Ville Vakkuri
Jyväskylän yliopisto

KV katsaus IEEE

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

IEEE on ottanut aktiivisen aseman tekoälyn (AI) ja autonomisten järjestelmien (AS) eettisesti kestävän suunnittelun ajamisessa. IEEE on kansainvälinen tekniikan alan järjestö ja yksi maailman suurimmista tekniikan alan järjestöistä. Laajimmin IEEE on tunnettu tieteellisen toiminnan edistämisen rinnalla standardien määrittelyorganisaationa. Eettisesti kestävän suunnittelun edistäminen tapahtuu kahden pääkanavan kautta: 1) Yhteisvalmisteltu visio Ethically Aligned Design – Version I ja Version II sekä 2) eettisen suunnittelun standardien valmistelu IEEE P7000 Standards Projects.

Ethically Aligned Design

Version I, julkaistu 2016 joulukuu, Version II, julkaistu 2018 toukokuu, Final Version, 2019?

Mukana valmistelussa yli 600 tutkimuksen, teollisuuden, kansalaisjärjestöjen ja valtioiden edustajaa.

Nykyinen versio II käsittelee teknologian hyötyjä sekä varoitusten kautta niitä haittoja, joita tekoälyn hyödyntämiseen liittyy. Näitä ovat muun muassa uhka yksityisyydelle, syrjintä, taitojen heikentyminen automaation myötä, taloudelliset vaikutukset, yhteiskunnan kriittisten järjestelmien turvallisuus, pitkän tähtäimen vaikutukset sekä yhteiskunnallinen hyvinvointi.

Keskeisen ajatus on se, että tekoäly ja autonomiset järjestelmät hyödyttävät parhaiten ihmiskuntaa, kun ne ovat linjassa, *aligned,* arvojen sekä eettisten periaatteiden kanssa.

Ethically Aligned Designin yleiset tavoitteet ovat:

- Jokainen tekoälyn sekä autonomisten järjestelmien suunnitteluun ja käyttöön osallinen olisi koulutettu, käytännössä harjoitellut sekä kyvykäs huomioimaan eettiset arvioinnit osaksi toteutusta siten, että ne tukevat ihmisten hyvinvointia
- Edesauttaa julkista keskustelua tekoälyn ja autonomisten järjestelmien eettisistä ja yhteiskunnallisista ulottuvuutta määrittelemällä arvoja sekä periaatteita, jotka ajavat ihmisisten hyvinvointia
- Tukea tekoälyn ja autonomisten järjestelmien standardien kehitystä laajemmin
- Ajaa standardointityössä esiteltyjä periaatteita osaksi kansallisia ja kansainvälisiä tekoälylinjauksia

Pääperiaatteet

- **Human Rights**: Ensure they do not infringe on internationally recognized human rights
- **Well-being:** Prioritize metrics of well-being in their design and use
- **Accountability:** Ensure that their designers and operators are responsible and accountable
- **Transparency:** Ensure they operate in a transparent manner
- **Awareness of misuse:** Minimize the risks of their misuse

Periaatteiden taustamotivaationa on toiminut seuraavat kolme linjausta:

1. Ihmisoikeuksien, ilmeisesti YK:n ihmisoikeusjulistuksen kunnioittaminen
2. Ihmisten ja luonnon hyvän priorisointi tekoälyn ja autonomisten järjestelmien käytössä
3. Riskien ja negatiivisten vaikutusten vähentäminen, ensisijaisesti varmistamalla tekoälyratkaisuiden ja autonomisten järjestelmien läpinäkyvyys sekä niihin liittyvät vastakysymykset.

Ville Vakkuri
Jyväskylän yliopisto

Poimintoja suosituksista:

- Tekoälyn käyttöä säätelevien regulaatioiden, standardien tulisi tukea ihmisoikeuksia, vapauksia, ihmiselämän arvokkuutta, yksityisyyttä ja tukea luottamuksen syntyä tekoälyratkaisuihin.
- Sääntelevien tahojen tulisi huomioida erilaiset kulttuuriset normit.
- Lähitulevaisuudessa tekoälylle ei tulisi myöntää ihmisenkaltaisia oikeuksia ja tekoäly tulee pitää ihmisen kontrollissa.
- Ihmisten hyvinvointi tulee huomioida sekä arvioida järjestelmäsuunnittelussa riippumatta millaisia ratkaisuja toteutetaan
- Lainsäätäjien tulisi määrittää ja tarkentaa, mitä tekoälyyn liittyviä vastuukysymyksiä on, jotta kehittäjät voisivat huomioida ne ennakoivasti suunnittelussa: Oikeuksien ja vastuiden ymmärtäminen.
- Kehittäjien tulisi huomioida kulttuurinen monimuotoisuus järjestelmien toteutuksessa
- Vastuukysymysten määrittämisessä työskentelyn tulisi huomioida kaikki tekoälyjärjestelmien käyttöön osalliset: käyttäjät, kehittäjät, viranomaiset, lainsäädäntö, kansalaisjärjestöt jne.
- Tekoälyratkaisuista tulisi koota rekisteri, jossa selviäisi:
    - Intended use, Training data/training environment, Sensors/real world data sources, Algorithms, Process graphs, Model features (at various levels), User interfaces, Actuators/outputs, Optimization goal/loss function/reward function
- Läpinäkyvyydelle (transparency) tulisi kehittää asteikko, joiden perusteella voitaisiin järjestelmien läpinäkyvyyttä mitata ja arvioida.
- Neliportainen transparency takaa järjestelmän toiminnan luotettavuuden
    - Käyttäjä: ymmärtää mitä järjestelmä tekee
    - Korjaaja: pystyy paikantamaan ja korjaamaan virheet
    - Lainsäädäntö: onnettomuustilanteissa voidaan löytää vastuulliset ja välttää virheisen toistumista
    - Yhteiskunta: järjestelmät saavuttavat riittävät yhteiskunnallisen hyväksyttävyyden
- Tekoälyn käyttöön liittyvien riskien tietoisuuden kasvattaminen
    - Etiikan ja tietoturvan lisääminen osaksi opetussuunnitelmia
    - Pelon lievittäminen faktoja tarjoamalla, asiantuntijanäkökulmat
    - Valtionhallinnon ja lainsäädännön tulisi huomioida oman asiantuntemuksen kehittäminen

Ville Vakkuri
Jyväskylän yliopisto

IEEE P7000 Standards Projects

IEEE P7000™ - Model Process for Addressing Ethical Concerns During System Design

IEEE P7001™ - Transparency of Autonomous Systems

IEEE P7002™ - Data Privacy Process

IEEE P7003™ - Algorithmic Bias Considerations

IEEE P7004™ - Standard on Child and Student Data Governance

IEEE P7005™ - Standard for Transparent Employer Data Governance

IEEE P7006™ - Standard for Personal Data Artificial Intelligence (AI) Agent

IEEE P7007™ - Ontological Standard for Ethically Driven Robotics and Automation Systems

IEEE P7008™ - Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems

IEEE P7009™ - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

IEEE P7010™ - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems