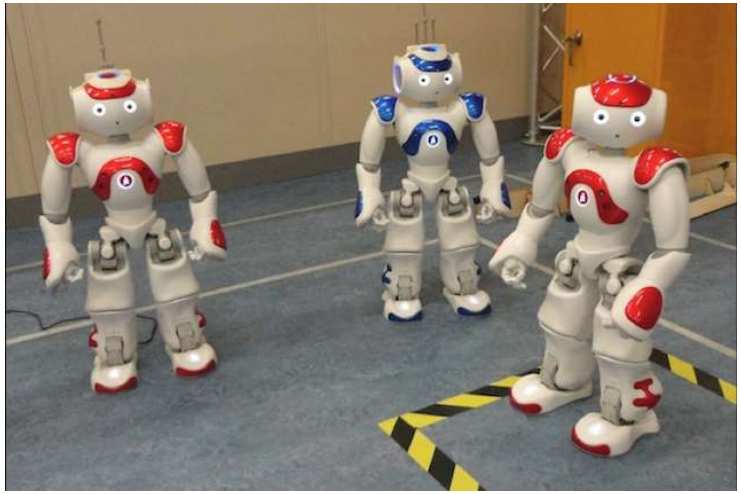
Autonomiset järjestelmät

- itseohjautuvat autot, autonomiset asejärjestelmät
- voivat toimia hyvin konteksteissa, joihin ne on suunniteltu, arvaamattomasti toisissa
- järjestelmässä voi olla bugeja

Yleinen tekoäly – liian hyvä, vaarallinen?

- huhu: OpenAI on yhdistänyt kielimallin ja logiikkamalli
- ei ”singulariteettia”, mutta sosiopaattisia toimijoita

# Kone-etiikka ja moraaliset koneet



Kone-etiikka

- tekniikan suunnittelu niin, että se toimii käyttöyhteydessään tavoilla, joita pidämme eettisesti oikeina
- mitkä ovat eettiset tavat toimia missäkin tilanteessa?
- jos tekoäly tekee itsenäistä päätöksentekoa vaihtuvassa joukossa tilanteita, sääntöjen on oltava yleisempiä

Koneet moraalisina toimijoina

- miten opettaa moraalisääntöjä tekoälylle?
- “top-down” (säännöt ensin) ja “bottom-up” (sääntöjen evolutiivinen oppiminen) molemmat osoittautuneet toimimattomiksi, hybridimallitkaan eivät ole olleet tyydyttäviä
- pitäisikö koneelle kehittää ensin muita valmiuksia, joiden varaan moraalikognitio rakentuu ihmiselläkin? (moraalin evoluutio keinotekoisen moraalin malliksi) (TK)
- voiko kone edes periaatteessa olla moraalinen toimija? (esim. Raul Hakli & Pekka Mäkelä: ei)

# Vastuu virheistä ja vastuullinen suunnittelu

Kuka on vastuussa virheistä ja odottamattomasta toiminnasta?

- **algoritmi** ei ole moraalinen toimija
- algoritmin **suunnittelija**? (ennakoimattomuus)
- algoritmin **käyttäjä**? (ei tunne algoritmin toimintaa)
- kaupallisen tuotteen **kehittäjä**?

Jos kukaan ei ole vastuussa, vastuu siirtyy siihen, että teknologia on otettu käyttöön tavalla jolla se on otettu

- teknologiseen determinismiin ei voi vedota

Teknologian vastuullinen suunnittelu:

- teknologian turvallisuuden ja luotettavuuden maksimointi
- käyttökontekstien huomioiminen suunnittelussa
- riskien kontekstuaalinen huomioiminen: käyttökontekstien valinta
- teknologian mukauttaminen käytäntöihin, ei käytäntöjen teknologiaan

