Merkintöjä tekoälystä

TEP-seminaari 23.11.2019
Pekka Orponen
Aalto-yliopisto

Tekoälyn lyhyt historia (1/3)

- Kysymys koneiden älykkyydestä esillä tietokoneiden alkuvaiheista lähtien
  - A. Turing: *Computing machinery and intelligence* 1950 (Turingin testi)
  - C. Shannon: *Programming a computer for playing chess* 1949; oppiva labyrinttihiiri *Theseus* 1950
- Darthmouthin kesäseminaari 1956
  - ”Artificial Intelligence” (J. McCarthy)
- Ensimmäinen näytös 1956-1980
  - menestys 1956-1974: hakupäättely, neuroverkot (perceptron/F. Rosenblatt 1958); valtava tutkimus- ja rahoitusinnostus
  - pettymys 1974-1980: symbolinen päättely liian raskasta, perceptron-neuroverkot liian rajoittuneita; rahoituksen romahdus

Tekoälyn lyhyt historia (2/3)

- Toinen näytös 1980-2010
  - 1980-1985: logiikkapohjaiset mallit ja päättely, ”asiantuntijajärjestelmät”
    - ensin innostus ja rahoitus, sitten pettymys kun järjestelmät liian rajallisia (arkijärjen ongelma)
  - 1985-1995: neuroverkkojen toinen tuleminen (backpropagation-verkot/G. Hinton et al. 1982)
    - ensin valtava innostus, sitten kiinnostuksen hiipuminen: verkkojen opettaminen vaikeaa, sovellusalueet rajallisia
  - hiljaiselo 1995-2010:
    - asiantuntijajärjestelmät haudattu, mutta logiikkapohjaisen mallintamisen ja päättelyn tutkimus jatkuu
    - neuroverkot passé, mutta automatisoidun tilastollisen päättelyn (”koneoppimisen”) tutkimus jatkuu
    - noin 2005 alkaen datatieteen ja ”big datan” nousu

Tekoälyn lyhyt historia (3/3)

- Kolmas näytös 2010-
  - automatisoitu tilastollinen päättely (koneoppiminen) suurista aineistoista alkaa tuottaa kaupallisesti arvokkaita tuloksia
  - syvät neuroverkot: ”backpropagation strikes back”
    - G. Hinton et al. 2006
  - jälleen valtava tutkimus- ja rahoitusinnostus
  - onko jokin muuttunut?
    - saatavilla olevan datan valtava määrä
    - tietokoneiden laskentatehon kasvu
    - aito ja jopa huomattava kaupallinen merkitys

Tekoäly: mitä se on?

- Kaksi päälähestymistapaa:
  - logiikkapohjaiset mallit ja niihin perustuva eksakti päättely (nykyisin ns. SAT-ratkaisimet)
  - tilastolliset mallit ja niihin perustuva tilastollinen päättely (nykyisin koneoppiminen, neuroverkot)
  - myös näiden yhdistämistä tutkitaan, mutta yllättävän vähän (koulukuntaerot?)
- Tilastolliset lähestymistavat tällä hetkellä huomattavasti suuremman kiinnostuksen kohteena

Vertailua

Logiikkapohjainen

+ Täsmälliset mallit, perusteltavuus, läpinäkyvyys

- Mallien rakentaminen vaatii ammattitaitoa

- Ratkaisimet laskennallisesti raskaita

+/- Datan hyödyntämistä tutkittu vähän

+/- Teknologia soveltuu paremmin suunnitteluun kuin päättelyyn

Tilastollinen

+++ Vahvat tilastolliset mallit ”oppivat” suurista datamääristä hämmästyttävän hyvin

+ Teknologia pystyy hyödyntämään alati kasvavia datamääriä

+ Teknologia on helposti käyttöön otettavaa ja opetusvaiheen jälkeinen päättely tehokasta

- Teknologia ei sovellu suunnitteluun eikä sellaisenaan mihinkään missä vaaditaan eksakteja tuloksia

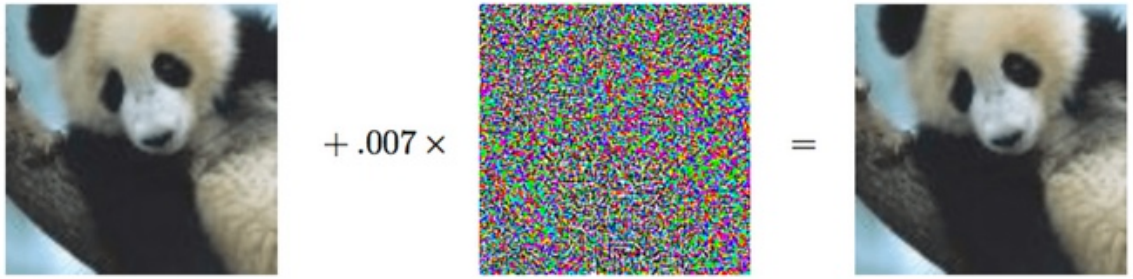
-- Mallipäättelyn tulokset ovat aina vain tilastollisia, eikä niihin usein liity edes luotettavuustietoa

-- Päättely on sidoksissa käytettyyn opetusdataan

+/- Monimutkaisten mallien opettaminen vaatii tällä hetkellä erittäin suuria datamääriä ja laskentaresursseja

Tekoäly: mitä se ei ole (1/2)?

