IMPULSSEJA Kesäkuu 2018

SUPERTEKOÄLYN SOSIAALIETIIKKAA

MARKKU VALTANEN

Supertekoäly on keinotekoisesti tuotettua älyä, joka ylittää ihmisen älykkyyden. Jossain vaiheessa se voi kyetä kehittämään itse itseään.

Supertekoälystä voi olla suureen hyvään, mutta siihen liittyy myös suunnattomia riskejä. Sosiaalietiikan näkökulmasta supertekoäly näyttäytyy normatiivisena erityistapauksena.

Edistyksellinen supertekoälypolitiikka tavoittelee hyötyjen oikeudenmukaista jakamista, riskien poistamista sekä sitä, että supertekoälyn aikakausikin on ihmisen aikakausi. Välineenä edistyksellisellä politiikalla täytyy olla kyky tuottaa katkoksia niihin yhteiskunnallisiin ja kulttuurisiin jatkumoihin, joilla teknologiaa nykymaailmassa kehitetään.

SISÄLTÖ

Kalevi Sorsa -säätiö on sosiaalidemokraattinen ajatushautomo, joka ylläpitää yhteiskunnallista, tasa-arvoa ja demokratiaa edistävää keskustelua.

JOHDANTO

Mikä on supertekoäly?

TEKOÄLY ON KONEELLE tuotettu kognitiivinen kyky, jolla se pystyy ratkaisemaan ongelmia. Supertekoälyksi voidaan kutsua tekoälyä, joka on ihmistä älykkäämpi. Sillä on tällöin kyky ratkaista mitä tahansa ongelmia, eli sen älykkyys on yleistä älykkyyttä eikä rajoitu vain tehtäviin, joita varten se on kehitetty. Tekoäly kykenee voittamaan ihmisen shakissa, mutta se ei osaa keittää kahveja vastustajalleen.

Supertekoälyn idea on puolisen vuosisataa vanha."[1] I. J. Good määritteli supertekoälyn vuoden 1965 kirjoituksessaan

> koneeksi, jonka älylliset kyvyt ylittävät nerokkaimmankin ihmisen kyvyt. Koska koneet on suunniteltu käyttämään älyä, äärimmäisen älykäs kone voisi suunnitella vieläkin älykkäämpiä koneita. Tästä epäilemättä seuraisi 'älyräjähdys' (intelligence explosion), ja ihmisen älykkyys jäisi siitä kauas jälkeen.[2]

Älyräjähdys, eli itseään toistava prosessi, jossa koneäly kehittäisi itsestään alati parempia versioita (esimerkiksi omaa lähdekoodiaan ohjelmoimalla) olisi kehityskulku, jonka seurauksena supertekoälyn älykkyyden taso voisi paitsi nousta kiihtyvällä tahdilla, myös karata inhimillisen käsityskyvyn ulkopuolelle.[3] Ajatus etenee, että lopputuloksena voisi olla pitelemättömäksi muuttunut, eli alkuperäisten kehittäjiensä kontrollista irtautunut keinoäly. Viimeisimmän varoituksen älyräjähdyksen riskisyydestä antoi Yhdysvaltain ilmavoimien tutkija John Geis tämän kirjoittamispäivän Helsingin Sanomissa: "Joku vuosi sitten puhuttiin, että tähän menee vielä 30 vuotta, mutta nyt on arvioita, että se tapahtuu jo 15 vuoden päästä."[4]

Kontrolliongelmaksi kutsutaan vaikeuksien vyyhtiä, joka liittyy haasteeseen pitää supertekoäly ihmisen hallinnassa. Kirjallisuudessa on esitetty monia kontrollin keinoja, jotka supertekoälypiirien tunnetuimpiin visionääreihin kuuluva Nick Bostrom[5] on jakanut kahteen ryhmään: supertekoälyä voidaan hallita ensinnäkin katkaisemalla siltä pääsy sen tarvitsemiin resursseihin (töpseli irti seinästä -idea), tai toiseksi sisällyttämällä sen ohjelmakoodiin estot tehdä koskaan mitään ihmisille haitallista.[6] Luottamus siihen, että itseään paremmaksi kehittävä keinotekoinen älykkyys ei milloinkaan muuttuisi hyvästä pahaksi, perustuu otaksumaan, ettei hyväksi ohjelmoitu älykäs intelligenssi koskaan valitsisi kehittää itsestään versiota, joka rikkoisi noita alkuohjelmansa arvoja.[7]

