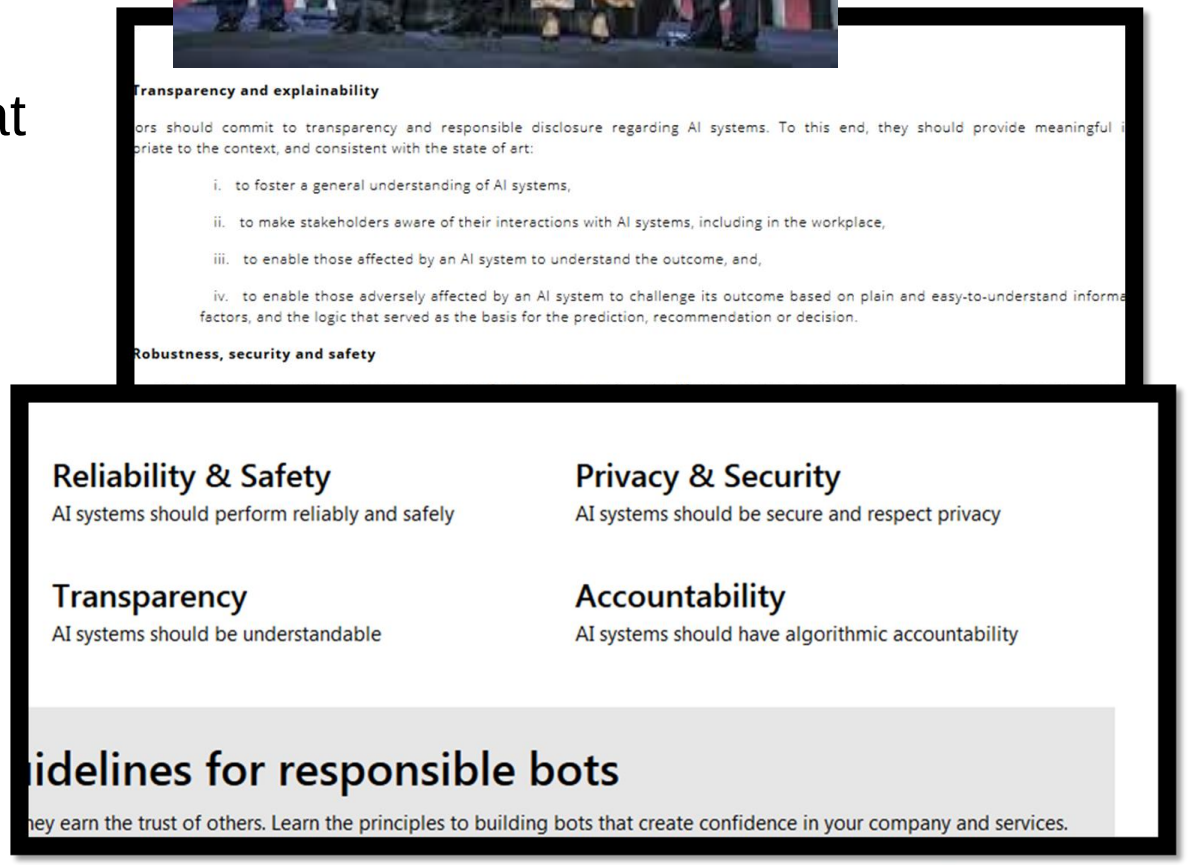
VALTIOVARAINMINISTERIÖ
FINANSMINISTERIET