- Yleisälykkyyden t. arkijärjen ongelma:
  - Nykyiset menetelmät alkavat olla erittäin hyviä ratkomaan rajattuja erityistehtäviä, mutta osaaminen ei siirry muunlaisiin tehtäviin
  - Tilastolliset menetelmät oppivat hyvin, mutta eivät ymmärrä maailmaa → päätöksenteon ”hauraus”



Goodfellow et al.
ICLR 2015

Tekoäly: mitä se ei ole (2/2)?

- Kehollisuuden ongelma:
  - H. Dreyfus, *What Computers Can't Do* (1972)
  - Ihmisen kaltainen älykkyys saattaa vaatia, että ”elää ihmisenä maailmassa” (Heidegger)
  - Esim. tuoli-käsitteen määrittely?
    - looginen: neljä jalkaa ja istumataso
    - tilastollinen: esimerkkikuvista hahmotettu yhteinen piirre
    - ”hermeneuttinen”: jokin, millä ihminen voi istua

Tekoälyn käytön riskejä (1/2)

- Peruslähtökohta: koneoppimis- ja neuroverkkomenetelmät eivät ole älyä eivätkä magiaa, vaan **sovellettua (laskennallista) tilastotiedettä** (”statistics on steroids”)
- Kaikessa argumentaatiossa kannattaa siten kysyä, miten se soveltuu tilastotieteeseen
- Keskeiset teknologisten riskien syyt tällä hetkellä:
  - Menetelmien läpinäkymättömyys
  - Riippuvuus opetusaineistosta
- Pidemmälle katsoen myös isoja taloudellisia, sosiaalisia ja kulttuurisia haasteita

Tekoälyn käytön riskejä (2/2)

- Tekoälyn käytön riskikysymykset eivät ehkä ole eettisiä vaan juridisia: mikä on oikea säädöspohja ja hyvät käytännöt tämän tilastollisen teknologian käytölle (vrt. ydinvoimateknologia, ilmailuteknologia, rakennusteknologia):
  - Mitkä ovat käyttöön otettujen järjestelmien luotettavuusvaatimukset ja miten niitä auditoidaan?
  - Kuka tai mikä organisaatio vastaa mistäkin virhetoiminnosta, kuinka pitkään ja missä rajoissa? Mitkä ovat sanktiot?
  - Mitkä ovat järjestelmien sallitut käyttötilanteet? Sanktiot?
  - Millaisissa tilanteissa viimekätinen päätöksenteko voidaan jättää automaattiselle järjestelmälle, ja milloin päätökselle tarvitaan vastuullisen käyttäjän vahvistus?

- Toimiva esimerkki tällaisesta tietojärjestelmien sääntelystä: GDPR
  - GDPR 22(1): ”Rekisteröidyllä on oikeus olla joutumatta sellaisen päätöksen kohteeksi, joka perustuu pelkästään automaattiseen käsittelyyn, kuten profilointiin, ja jolla on häntä koskevia oikeusvaikutuksia tai joka vaikuttaa häneen vastaavalla tavalla merkittävästi.”

Riskien hallinta (1/2)

- *Informatics Europe/ACM Europe Recommendation on Machine-Learned Automated Decision Making* (2018):

1. Establish means, measures and standards to assure ADM systems are fair.
2. Ensure ethics remain at the forefront of, and integral to, ADM development and deployment.
3. Promote value-sensitive ADM design.
4. Define clear legal responsibilities for ADM's use and impacts.
5. Ensure the economic consequences of ADM adoption are fully considered.
6. Increase public funding for non-commercial ADM-related research significantly.
7. Foster ADM-related technical education at the university level.
8. Complement technical education with comparable social education.
9. Expand the public's awareness and understanding of ADM and its impacts.

Riskien hallinta (2/2)

- *ACM US & ACM Europe Public Policy Statement on Algorithmic Transparency and Accountability (2017):*

1. *Awareness:* Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

2. *Access and redress:* Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.

3. *Accountability:* Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

4. *Explanation:* Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.

5. *Data Provenance:* A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorized individuals.

6. *Auditability:* Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.

7. *Validation and Testing:* Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.

Lähteitä

1. S. J. Russell & P. Norvig, *Artificial Intelligence: A Modern Approach, 4th Ed.* Pearson Education, 2020 (forthcoming). http://aima.cs.berkeley.edu/
2. Y. LeCun, Y. Bengio & G. Hinton, Deep learning. *Nature 521* (2015), 436-444. https://doi.org/10.1038/nature14539
3. H. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence.* Harper & Row, 1972 (2nd Ed. 1979, 3rd Ed. 1992). https://archive.org/details/whatcomputerscan017504mbp
4. I. J. Goodfellow, J. Shlens & C. Szegedy, Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations (ICLR 2015),* San Diego CA, May 2015. https://arxiv.org/abs/1412.6572
5. M. Hildebrandt, *Law for Computer Scientists and Other Folk.* Oxford University Press, 2020 (forthcoming). https://lawforcomputerscientists.pubpub.org
6. *When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making.* Informatics Europe & ACM Europe Council, 2018. https://www.informatics-europe.org/publications.html
7. *Statement on Algorithmic Transparency and Accountability.* ACM US Public Policy Council & ACM Europe Council, 2017. https://www.acm.org/articles/bulletins/2017/january/usacm-statement-algorithmic-accountability