Kirjallisuudessa on useita ehdotuksia teknologioiksi, joiden varassa supertekoäly voitaisiin toteuttaa, mutta kaksi niistä näyttää muita uskottavammilta. Voitaisiin laatia riittävän kyvykäs tietokoneohjelma, joka kykenisi lopulta itsenäiseen älykkääseen toimintaan. Ohjelman kirjoittamisen haasteet tosin ovat periaatteellisia, sillä todellisuudessa emme vielä tiedä, mitä älykkyys oikeasti on.[8] Toinen vaihtoehto olisi mallintaa jo olemassa olevien älykkäiden rakenteiden toimintaa. Evoluutio on miljoonien vuosien aikana kehittänyt ihmisaivot, huiman monimutkaisen sähkökemiallisen järjestelmän, josta voisi olla malliksi koneälylle. Jos hermoverkon rakenne ja toiminta

1 Norvig & Russell 2016, 1037-8.

2 Good 1965, 33.

3 Bostrom 2014, 18-9, 38; Älyräjähdyksestä eli singulariteetista ks. Eden 2013.

4 HS 16.6.2018, A 23-24.

5 Bostrom on ruotsalaissyntyinen filosofi ja Oxfordin yliopiston professori. Hän johtaa yliopiston The Future of Humanity -instituuttia, jota oli perustamassa vuonna 2005.

6 Kontrolliongelmasta ks. Bostrom 2014, 129-.

7 Supertekoälyn päämäärärjärjestelmän teknisistä ja eettisistä kysymyksistä ks. esim. Yudkowsky 2001.

8 Bostrom 2014 35-39.

voitaisiin skannata tarpeeksi tarkasti ja mallintaa uudelleen tietokoneohjelmana, ei älykkyyden kaikkia periaatteita tarvitsisi periaatteessa keksiä itse, taikka edes täysin ymmärtää.[9] Molemmille mainituille teknologioille on yhteistä se, että tietokoneiden laskentatehon vaatimukset ovat hyvin suuria ja tyydytettävissä vain hyvin tehokkailla laitteilla tulevaisuudessa.

9 Bostrom 2014 39-44; muista supertekoälyn alustoista ks. 44-52. Aivojen mallintamisen kysymyksistä ks. Sandberg 2013.

SUPERTEKOÄLYN AIKAKAUSI MERKITSISI ENNENNÄKEMÄTTÖMIEN MAHDOLLISUUKSIEN, MUTTA MYÖS VALTAVIEN, JOPA IHMISKUNNAN OLEMASSAOLOA UHKAAVIKSI YLTYVIEN RISKIEN AIKAKAUTTA.

Supertekoäly astuu yhteiskuntaan

Yhteiskuntaan sijoitettuna supertekoäly olisi käänteentekevä teknologinen uutuus, joka aloittaisi uuden vaiheen ihmiskunnan historiassa. Supertekoälyn aikakausi merkitsisi ennennäkemättömien mahdollisuuksien, mutta myös valtavien, jopa ihmiskunnan olemassaoloa uhkaaviksi yltyvien riskien aikakautta.

Supertekoälyn kehittämisestä on potentiaalisesti hyvin suurta hyötyä. Tuotanto laajenisi ja tuotannolliset prosessit tehostuisivat niin, että inhimillinen puute lakkaisi. Sairauksiin saataisiin uusia hoitoja. Ihmisen kykyjä voitaisiin parannella. Kaikki asiat, missä koneälyn ylivoimaisista ongelmanratkaisukyvyistä olisi apua, hoituisivat nykyistä tehokkaammin. Jotkin utooppiset kuvaukset supertekoälyn potentiaalisista hyödyistä eivät kunnioita muita rajoja kuin fysiikan lait, mutta yhtä totta on sekin, että monet nykypäivän keksinnöistä olisivat muutama aika sitten eläneelle ihmiselle yhtä suuren epäuskon aihe kuin supertekoäly nyt meille. Jo tavanomaisen tekoälyn (jolla tarkoitan erityisiin tehtäviin ohjelmoitua monimutkaista automaatiota, ajankohtaisena esimerkkinä itseohjaavat autot), lupaukset ovat suuret, joten on helppo ymmärtää, mitä monet katsovat voitavan odottaa sitä kehittyneemmiltä koneälyn muodoilta.