Tekoälyn eettiset ohjeistukset: 2021 teemoja

Esittäjän nimi, tehtävänimike
pp.kk.vvvv
Tilaisuuden nimi

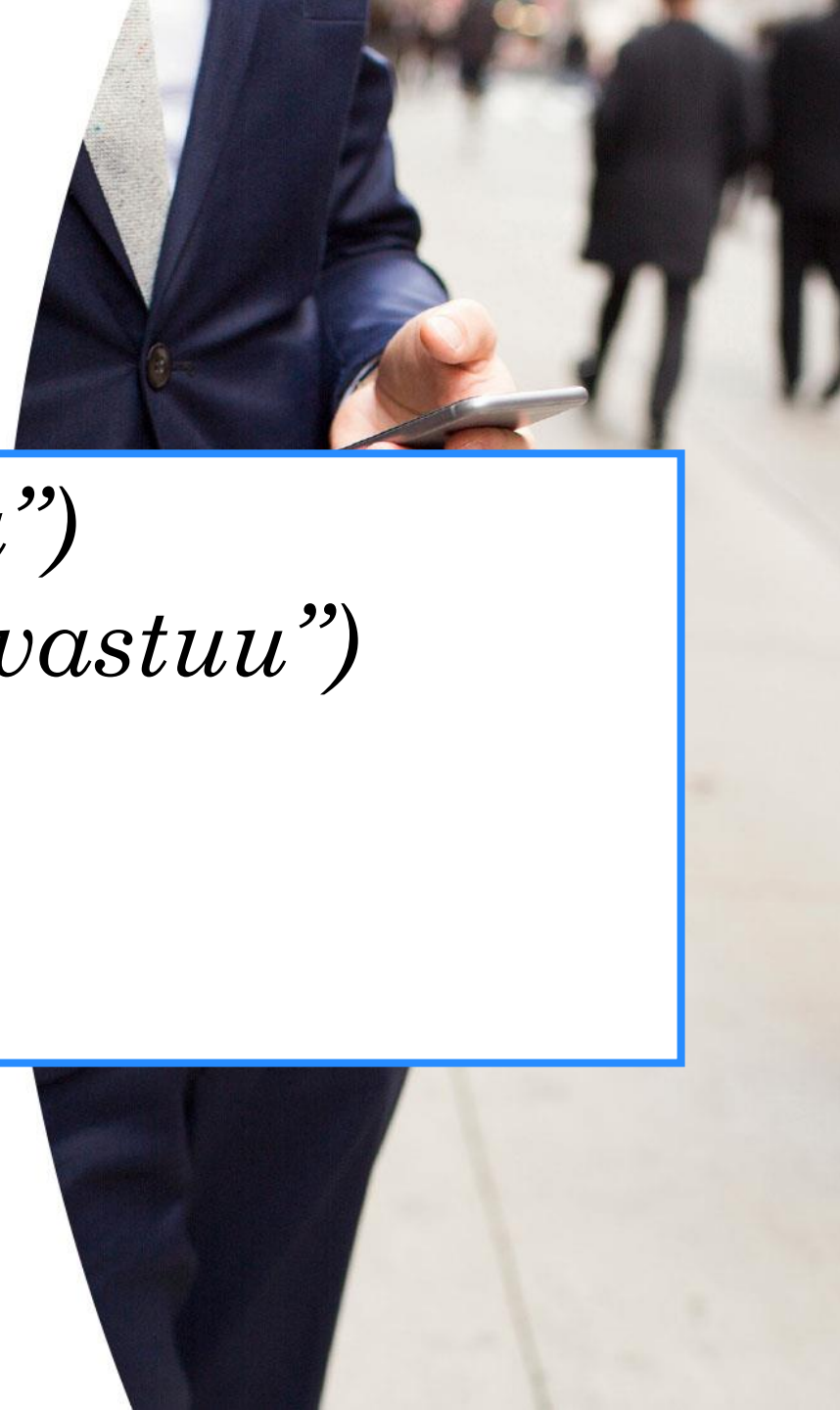
Tekoälyn etiikka 1.0: Sata ja yksi eettistä ohjeistusta

- 2016/2018 – 1. sukupolven eettiset ohjeistukset
  - 180 + ohjeistusta, ad hoc- komiteat (yritykset, organisaatiot,…)
  - Kokoelma irrallisia normeja, ei systemaattinen normisto
  - Ei usein erotella data/algoritmi-keskustelua
- Pääteemat:
- 2018-2020:
  - Pääteema: ”Trustworthy AI”
  - Muuttumassa: ”Sustainable” AI



VALTIOVARAINMINISTERIÖ
FINANSMINISTERIET

Tekoälyn etiikka 1.0:
Sata ja yksi ohjeistusta – top 5



- *non-maleficence (“älä tee pahaa”)*
- *responsibility/accountability (“vastuu”)*
- *Transparency (“läpinäkyvyys”)*
- *justice and fairness (“reiluus”)*
- *Privacy (“yksityisyys”)*

Tekoälyn eettiset ohjeistukset 1.0: Yleistä

- Eurooppalaisen moraalifilosofian ja yhteiskuntafilosofian vaikutus
  - arvo- ja normipohja
  - perinteinen päättely (esim. ns. kantilainen vastuullisuus; akti-omissio- epäsymmetria)
  - Ihmisoikeudet normien pohjana: Ei erottelua positiiviset/negatiiviset oikeudet, yleensä painotus jälkimmäisessä
- Ei varsinaisesti uusia teemoja
  - Normien ja arvojen resubsumptio; käytännön muutos – nostaa esiin uudessa kontekstissa kysymyksen (”coded bias”, ”accountability”…)
  - Mm. soveltava etiikka: bioetiikka, teknologian etiikka, neuroetiikka…
  - Tekoälyspesifejä: Läpinäkyvyys, data ja tiedollinen käyttö (esim. ”asianmukaisuus” ja ”tarkoituksenmukaisuus” tietojen hallinnassa)

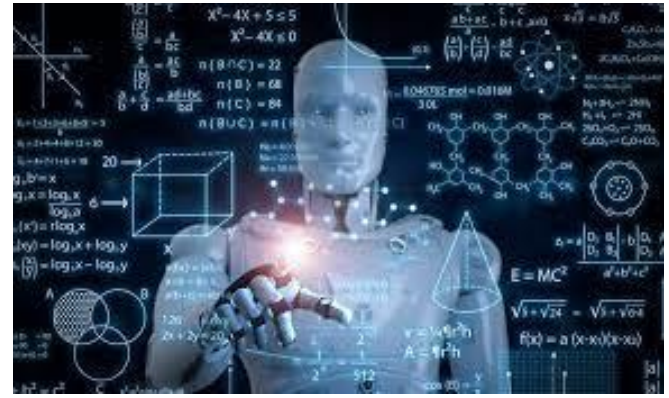
Tekoälyn eettiset ohjeistukset 1.0: Sata ja yksi ohjeistusta

- Kritikki mm.:
  - ”10 käskyä” (”Älä tee sitä, älä tätä”)
    - Sirpaleisuus, ei akateeminen systemaattisuus, normien ristiriitaisuus käytännössä
    - Painopiste: Harmien ennaltaehkäisy, riskien ylikorostuminen
    - Teknosolutionismi: Tekniset ratkaisut, ”pikafiksit”?
    - Regulatiorismi: Ongelmat sosioteknisiä, pehmeä sääntely ei riitä, kova regulaatio kieltoina – myös pikafiksi?
  - Abstraktioista käytäntöön
    - Ohjeistukset liian abstrakteja, edettävä tapausesimerkkien kautta
  - ”From ethics washing to ethics bashing”
    - Aggressiivinen liikehdintä esim. somessa
    - Poliittinen painotus, tulehtunut yhteiskunnallinen tilanne

Tekoälyn eettiset ohjeistukset 2.0: 2019->

1. Jäsentyminen: Systemaattinen tutkimus, kattava analyysi
   - Esim. Unesco, hallitustenväliset neuvottelu 2021, CAHAI2. Riskeistä riski-hyöty-vertailuun:
2. Harmeista riski-hyöty- vertailuun:
   - Riski: Miten estetään X (esim. yksityisyydensuoja/tietojen asianmukaisen käsittely) vaarantuminen?
     - Y-filosofia: oikeuksien negatiivinen muotoilu
     - Johtaa usein pikafikseihin, ”teknosolutionismi”, pinnallisuus, regulaatiohankkeet
   - Hyöty: Miten mahdollistetaan paremmin X:n tosiasiallinen toteutuminen (esim. yksityisyyden suoja -> synteettinen data)
     - Y-filosofia: oikeuksien positiivinen muotoilu
     - Monimutkaisempi analyysi; ”miten mahdollista” -> periaatetaso, innovaatioystävällisyys

Tekoälyn eettiset ohjeistukset 2.0: Abstraktioista konkretiaan – ja takaisin



- Terminaattoreista tosiasialliseen käyttöön esim. viranomaistoiminnassa
  - Esim. automaattinen päätöksenteko: oikeus tulla informoiduksi (GDPR), tietosuoja
  - Kaksi esimerkkiä:
    - **Oikeus saada selitys:**
    - Mitä tarkoittaa konkreettisesti?
    - Läpinäkyvyys, selitettävyys, ymmärrettävyys…?
    - **Tietosuoja ja asianmukainen tietojen käyttö:**
    - Data vai datasta tehtävät päätelmät?

