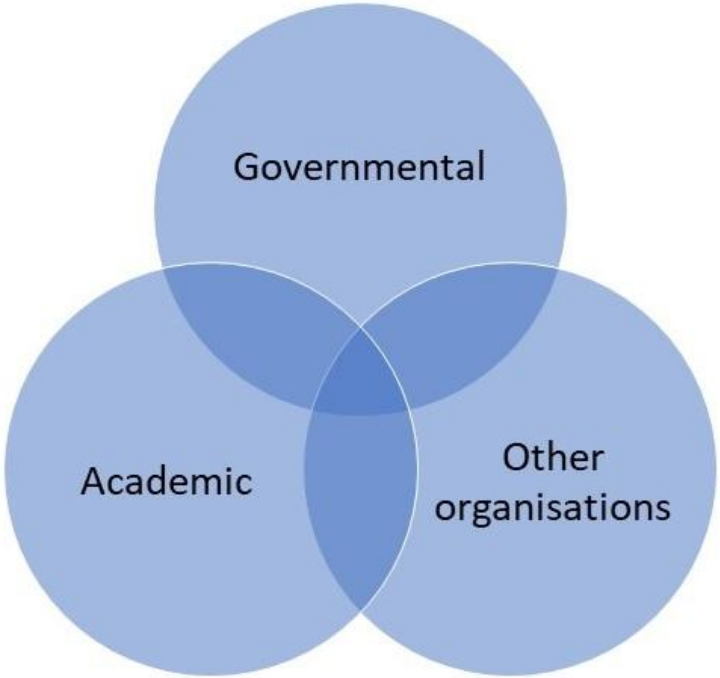
Tekoälyn etiikka:

Mitä kansainväliset organisaatiot ja kansalaisjärjestöt suosittelevat?

Turku AI Society
Juho Vaiste, 4.6.2018

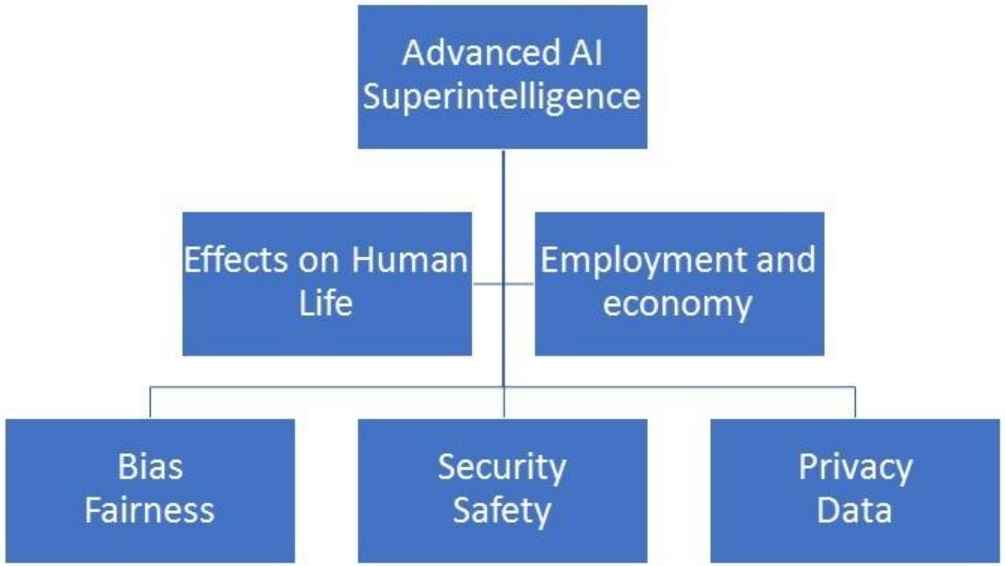
Tekoälyn ohjeistukset ja raportit



TURKU AI SOCIETY

AI Ethics vs Tech Ethics

Intelligence



Transparency

Rights

Unmanipulated

Organisaatiot ja lähteet

- AINow (New York University)
- AI100 (Stanford)
- The Royal Society
- World Economic Forum
- Internet Society
- MIRI (Machine Intelligence Research Institute)
- Future of Humanity Institute (Oxford)