Supertekoälyn riskit muodostuvat monesta lähteestä, eivätkä kaikki niistä suinkaan toista tieteiskirjallisuuden visiota ihmiskunnan robottiarmeijoillaan orjuuttavasta ilkeästä konemielestä. Supertekoälyn tuottamiseen sisältyy virheiden mahdollisuus, vaikkapa se, että älykästä koodia pakenisi tutkimuslaboratorioista. Supertekoälyn toiminnalla voisi myös olla tahattomia oheisvaikutuksia, esimerkiksi että se alkaisi kuluttaa runsaasti tietoteknisiä tai muita resursseja. Itseään alati kehittävä supertekoäly voisi myös matkalla päämääriinsä valita ihmiskunnalle epäedullisia välitavoitteita. Ja lopuksi on esitettävä kysymykset supertekoälylle ohjelmoitavasta moraalista. Eettisen harkintakyvyn ohjelmoiminen supertekoälylle ei ole periaatteellisestikaan yksinkertainen tehtävä, ei vaikka ihmiskunta olisi yksimielinen siitä, mitä koneelta halutaan. Uuteen kertaluokkaan supertekoälyn riskit kohoavat, jos sen potentiaali yhdistetään johonkin edistyneeseen tuotantoteknologiaan, kuten vaikkapa nanoteknologiaan.[10]

10 Supertekoälyn äärimmäisistä riskeistä, ks. esim. Yudkowsky 2008.

Yhteiskunnallisten suhteiden keskelle sijoitettuna supertekoäly näyttäytyy paitsi teknologisena artefaktina, jonka kanssa ihmiskunta joutuu järjestämään suhteensa, myös ihmisten välisiä suhteita määrittävänä instituutiona. Yhteiskunnalliseen todellisuuteen syntyvä supertekoäly on valtavan potentiaalin keskittymä, joka tarjoaa

suunnattomia määriä vaurautta ja valtaa haltijoilleen. Supertekoälyteknologian ensimmäisenä kehittävä taho saa huomattavan strategisen edun siitä, että saa tuon vallan ja vaurauden virtausten lähteen haltuunsa: Supertekoälyteknologian ensimmäisenä kehittänyt yritys omistajineen korjaisi siitä suunnattoman taloudellisen voiton. Supertekoälyn ensimmäisenä kehittänyt valtio tai valtioiden verkosto etenisi vastedes nopeammin kuin muut. Supertekoälyhankkeen ensimmäisenä maaliin vienyt sotilasmahti voisi halutessaan kääntää strategisen edun rajoittamattomaksi kyvyksi määrätä muiden kohtaloista väkivalloin.

Supertekoäly on siis tulevaisuuteen sijoittuva teknologinen innovaatio, jonka tulo todelliseksi on sekin epävarmaa. Tekoälytutkijoiden käsitystä siitä, milloin he uskovat ensimmäisen supertekoälyn tulevan kehitetyksi, on kysytty ja näkemykset ovat sekä hyvin kaukana toisistaan, että molemmissa päissä sitäkin varmempia. Keskiarvo sijoittuu yhtä kaikki jonnekin kuluvan vuosisadan puolivälin tienoille.[11] Mene ja tiedä.

