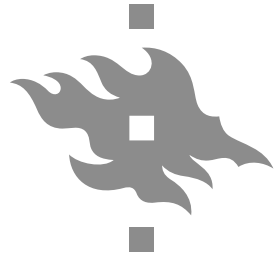
Eettisesti vastuullisen tekoälyn filosofiset perusteet

Raul Hakli (Helsingin yliopisto)

Ennakointiakatemia 6.11.2024

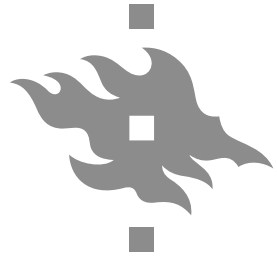


Tekoäly ja toimijuus

- Tekoälyllä tarkoitetaan tutkimusalaa, joka tutkii ja kehittää menetelmiä ja järjestelmiä, jotka suorittavat "autonomisesti" tehtäviä, jotka ihmisten suorittamina edellyttäisivät älykkyyttä. Tällaisia järjestelmiä voi kutsua tekoälyjärjestelmiksi.
- "Autonomialla" tarkoitetaan tässä (karkeasti) sitä, että järjestelmä on luontevaa nähdä *toimijana* (agenttina), joka toimii tavoitteellisesti ja itsenäisesti (ilman jatkuvaa ohjausta tai valvontaa).
- Toimijalla tarkoitetaan filosofiassa jotakin, jolla on *uskomuksia* ympäristöstään, *haluja* tai *tavoitteita*, joita kohti se pyrkii, ja kyky tehdä *päätöksiä* ja *tekoja*, jotka sen uskomusten valossa johtavat kohti tavoitteiden toteutumista.

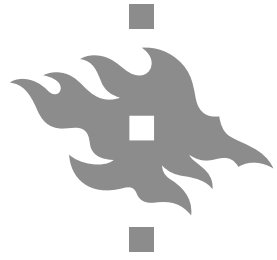
Esimerkkejä toimijoista

- Toimijoina voidaan pitää esimerkiksi:
  - aikuisia ihmisiä
  - lapsia
  - eläimiä
  - organisaatioita, kuten yritykset tai valtiot
  - koneita ja automaatteja, kuten termostaatti
  - tekoälyjärjestelmiä ja robotteja
- Kolme ensimmäistä luokkaa ovat *luonnollisia* toimijoita ja loput *keinotekoisia* toimijoita
- Joskus toimijat ymmärretään väljemmin (esim. kaikki joilla on kausaalista vuorovaikutusta ympäristön kanssa, esim. kemialliset agentit) ja joskus tiukemmin (esim. vain aidosti intentionaaliset ja ympäristöstään tietoiset olennot)



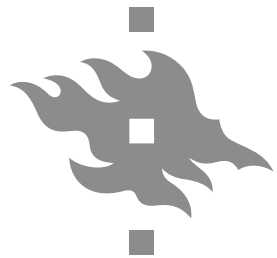
Etiikka ja moraalinen toimijuus

- Etiikka on filosofian ala, joka tutkii moraalisia valintoja ja arvoja: miten pitäisi elää ja toimia, mikä on oikein, mikä väärin
- Etiikka on läheisesti yhteydessä toimijuuteen: Eettiset kysymykset koskevat vain toimijoita, jotka kykenevät tekemään valintoja
- Etiikka ei kuitenkaan koske kaikkia toimijoita. Vain osalla on kyky pohtia eettisiä kysymyksiä tai tehdä moraalisia valintoja.
- *Moraaliset toimijat* ovat toimijoita, joilla on tällaiset kyvyt:
  - kyky tehdä vapaita valintoja,
  - kyky tunnistaa ja arvioida moraalisia perusteita ja
  - kyky toimia näiden arvioiden mukaisesti.
- Usein vaatimuksina mainitaan tietoisuus, emootiot tms.
- Keinotekoiset toimijat eivät toteuta ehtoja.



Vastuunkantokyky ja vastuullisuus

- Moraaliset toimijat kykenevät kantamaan *moraalista vastuuta*, he ovat *vastuunkantokykyisiä*.
- Jotta moraalinen toimija olisi lisäksi *vastuussa* jostain teostaan tai tekemättä jättämisestään, vaaditaan lisäksi, että
  - hänellä on riittävä *tieto ja ymmärrys* tekonsa luonteesta ja seurauksista, ja
  - hän toimii *tarkoituksellisesti ja vapaaehtoisesti*.
- Sillä, että moraalinen toimija on vastuussa teostaan, tarkoittaa, että voimme arvioida hänen tekoaan moraalisesti: voimme *moittia* tai *kehua* häntä teostaan.