# Arvosensitiivinen suunnittelu

Arvojen huomioiminen suunnittelussa, kehittelyssä ja käyttöönotossa

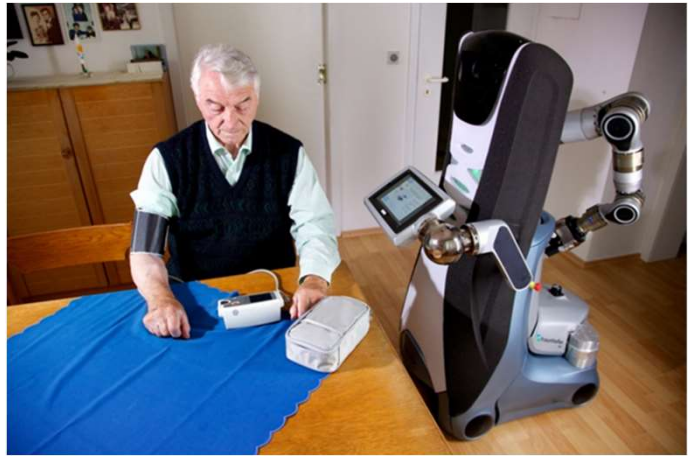
- moraaliset ja kulttuuriset arvot, säännöt ja periaatteet
- pyrkimyksenä teknologian sujuvampi integraatio ja toiminnallisten haittojen minimointi
- eettisten riskien minimointi ja käytäntöjen eettisyyden pohtiminen

Selvitettävä sen kokonaisuuden toiminta, johon teknologia tuodaan

- esim. hoivarobotit: ne hoitokäytännöt, joihin robotti tuodaan
- käytännön design ja suhde muihin käytäntöihin, automatisoitavan elementin funktio kokonaisuudessa
- “toissijaiset” funktiot (kuten vuorovaikutus hoitajan kanssa)

Kokonaisuuden arvorakenne

- miten arvot ilmenevät käytännöissä
- miten teknologia muuttaa arvojen ilmenemistä



# Arvosensitiivinen suunnittelu

Analyysin elementtejä:

- **konteksti**: esim. hoivatyön rakenne, organisointi ja resurssit, sosiaalinen konteksti
- **käytännöt**: funktiot ja miten käytäntöjen muuttaminen muuttaa muuta rakennetta
- **osapuolet**: keitä muutokset koskevat suoraan tai epäsuoraan (esim. potilaat, hoitajat, lääkärit, omaiset, administraatio)
- **moraaliset elementit**:
  - osapuolten ja käytäntöjen arvot ja niiden ilmeneminen konkreettisissa tilanteissa
  - arvot ja tarpeet, eivät preferenssit ja toiveet
  - esim. hoitoarvot: kyky huomata potilaan tarpeet ja vastata niihin oikealla tavalla, potilaan itsemääräämisoikeus, itsenäisen toimintakyvyn maksimointi, yksityisyys ja ihmisarvon tunne
  - myös paikka miettiä käytäntöjen eettisyyttä ja niiden kehittämistä

# Arvosensitiivinen suunnittelu

Suunnittelun vaiheet:

- **retrospektiivinen analyysi**: käytäntöjen elementtien arvioiminen (voi olla kriittistä)
- **prosepektiivinen analyysi**: oletetut vaikutukset
- **empiirinen analyysi**: prototyyppien suunnittelu ja testaaminen
- **tekninen analyysi**: tuotteen suunnittelu ja käyttöönotto
- **vaikutusten analyysi**: millaiseksi käytännöt ovat muodostuneet

# Sosio-teknologiset järjestelmät ja institutionaalinen suunnittelu

Sosiologinen teknologiantutkimus: teknologiat sosio-teknologisia hybridejä

- teknologialla on käyttökonteksti, joka on osa laajempaa käytäntöjen kontekstia

Uudet teknologiat ja yhteiskunnallinen muutos:

- muutokset elämäntavoissa, sosiaalisissa suhteissa ja käytännöissä, työelämässä, instituutioiden toiminnassa jne.
- kyse ei ole olemassa olevien käytäntöjen hallitusta muutoksesta, vaan uusien sosio-teknologisten hybridien syntymisestä
- teknologia ja sosiaalisuus ovat osa ihmisyyttä, samoin teknologinen, sosiaalinen ja yhteiskunnallinen (kulttuuri)evoluutio

Tarvitaan arvosensitiivistä institutionaalista suunnittelua

- yhteiskunnallisten instituutioiden ja niiden suhteiden tarkastelu arvojen, yhteiskunnallisten tavoitteiden ja eettisten riskien minimoimisen näkökulmasta
- ei välttämättä valtiollista ohjausta vaan teknologian regulointia ja mahdollistavien rakenteiden luomista
- etiikan ja politiikan ongelmallinen suhde