Seuraavassa esittelen sosiaalieettisiä havaintojani, joita olen supertekoälystä käytyyn akateemiseen ja populaariinkin kirjallisuuteen tutustuessani tehnyt.[12] Nojaudun lähtökohtaan, että teknologiaa kehitetään ja käytetään aina jossain yhteiskunnallisessa ja kulttuurisessa tilanteessa, ja sen myötä suuntaan huomioni ennen muuta sellaisiin yhteiskunnallisiin ja kulttuurisiin jatkumoihin, joita pidän ongelmallisina supertekoälyn kehittämistä ajatellen. Perustelen tämän osoittamalla, että supertekoäly yhteiskunnallisena instituutiona on normatiivisessa mielessä erityistapaus. Lopussa piirrän ääriviivoja politiikalle, joka muodostaisi mielestäni vallitsevia sosiokulttuurisia ehtoja paremman lähtökohdan.

11 Bostrom 2014, 31-34; Supertekoälyasiantuntijoiden vastauksista heille esitettyyn milloin-kysymykseen, ks. Grace & Salvatier & Dafoe 2018.

12 Bostromin vuonna 2014 ilmestynyt teos Superintelligence. Paths, Dangers, Strategies on näkemyksellinen superälytematiikan kokonaisesitys, joka kokoaa yhteen kirjoittajansa ja monen muun tuottamia tutkimustuloksia. Teos on oiva lähtökohta supertekoälyn sosiaalietiikan lähtökohtien hahmotteluun ja sellaisena se palvelee tämän tarkastelun keskustelukumppanina.

Yleisesityksistä ks. myös edellä mainittu Norvig & Russell 2016 tai Kaplan 2016. Supertekoälyteorian skeptisismistä ks. esim. Boden 2016.

SUPERTEKOÄLYN NORMATIIVINEN LUONNE

Seurausten ainutlaatuinen mittakaava

MINKÄLAINEN ASIA SUPERTEKOÄLY on sosiaalieettisestä näkökulmasta? Minkälainen olemus keinotekoiselle intelligenssille syntyy, kun se asetetaan yhteiskuntaan? Supertekoälyn normatiivista luonnetta voi tehdä näkyväksi kysymällä, miten supertekoälyteknologia vaikuttaa hyvän ja pahan, tässä tapauksessa hyödyn ja riskin, jakautumiseen yhteisön jäsenten kesken.

Johdannossa tuli selväksi, että supertekoälyteknologian hyödyt ovat parhaimmillaan – ja sen riskit pahimmillaan – mittakaavaltaan tähtitieteellisiä. Silkka ennennäkemätön mittakaava tekee supertekoälystä erityislaatuisen yhteiskunnallisen tekijän. Supertekoälyteknologian kuviteltavissa olevat hyödyt ylittävät kaikkien edeltävien ja kenties myös kuviteltavissa olevien teknologioiden hyödyt. Tällöin on selvää, ettei mikään eettinen harkinta ajassa ole vielä voinut ottaa sen oikeudenmukaisen jakamisen kysymystä ratkaistavakseen. Koska supertekoälyteknologian potentiaaliset hyödyt ovat omaa mittaluokkaansa, niiden oikeudenmukainen distribuutio on kysymys, joka avautuu meidän aikanamme uutena.

Supertekoälyteknologian riskit ovat suurimmillaan eksistentiaalisia riskejä. Ne vertautuvat muihin ihmiskunnan olemassaoloa uhkaaviin riskeihin, mikä nostaa supertekoälyteknologian samaan luokkaan täysimittaisen joukkotuhoaseilla käytävän sodan taikka planetaarisen mittakaavan luonnonkatastrofin kanssa.

Potentiaalisten hyötyjensä osalta supertekoälyteknologia on siis omaa luokkaansa. Sille ei löydy historiallisia vastineita, ja lähin vertailukohta, kolmas teollinen vallankumous eli informaatio- ja viestintäteknologian tuleminen, jää sekin kauas sen alapuolelle. Riskiensä osalta se vertautuu muihin, olemassa oleviin globaalin mittakaavan uhkiin.

KOSKA SUPERTEKOÄLYTEKNOLOGIAN POTENTIAALISET HYÖDYT OVAT OMAA MITTALUOKKAANSA, NIIDEN OIKEUDENMUKAINEN DISTRIBUUTIO ON KYSYMYS, JOKA AVAUTUU MEIDÄN AIKANAMME UUTENA.