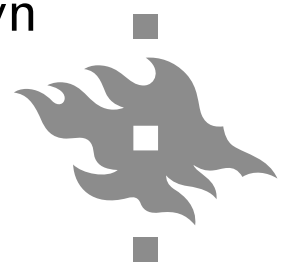


Etiikan rooli päätöksenteossa

- Etiikka kuuluu lähes kaikkeen päätöksentekoon. Vain harvat päätökset ovat ilmiselvästi eettisesti merkityksettömiä.
- Se on henkilökohtaista eikä sitä voi ulkoistaa muille (vrt. etiikkakomiteat)
- Se on monimutkaista eikä sitä voi pukea yksinkertaisiksi säännöiksi tai periaatteiksi (vrt. yritysten eettinen valkopesu tai etiikan trivialisointi tarkistuslistoiksi)
- *Eettisesti vastuullinen toiminta* on toimintaa, joka on sensitiivistä moraalisille perusteille ja niiden mukaista.
- Etiikan teoriat tarjoavat välineitä löytää moraalisia perusteita, esim. asiantilat, joista seuraa velvollisuuksia, tai tekojen luonne tai seuraukset. Esimerkkejä:
  - se, että lapsi on hädässä on asiantila, josta seuraa velvollisuus auttaa
  - se, että jonkin tekoälyjärjestelmän käyttö manipuloi käyttäjiä tai lisää eriarvoisuutta, on peruste sen käyttöä vastaan

Vastuuaukot

- Tehtävien automatisointi ja päätöksenteon siirtäminen koneiden tehtäväksi (tai monimutkaisten byrokraattisten sääntöjärjestlemien ratkaistavaksi) voi vähentää ihmisten vastuuta päätöksistä, sillä se vähentää heidän tietämystään tai kontrollia päätöksentekoprosessista.
- Koska vastuu ei siirry keinotekoisille järjestelmille, voi syntyä *vastuuaukkoja*.
- Päätöksenteon siirtäminen *tietoisesti* keinotekoisten toimijoiden harteille on vastuun pakoilua.
- Voi argumentoida, että vastuu ei tällöin katoa mihinkään, sillä vastuu voidaan jäljittää päätökseen käyttää esim. tekoälyjärjestelmää.
- Vastuullinen tekoäly tarkoittaa *ihmisten* vastuuta tekoälyn käytöstä ja kehityksestä.

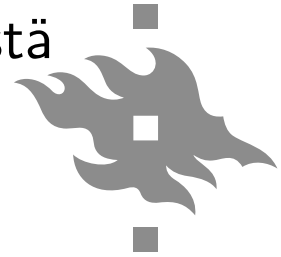


Instituutioiden vastuullisuus

- Keinotekoiset toimijat voivat jossain määrin simuloida vastuullista toimintaa.
- Instituutioiden vastuullisuus täytyy ymmärtää niiden ihmisten vastuullisena toimintana, joiden toiminta konstituoi instituution toiminnan: Toimiminen institutionaalisessa roolissa.
- Yksilöiden vastuullisen toiminnan kautta instituutiot voivat olla sensitiivisiä moraalisille perusteille, samoin kuin instituution arvoille ja tavoitteille.
- Instituution sääntöjen ja toimintakulttuurin tulee ohjata (tai oikeastaan pakottaa) ihmiset toimimaan vastuullisesti.

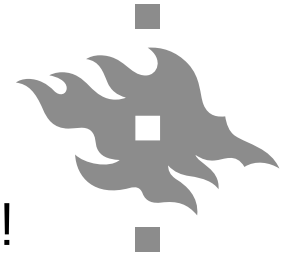
Vastuullisuus koulutusinstituutioissa (1)

- Koulutuskontekstissa moraalisia perusteita voisivat olla esimerkiksi se, että opiskelijoita kohdellaan tasaveroisesti tai että he saavat heille kuuluvaa opetusta.
- Tekoälyn tai minkä tahansa teknologian käyttöönotossa täytyy ottaa huomioon kaikki relevantit moraaliset perusteet ja lisäksi instituution omat arvot ja sitoumukset: sen perustavat tavoitteet ja periaatteet, esimerkiksi korkeatasoisen koulutuksen tarjoaminen.
- Teknologian käytössä täytyy siis puntaroida instituution arvoja ja tavoitteita sekä moraalisia perusteita. Vastuullinen koulutusinstituutio tekee päätökset teknologian käytöstä sen mukaan mikä tapa parhaiten palvelee instituution tavoitteita ja on eettisesti arvokkainta.
- Teknologiaoptimismin ja markkinapuheiden sijaan kriittistä harkintaa: Onko tieteellistä näyttöä teknologian todellisista hyödyistä?