Esimerkki: ”Oikeus saada selitys”

- 2016-2018: läpinäkyvyys (ex ante – selitys?)
  - Tekninen irrealismi
  - Ei ”oikeutus” päätöksille; vain systeemin kuvaus
  - ”mekanistinen/funktionaalinen selitys”
  - ”Black box”

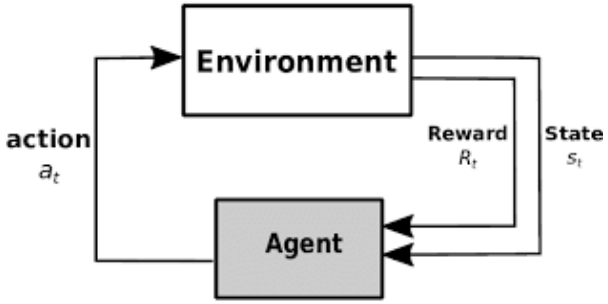
Henkilö X hakee sosiaalietuutta. Järjestelmä käsittelee hakemuksen automaattisesti. Päätös kielteinen, koska z ja w. Henkilöllä oikeus saada selitys, miksi.
- Mikä ”miksi”-kysymys?

Esimerkki: ”Oikeus saada selitys”: Transparenssista kontrastiiviseen oikeutukseen

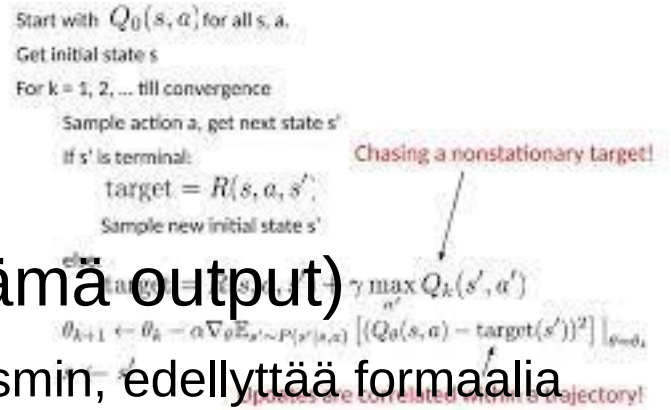
- ”Etiologinen”, kontrastiivinen selitys
  - Matsudakis 2018 yms
- Miksi päätös tämä?
  - Koska olosuhteet X ja Y ja säännöt p, r, q - > ei etuutta
  - Kontrastiivisuus: jos Z, niin päätös olisi ollut toinen
  - Oikeus valittaa osoittamalla, että jokin virhe
- Kritiikki:
  - kuvaa olosuhteet, ei tosiasiallista syiden ja seurausten ketjua (=kuinka päätös tehtiin?)
  - Jos virhe/systemaattinen virhe (esim. käsittelyjärjestelmä) – virheen osoittaminen-> selitettävyys
  - Edellyttää kuvausta järjestelmästä -> selitettävyys

Henkilö X hakee sosiaalietuutta.
Järjestelmä käsittelee hakemuksen automaattisesti.
Päätös kielteinen, koska z ja w.
Henkilöllä oikeus saada selitys, miksi.
- Mikä ”miksi”- kysymys?

Oikeus saada selitys: Oikeutuksesta toiminnalliseen selittämiseen



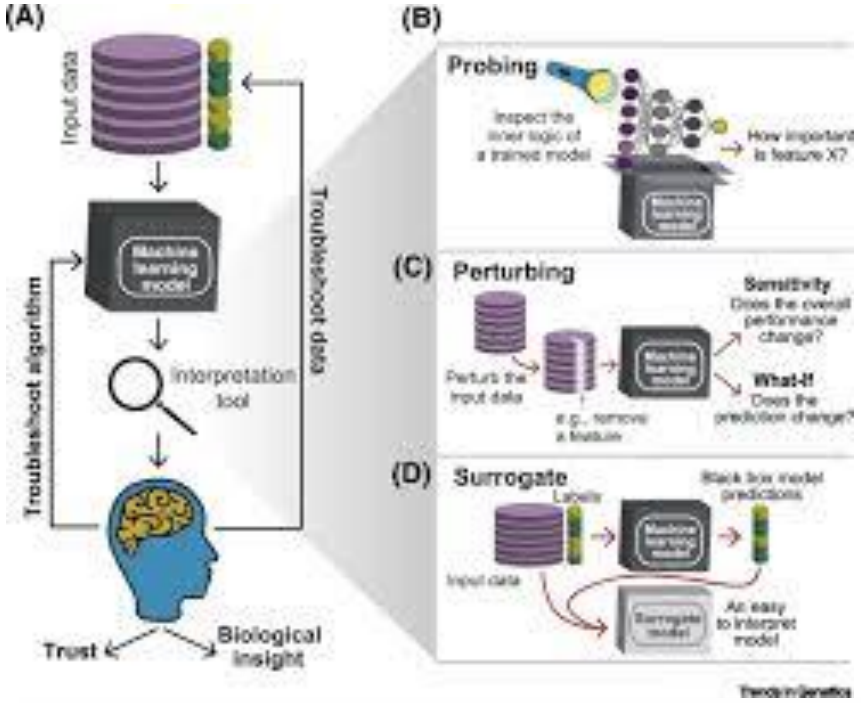
- Selittämisen tasot:
- Formaali taso
  - Matemaattinen kuvaus algoritmista – mitä ongelmaa laskee
  - Täydellinen ylätason ”selitys”:
    - vastaa kysymykseen ”miksi” matemaattisesti
    - ei vastaa kysymykseen ”kuinka”