Hyötyjen jakamisen välttämättömyys

Edellä kuvatun supertekoälyteknologian luonteen vuoksi sen hyötyjen ja haittojen jako on välttämätöntä tehdä tietyllä tavalla. Jaolla tarkoitetaan tässä teknologian seurausten jakautumista yhteiskunnan jäsenten kesken ja siihen voidaan vaikuttaa tietoisesti. Jako-oikeudenmukaisuus tarkastelee hyötyjen ja taakkojen jakautumisen

perusteluja hyvän ja pahan, oikeuden ja velvollisuuden näkökulmista.

Ensimmäinen jako-oikeudenmukaisuuden näkökulmasta tehtävä havainto supertekoälyteknologian luonteesta on Bostromin huomautus, että kaikkien ihmisten voi katsoa altistuvan supertekoälyn riskeille. Miten tahansa ihmistä älykkäämpi kone toteutetaankaan, yhteiskuntaan sijoitettuna sen voi olettaa olevan vaikutuksiltaan niin läpitunkeva teknologia, että sen riskit koskettavat potentiaalisesti kaikkia. Yksin tämä riskien kompensaation vaatimus riittää perustelemaan jonkintasoisen ”laajan distribuution” eli supertekoälyn hyötyjen minimijaon kaikille ihmisille.[13]

Laajan jaon perusteluja on kahdentyyppisiä, prudentiaalisia eli järkiperäisiä sekä normatiivisia eli arvoperäisiä.[14] Erona on, että järkiperäisten perustelujen oletetaan sopivan kaikille aihetta rationaalisesti tarkasteleville, kun taas normatiiviset perustelut edellyttävät yksimielisyyttä niistä arvoista, joiden katsotaan olevan riittävän vahvoja oikeuttamaan jakoon puuttumisen. Tällainen erottelu ei ole ainoa mahdollinen, eikä yleensä tehtävissä kovin selvästi, mutta riittää käsillä olevan analyysin tekoon.

Laajan jaon järkiperusteet hyödyntävät oletusta, että supertekoälyteknologian hyötyjen ja haittojen jakautuminen vaikuttaa hankkeiden riskeihin. Laaja jako alentaa supertekoälyhankkeiden riskien tasoa, sillä supertekoälyhankkeiden riskisyys nousee suhteessa teknologiaa kehittävien osapuolten (yksilöiden, tutkimushankkeiden, yritysten, valtioiden) välille viriävän kilpailullisuuden tasoon.[15] Riskien hallinta on julkisen vallan velvollisuus, joten onnistuneen supertekoälyhankkeen tuottamien hyötyjen jaon laajentaminen järkiperäisesti perusteltua. Kilpailullisuuteen ja sen haitallisuuteen palataan vielä seuraavassa luvussa.

Edistyksellinen poliittinen positio harvemmin kaivaa esiin normatiiviset perustelunsa sille, miksi hyvää tulee jakaa uudelleen. Ainakin tässä maassa uudelleenjako periaatteena hyväksytään kaikkien eduskuntapuolueiden ohjelmassa, mittakaava tosin vaihtelee sitäkin enemmän. Vasta-argumentti uudelleenjaolle voi olla periaatteellinen, kieltäen sen moraalisen oikeutuksen tykkänään vetoamalla oikeudenmukaisuuteen reiluutena. Tällöin ajatellaan, että hyvän jakautumisen täytyy olla sensitiivinen sille, että ihmisten kunnianhimo ja ahkeruus on erilaista, joten toiminnan palkkioidenkin tulee vaihdella (myös) sen perusteella. Teknologian kehittämisen kohdalla reiluus olisi ehkä valmiutta tunnustaa teknologian kehittäjien ja sponsorien ensisijainen oikeutus sen tuottamiin etuihin. On oikein palkita toimeliaisuus sen hedelmillä, ja ilman palkinnon houkutinta ei siihen olisi ryhdyttykään.

Supertekoälyteknologian seurausten ainutlaatuinen mittakaava suhteellistaa argumentin oikeutusta olennaisesti. Reiluusvaade tyydyttyy Bostromin mukaan aina, koska supertekoälyn hyödyt, siis myös sen taloudelliset edut, ovat kertaluokaltaan ennennäkemättömän suuret.[16] Näin kohtuullistettu reiluusvaade ei jäisi tyydyttämättä, vaan se voitaisiin tehdä – jopa hyvinkin korkeana – sillä hyödyn jatkumon osalla, johon supertekoälyteknologian tuottamisessa onnistuneen osapuolen voi katsoa olevan moraalisesti oikeutettu. Palkkion absoluuttinen koko voitaisiin myös mitoittaa ordinaalisesti ylittämään jokin korkein siihenastisista teknologiahankkeista koituneen taloudellisen edun taso.

Jako-oikeudenmukaisuuden nimissä lausuttu reiluusvaade menettää tässä katsannossa oikeutuksensa ylempään osaan menestyneen supertekoälyhankkeen tuottamasta edusta. Supertekoälyn hyötyjen ja sen itsensä ainutlaatuisuus perustelee tämän eri tavoin. Yllä kuvattu hyötyjen mittakaava vie ajatuksen siihen suuntaan, että supertekoälyssä on tässä suhteessa kyse historiallisesti niin uudesta tilanteesta, että se tuottaa uuden tilanteen myös hyvän uudelleenjaon mielessä. Hyödyt ylittävät ensimmäistä kertaa rajan, jonka puitteissa ihmiskunta on tähän asti niiden jakamisesta neuvotellut, saatikka (kuvitteellisessa mielessä) sopinut. Voidaan myös argumentoida, että supertekoälyn tuottaman hyödyn avulla kenen tahansa ihmisyksilön toiveet ovat lopullisesti (jollain vaurauden tasolla) tyydytettävissä, mikä heikentää

13 Bostrom 2014, 291.

14 Bostrom 2014, 290-1.

15 Bostrom 2014, 288-290.

16 Bostrom 2014, 291.

yksityistä vaadetta kaikkeen superälyteknologiasta syntyvään etuun. Bostrom ilmaisee ajatuksen siten, ettei kukaan voi haluta koko universumia, jos kuitenkin saa oman galaksin.[17]

Entä onko keinotekoista älyä edes oikeutta omistaa? Voiko evoluution tuottamaa ilmiötä, älykkyyttä, rinnastaa teknologisiin luomuksiin, vai olisiko sen shareholder-omistajuutta syytä katsoa uudesta näkökulmasta? Samantyyppisin perustein on vaadittu muidenkin luonnonresurssien, kuten vaikkapa ihmisen genomin, omistajuuden säätelyä.[18]

Bostrom ehdottaa, että supertekoälyn kehittäjien tulisi itse alkaa sitoutua yllä kuvattuihin periaatteisiin ottamalla käyttöön vapaaehtoinen windfall-ehto, joka määrittelisi, miten onnistuvan hankkeen voitosta jaettaisiin tietyn rajan ylittävä osuus laajemmin. Hankkeen sponsorit saisivat reilun korvauksen, mutta sen ylittävän osuuden ne sitoutuisivat siirtämään osaksi laajaa distribuutiota. Ehdon käyttöönottaminen olisi helppoa, koska todennäköisyys onnistua tuottamaan älyräjähdykseen johtava keinotekoinen äly on yksittäisessä hankkeessa lähinnä sattuman mittaluokkaa. Valtioiden kohdalla kyseeseen voisi tulla sitoutuminen jakaa onnistuneen hankkeen hedelmiä maailmanyhteisön kanssa siinä vaiheessa, kun superälyteknologian kehittäjävaltion bruttokansantuote on hyötynyt jonkin sovittavan tason verran.[19]

Ihmiskunta on siis supertekoälyn aikakauden lähestyessä siirtymässä uuteen tilanteeseen: tämän teknologian vaikutukset hyötyjen ja taakkojen suhteelliseen jakautumiseen ovat ennennäkemättömät. Tilanne pakottaa yhteisön tarkastelemaan keskinäisiä eettisiä sitoumuksiaan uudelleen. Uudet sitoumukset täytyy tehdä ennen kuin meitä tulevaisuudesta erottava tietämättömyyden verho avautuu, ennen kuin on jo tietoa ensimmäisestä maaliinpääsijästä.[20] Teoreettisia sitoumuksia voidaan muotoilla poliittisen teorian piirissä ja sitä mukaa kun supertekoälyhanke menee eteenpäin, tuoda niitä asteittain lainsäädännön tasolle.

17 Bostrom 2014, 291.

18 Bostrom 2014, 89-90.

19 Bostrom 2014, 293.

20 Bostrom 2014, 293.

ON SELVÄT JÄRKIPERUSTEET SILLE, ETTÄ SUPERTEKOÄLYN POTENTIAALISTEN HYÖTYJEN KASAUTUMINEN HARVOJEN KÄSIIN OLISI HYVÄ ESTÄÄ.

Supertekoäly on normatiivinen erityistapaus

Supertekoälyteknologialla on siis normatiivinen erityisluonne. Tämä paljastuu peilaamalla sekä supertekoälyteknologiaa että sen kehittämishanketta jako-oikeudenmukaisuuden käsitteeseen. Supertekoälyllä on piirteitä, jotka tekevät siitä jako-oikeudenmukaisuuden näkökulmasta omaan luokkaansa kuuluvan erityistapauksen.

Supertekoälyteknologian erityisluonne perustuu sen potentiaalisten hyötyjen ja haittojen (riskien) tähtitieteelliseen mittakaavaan. Supertekoälyn potentiaaliset hyödyt ovat niin massiivisia, että niistä voidaan palkita kaikki hankkeen kehittämiseen osallistuneiden kohtuulliset tuotto-odotukset runsaskätisesti. On selvät järkiperusteet sille, että potentiaalisten hyötyjen kasautuminen harvojen käsiin olisi hyvä estää. Supertekoälyn riskit oikeuttavat hankkeen kehittämiseen puuttumisen ulkoapäin. Koska riskit ovat pahimmillaan eksistentiaalisia, tulee niiden toteutumisen todennäköisyys painaa mahdollisimman lähelle nollaa. Tämä velvoittaa riskeille altistuvan yhteisön puuttumaan asiaan tehokkaasti ja uskottavasti.

Supertekoälyn potentiaalisten hyötyjen mittakaava on niin suuri, että se antaa mahdollisuuden vaikuttaa kehittämishankkeen riskien alentamiseen ehkäisemällä

kilpailudynamiikan syntyä. Yhteistyö näyttäytyy välttämättömänä ehtona kilpailudynamiikalle. Yhteistyönä etenevän kehittämishankkeen normatiivinen maksiimi voisi olla tuottaa supertekoälyteknologiaa siten, että hanke lisää yhteisön luottamusta supertekoälyn aikakauteen enemmän kuin mitä sen vaihtoehdot tekisivät. Tähän ideaan palataan seuraavassa luvussa.

SUPERTEKOÄLYN TUOTTAMISEN HAITALLISET JATKUMOT

TEKNOLOGIAA TUOTETAAN AINA kulloisessakin yhteiskunnallisessa ja kulttuurisessa ympäristössä, eikä supertekoälyteknologia ole poikkeus. Minkälainen synnyttäjä aikakautemme on supertekoälyn kaltaisille kehittyneille teknologioille ja mitä se merkitsee sosiaalieettisestä näkökulmasta?

Lähtökohtanani on väite, että supertekoälyteknologian tuottaminen nykyisten yhteiskunnallisten suhteiden ja kulttuuristen arvostusten vallitessa lisää riskiä, että supertekoälyn hyödyt jäävät toteutumatta ja riskien todennäköisyys kasvaa.

SUPERTEKOÄLYTEKNOLOGIASSA SAATTAA OLLA LIIKAA VALTAPOTENTIAALIA, JOTTA SITÄ OLISI LUPA PITÄÄ KOVIN HELPPONA NEUVOTTELUN KOHTEENA.