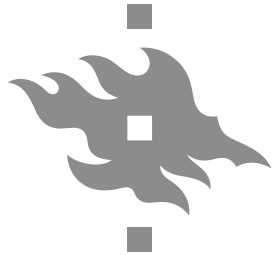
Vastuullisuus koulutusinstituutioissa (2)

- On esim. arvioitava milloin tekoälyn käyttö sallitaan opiskelijoille: Missä tapauksissa siitä on hyötyä oppimisessa? Kaiken salliva linja ei voi olla vastuullinen atropiavaaran takia.
- Toisaalta opiskelijoita pitäisi myös opastaa tekoälyn hyödyntämiseen ja kriittiseen arviointiin. Vaatii opettajilta perehtymistä.
- Myös opettajat voivat käyttää tekoälyä esim. opetussisältöjen tekemiseen tai opiskelijoiden työsuoritusten arviointiin. Joidenkin mukaan tämä säästää aikaa ja on ongelmatonta, koska opettajat osaavat jo asiat. Atropian vaara on kuitenkin edelleen olemassa. Lisäksi muita pohdittavia asioita kuten tasavertaisuus, vinoumat, yksityisyys, ymmärtämisen kaventuminen, mittaamisen ongelmat jne.
- Ja koulutusinstituutiot voivat käyttää tekoälyä vaikkapa opettajien arviointiin. Samat ongelmat kuin edellä.
- Lisäksi syytä pohtia tekoälyn ekologista kuormittavuutta!



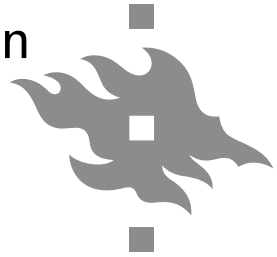
Vastuullisuuden toteuttaminen (2)

- Miten vastuullisuus toteutetaan? Ongelmana nk. *Collingridgen dilemma:* Teknologian suunnittelussa tai käyttöönotossa perustavanlaatuinen ristiriita tiedon ja kontrollin välillä.
- Erilaisia eettisiä työkaluja, mutta kaikki eivät tähtää perustesensitiivisyyteen. Päinvastoin, tarkistuslistat ja etiikkakomiteat toimivat melkeinpä päinvastoin ja joko trivialisoivat tai ulkoistavat eettisen harkinnan.
- On kuitenkin myös sellaisia välineitä ja lähestymistapoja, jotka pyrkivät nimenomaan lisäämään harkintaa useasta eri näkökulmasta. Tärkeää etukäteen tehtävä monipuolinen seurausten arviointi, jossa mukana kaikki relevantit osapuolet.
- Teoreettisista lähestymistavoista mainittakoon *arvosensitiivinen suunnittelu* ja konkreettisista työkaluista DEDA (Data Ethics Design Aid)



Vastuullisuuden toteuttaminen (2)

- Arviointikriteerinä moraaliset perusteet (esim. tekeekö tekoälyn käyttö ihmisten elämästä parempaa?) ja instituution konstitutiiviset arvot ja tavoitteet (esim. hyvän koulutuksen tarjoaminen, tiedon, taitojen ja ymmärryksen lisääminen)
- Tekoälyn käyttöä puolustetaan yleensä sillä, että se tehostaa ja nopeuttaa asioita tai tekee niiden tekemisen edullisemmaksi.
- Yksityisten yritysten kohdalla se että tuotetaan halvalla ja nopeasti voi olla arvojen ja tavoitteiden mukaista: konstitutiivinen tavoite on yksityinen etu, taloudellisen voiton maksimointi
- Sen sijaan julkisten instituutioiden tavoite on tuottaa yleistä hyvää kuten tietoa, taitoa ja ymmärrystä. Niiden konstitutiiviset arvot ovat yleensä pikemminkin laadullisia, esim. "korkealaatuinen koulutuksen tarjoaminen ja ymmärryksen syventäminen" pikemminkin kuin "auttavien taitojen antaminen mahdollisimman nopeasti"



Yhteenveto

- Inhimillisen työn – ajattelun, arvioinnin ja päätöksenteon – *korvaaminen* tekoälyllä vaatii tarkkaa harkintaa, oppilaitoksissa aivan erityisesti.
- Huomio moraalisiin perusteisiin ja instituution perustaviin arvoihin ja tavoitteisiin.
- Tehostamisen edut eivät välttämättä siirry laatuun tai ihmisten elämän parantamiseen.
- Tekoälyn toimintaperiaatteiden ymmärtämisen, vastuullisen käytön ja kriittisen arvioinnin *opettaminen* tärkeää. Kriittisyydellä tarkoitan ennen kaikkea eettisten, yhteiskunnallisten ja ekologisten näkökulmien painottamista.