- Koneellisen laskennan taso (kuinka/miksi tämä output)
  - Syy- ja seurausselitys, kuvaa laskennallisen mekanismin, edellyttää formaalia tasoa
  - Laskennallinen ex ante esim. koodin muodossa, mutta tosiasiallinen mekanismi usein mahdoton kuvata erityisesti DCNN- järjestelmissä

Oikeus saada selitys:
- selitettävyydestä auditointiin

- Nykyjärjestelmät: usein ei pääsyä syviin kerroksiin – ”black box by definition”
  - Seuraus: Vain testimenetelmiä – matemaattinen arvio, millä todennäköisyydellä kone erehtyy
  - Mutta: Testimenetelmien arvioinnin ongelmat?
    - Haibe-Kaines & al, 2020: ”Toistettavuuskriisi” kone-oppimisessa – ei yhtenäistä käytäntöä dokumentoinnissa, tulokset eivät usein toistettavissa

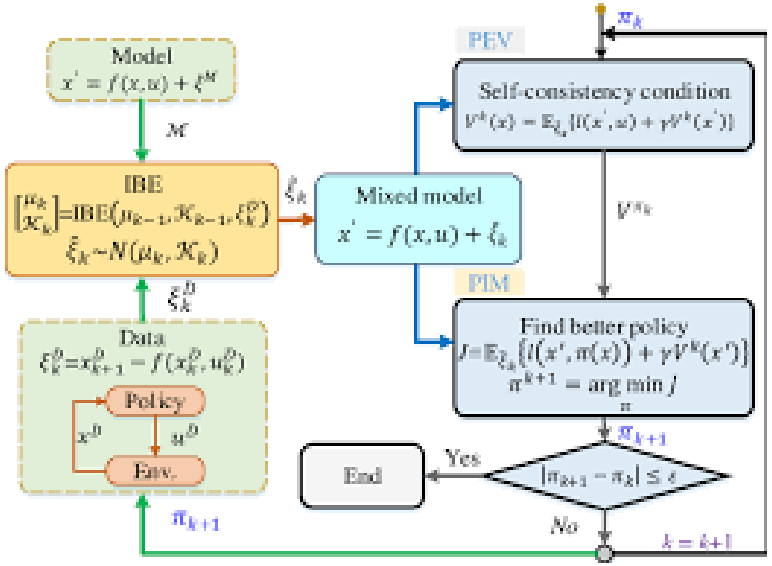


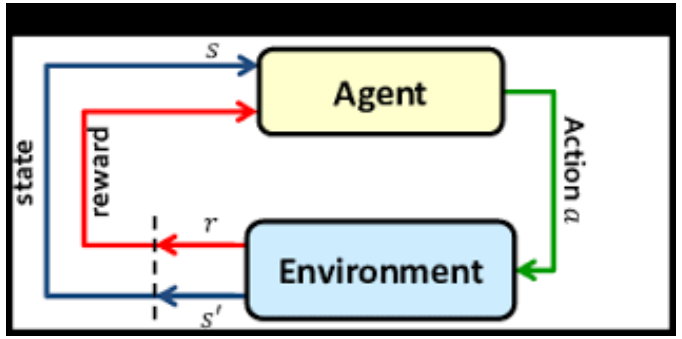
hardware, and its release is therefore not feasible”. Computational reproducibility is indispensable for robust AI applications[5,6] more complex methods demand greater transparency[7]. In the absence of code, reproducibility falls back on replicating methods from textual description. Although, the authors claim that *“all experiments and implementation details are described in sufficient detail in the Supplementary Methods section to support replication with non-proprietary libraries”*, key details about their analysis are lacking. Even with sufficient description, reproducing complex computational pipelines based purely on text is a subjective and challenging task[8,9].