- ACM (Association for Computing Machinery)
- Tivia (Tieto- ja viestintätekniikan ammattilaiset)

- IEEE (Ville Vakkuri esittelee)

TURKU AI SOCIETY

AINow Institute (New York University)

- Vahva painotus data-/algoritmivääristymien ja algoritmisen/koneellisen päätöksenteon ongelmiin
- Monimuotoisuus tekoälykehityksessä

- USA-keskeiset: monet ensi vaiheen AI-ongelmat ovat siellä paljon vakavampia (vt. Cathy O'Neil, valvonta, yksityiset laitokset, epätasa-arvo)

- *"We track the growing interest in ethical codes of conduct and principles, while noting that these need to be tied more closely to everyday AI design and development."*

TURKU AI SOCIETY

AINow: Algorithmic Impact Assessments

- Algorithmic Impact Assessments
  - Malli tai ohjeistus jokaisen julkisen päätöksentekoalgoritmin suunnittelu-, käyttöönotto- tai ostoprosessiin

1. Establishing Scope: Define “automated decision system”
2. Public notice of existing and proposed automated decision systems: Alert communities about the systems that may affect their lives
3. Internal agency self-assessments: Increase the capacity of public agencies to assess fairness, justice, due process, and disparate impact
4. Meaningful access: Allow researchers and auditors to review systems once they are deployed

.

TURKU AI SOCIETY

Stanford One Hundred Year Study

- Arvatenkin keskittyy tekoälyn mahdollisuuksiin. Etiikassa usein paljon kritiikkiä ja rajoitusta -> syytä muistaa myös hyvät puolet

- Kuitenkin; Ensimmäisiä kriittisiä lauseita virtuaali-/digitaalimaailmaan siirtymisestä

- Erityisesti Suomeen mietittäväksi:
  Low-resource Communities
  - Resurssien jako ja allokaatio
  - Jakamistalous, yhteiskunnallinen yrittäjyys

TURKU AI SOCIETY

Internet Society

- **Ethical Considerations in Deployment and Design:** AI system designers and builders need to apply a user-centric approach to the technology. They need to consider their collective responsibility in building AI systems that will not pose security risks to the Internet and Internet users.
- **Ensure "Interpretability" of AI systems:** Decisions made by an AI agent should be possible to understand, especially if those decisions have implications for public safety, or result in discriminatory practices.
- **Public Empowerment:** The public's ability to understand AI-enabled services, and how they work, is key to ensuring trust in the technology.

TURKU AI SOCIETY

Internet Society

- **Responsible Deployment:** The capacity of an AI agent to act autonomously, and to adapt its behavior over time without human direction, calls for significant safety checks before deployment, and ongoing monitoring.
- **Ensuring Accountability:** Legal accountability has to be ensured when human agency is replaced by decisions of AI agents.
- **Social and Economic Impacts:** Stakeholders should shape an environment where AI provides socio-economic opportunities for all.
- **Open Governance:** The ability of various stakeholders, whether civil society, government, private sector or academia and the technical community, to inform and participate in the governance of AI is crucial for its safe deployment.

TURKU AI SOCIETY

The Royal Society: Machine Learning

- Suositus raportille: ymmärrettävä ilman teknistä taustaa, kuitenkin kattava esitys koneoppimisesta

- Continued engagement between machine learning researchers and the public is needed
- Society needs to give urgent consideration to the ways in which the benefits from machine learning can be shared across society
- It is not appropriate to set up governance structures for machine learning per se. While there may be specific questions about the use of machine learning in specific circumstances, these should be handled in a sector-specific way

TURKU AI SOCIETY

World Economic Forum

- Selkeitä ylätason raportteja eri ongelmakohdista (Bias, Employment, Governance, Ethical Problems)
- Tarkasteltava kriittisellä CSR-silmällä

- How to Prevent Discriminatory Outcomes in Machine Learning
- Agile Governance: Reimagining Policy-making in the Fourth Industrial Revolution
- Eight Futures of Work: Scenarios and their Implications

TURKU AI SOCIETY

WEF: Eight Futures of Work: Scenarios and their Implications

- Workforce reskilling, education systems reform
- Enhanced digital access
- Agile safety nets
- Job protection incentives (?)
- Smart job creation incentives (?)
- Support to mass entrepreneurship

- Pääosin tartuttu jo toimeen Suomessa?

TURKU AI SOCIETY

WEF: Agile Governance

- Teknologian kehityksen nopeus on suuri ja valtavasti suurempi kuin aikaisempien teollisten vallankumousten
- Policy Labs, Regulatory sandboxes, Increasing agility through the use of technology, Promoting governance innovation

- Beyond government: industry self-regulation, ethical standards, etc.
- Ajateltava lisäkeinona ja tärkeänä osana käytäntöön saattamista

TURKU AI SOCIETY

Machine Intelligence Research Institute: Creating Friendly AI

- “Yudkowskyn koulukunta”

- Koskee vain yleistekoälyä, mutta politiikkaosion sisällöstä johdettavissa
  - Ei todennäköistä, että regulaatiolla varmistettaisiin Friendly AI:n toteutus
  - Koodariyhteisön kulttuuri sekä tietämys asioista ratkaisee
  - Tutkimusta avoimista kysymyksistä ja mikä olisi tapa rakentaa “friendly AI”

TURKU AI SOCIETY

Future of Humanity Institute (Oxford)

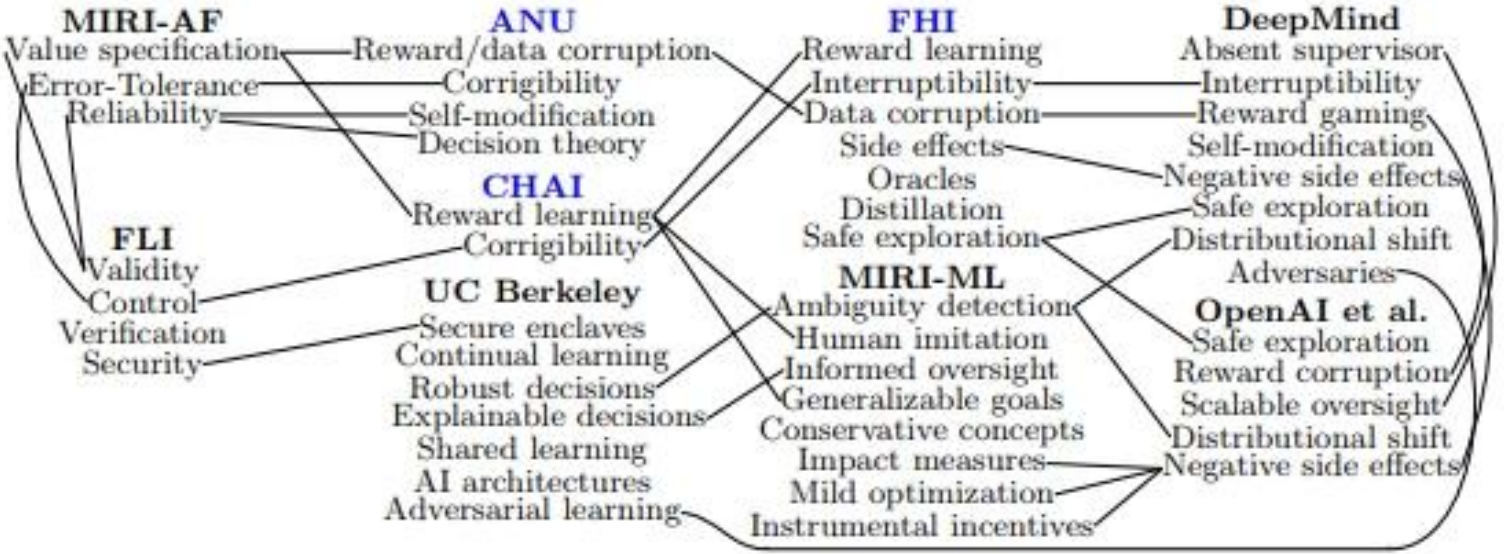
- Tunnetuin AGI- ja superintelligence-keskus
- Painopisteenä juuri nyt myös AI Governance, mutta nimenomaan yleistekoälyn diskurssissa

- Osallistuminen akateemiseen keskusteluun, mutta myös raportit:
  - Deciphering China's AI Dream
  - The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

- Miksi? Näyttää todennäköiseltä, että ainakin jonkin tason "advanced AI" saavutetaan
  → Vrt. ilmastonmuutos, on tärkeää, että työ aloitetaan mahdollisimman pian, voi olla, että tämän kanssa on kiireempi

TURKU AI SOCIETY

AGI Safety


MIRI-AF
Value specification
Error-Tolerance
Reliability
FLI
Validity
Control
Verification
Security
ANU
Reward/data corruption
Corrigibility
Self-modification
Decision theory
CHAI
Reward learning
Corrigibility
UC Berkeley
Secure enclaves
Continual learning
Robust decisions
Explainable decisions
Shared learning
AI architectures
Adversarial learning
FHI
Reward learning
Interruptibility
Data corruption
Side effects
Oracles
Distillation
Safe exploration
MIRI-ML
Ambiguity detection
Human imitation
Informed oversight
Generalizable goals
Conservative concepts
Impact measures
Mild optimization
Instrumental incentives
DeepMind
Absent supervisor
Interruptibility
Reward gaming
Self-modification
Negative side effects
Safe exploration
Distributional shift
Adversaries
OpenAI et al.
Safe exploration
Reward corruption
Scalable oversight
Distributional shift
Negative side effects

Ammattietiikan ohjeistukset antavat pohjaa AI-etiikalle

- ACM (Association for Computing Machinery)
- Tivia (Tieto- ja viestintätekniikan ammattilaiset)

- IEEE (Ville Vakkuri esittelee)

- IEEE:n vahva ote muuttuneeseen tulevaisuuteen. ACM ja Tivia päivittämässä?

TURKU AI SOCIETY

Muuta

- Valtiolliset: tällä hetkellä tiedossa Ranska, Iso-Britannia, EU ja USA. Pian Suomi, hienoa!

- Lait ja prinsiipit: löytyy nettisivuiltani käännettynä, tällä hetkellä tärkein projekti lienee Asilomarin lait
- Kokoelma tärkeimmistä akateemisista artikkeleista: elokuussa

- Aktiivista osallistumista akateemikoilta Twitterissä, #aiethics alla. Suomessa vielä heräilemässä.

TURKU AI SOCIETY

Suomen tekoälyetiikan yhteisö

- Akateeminen
  - Helsinki: MOIM, Rajapinta, HY, Aalto
  - Turku: Turku AI Society, Future Ethics -ryhmä
  - Tampere: Rose-projekti, Laitinen & Parviainen et al
  - Jyväskylä: Etiikan ryhmä, Vakkuri

- Julkinen: selonteot ja työryhmät, Tekoälyaika-kokonaisuus
- Organisaatiot: TIVIA, Millennial Board AI

- National Seminar of Theoretical AI (syksy 2018/kevät 2019)

TURKU AI SOCIETY

Kiitos!

aisociety.fi
juhovaiste.fi
@juhovaiste