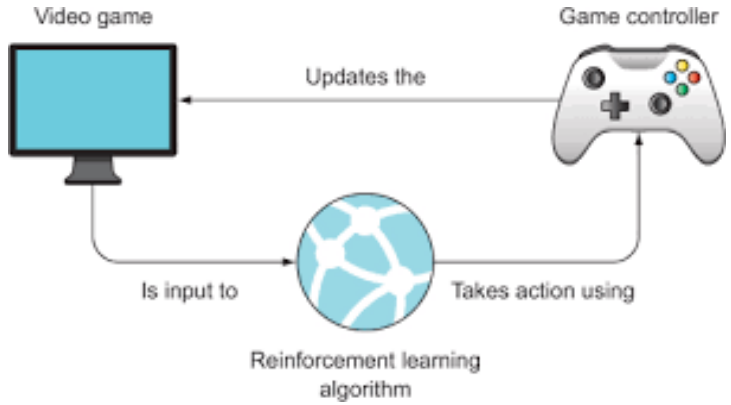
More specifically, the authors’ description of the model development as well as data processing and training pipelines lacks critical details. The definition of multiple hyperparameters for the model’s architecture (composed of three networks referred to as the Breast, Lesion, and Case models) is missing (Table 1). The authors did not disclose the parameters used for data augmentation; the transformations used are stochastic and can significantly affect model performance[10]. Details of the training pipeline were also missing. For instance, they state that

Oikeus saada selitys: 3. Ymmärrettävyys

- Tarkoituksenmukaisuus
- Saavutettavuus, lukutaitokysymykset, kulttuurinen osaamistaso
- Tutkimuskohde:
- Kognitiontutkimus (algoritminen lukutaito), datavisualisaatio, informaatiomuotoilu…







VALTIOVARAINMINISTERIÖ
FINANSMINISTERIET

Esimerkki 2: Tietojen asianmukainen käyttö

- Esim. ”Henkilötietojen asianmukainen käyttö” päätöksenteossa
- Asianmukaisen käytön määritelmä:
- Wachter & Mittelstadt (2019): Datasta tiedolliseen käyttöön
  - Koskeeko vain henkilöä koskevaa dataa? Vai myös päätelmiä, joita analytiikkamenetelmien avulla voidaan muodostaa?
  - Jos päätelmiä, niin huomattavan hankalaa?

Esimerkki 2: Tietojen asianmukainen käyttö

- Asianmukaisuuden rajaamisen kriteerit:
  - ”Minimi”: mahdollisimman vähän tietoja (riskipohjaisuus?)
  - Maksimi: mahdollisimman paljon tietoja (hyötypohjaisuus?)
- Asianmukaisuuden kriteeri:
  - Oltava oleellista/tarkoituksenmukaista henkilöä koskevan asian käsittelyn kannalta
  - Mitä on oleellisuus/tarkoituksenmukaisuus?

Esimerkki 2: Tietojen asianmukainen käyttö - Oleellisuus

- Oleellisuuskriteerit tiedollisissa päätelmissä huomattavan hankalia
- Esim. Mitä tietoja tarvitaan? Miksi?

A. ***Mahdollistaa mahdollisimman oikea päätöksenteko?***

   - Sarasevic & co (informaatiotutkimus): oleellisuus ”aiheenmukaisuutta”, kriteerit esim. lainsäädäntö
   - (Riskipohjaisuus: turhien tietojen käytön välttäminen?)

B. ***Vai mahdollistaa mahdollisimman oikea päätöksenteko mahdollisimman tehokkaasti?***

   - esim. Sperber, Wilson: oleellisuus aiheenmukaisuutta ja tiedonkäytön tehokkuutta
   - Tiedollinen hyöty/tiedonkäytön kustannukset: kokonaistehokkuus
   - (Hyöty: prosessien tehokkuus?)

Yhteenveto

- Tekoälyn eettiset ohjeistukset:
- Itse etiikka vanhaa, ohjeistusten tutkimus uutta
- Dynamiikka: abstraktioista käytäntöön, käytännöistä takaisin abstraktioihin

VALTIOVARAINMINISTERIÖ
FINANSMINISTERIET

Kiitos