Valheellinen turvallisuudentunne

Joillakuilla on taipumus suhtautua tulevaisuuteen vastaansanomattomalla huolettomuudella. Jos tutustuu kaikkeen, mitä supertekoälystä rinnakkaisideoineen on spekuloitu, voi helposti ajautua pitämään puhetta supertekoälyn aikakaudesta silkkana hupailuna, joka kelpaa korkeintaan sci-fi -viihteen innoitukseksi. Tästä asenteesta ei paljoa eroa suhtautumistapa, jossa supertekoälyn riskeistä puhumista pidetään tarpeettomana, koska yksinkertaisesti luotetaan siihen, että teknologinen tutkimus kuitenkin etenee riittävän hyväntahtoisten intressien kannattelemana. Siksi teknologisen tutkimuksen ei katsota kaipaavan erityistä kaitsentaa eettisestä näkökulmasta. Jonkun huolettomuutta voi ylläpitää ajatus, että keinotekoisen älyn hyväntahtoisuus olisi myönteisellä tavalla riippuvaista sen intelligenssin asteesta.

Periaatteellinen ongelma piilee siinä, ettei tätä oletusta voi tehdä. Bostrom katsoo kumppaneineen osoittaneensa, että hyväntahtoisuutta ei voida pitää älykkyydestä riippuvana ”suureena”, toisin sanoen koneälyjä voi eettiseltä asenteeltaan olla (periaatteessa) kaikenlaisia, olipa niiden intelligenssin taso mikä tahansa. Jos tämä teesi pitää paikkansa, ihmiskunta voi periaatteessa onnistua luomaan eettiseltä ohjelmaltaan myös ilkeän supertekoälyn taikka supertekoälyn, joka päätyy valitsemaan lopullisiksi tavoitteikseen ihmiselle epäsuotuisia päämääriä.[21] Ihmiselle vihamielinen tai haitallinen supertekoäly on siis loogisesti mahdollinen, mikä tekee kestämättömäksi lähtökohtaisen luottamuksen keinotekoisen älyn suopeaan asenteeseen meitä kohtaan.

Poliittisen ilmaisunsa huolettomuus saa lähinnä tietynlaisessa naiivissa liberalismissa, jossa yhteiskunnallinen neuvottelutilanne mielletään sarjaksi rationaalista näkökantojen vaihtoa, joka etenee kohti alati laajenevaa tukea yhteisesti jaetuille periaatteille, ja jossa neuvottelijat tulevat toinen toistaan puoliväliin vastaan.

21 Tästä nk. ortogonaalisuusteesistä ks. Bostrom 2012; Armstrong s.a.

Supertekoälyteknologiassa saattaa olla liikaa valtapotentiaalia, jotta sitä olisi lupa pitää kovin helppona neuvottelun kohteena. Sotateollisuussektori on yksi johtavista tavanomaisen tekoälyn kehittäjistä, mikä jo itsessään ilmentää sopimusparadigman poissaoloa. Onneksi on myös yrityksiä sotilaallisen tekoälyn kehittämisen kieltämiseen. Silti näyttää selvältä, että neuvottelua supertekoälyteknologiasta joudutaan käymään masentavan tietoisina siitä, että sen (väki)valtapotentiaali kohteen muun riskisyyden ohella tekee siitä vaativan neuvoteltavan.

Edellä mainitusta seuraa ensimmäinen huolestuttava havainto nykyajasta supertekoälyn tuottajana. Maailmanyhteisön neuvotteluinstituutiot ja niiden nykytila ei anna aihetta suurelle optimismille sen suhteen, että olisi tarjolla foorumia, jossa näkökantoja voitaisiin vaihtaa rationaalisesti ja jolta voitaisiin odottaa sitovia kantoja tai varsinkaan niiden tehokasta toimeenpanoa. Jos yksityisten toimijoiden sekä valtioiden supertekoälyhankkeet käynnistyvät tämäntapaisessa institutionaalisessa kontekstissa, näyttää huolestuttavasti siltä, että supertekoälyn riskejä ei tulla minimoimaan siinä määrin kuin kyseisen teknologian kohdalla olisi välttämätöntä.

Edelleen, vaikka supertekoälyn tuottaneiden tahojen pyrkimys olisi ollut hyvä, ei ole takeita siitä, että tuotetun keinotekoisen älyn eettinen asenne ihmiskuntaa kohtaan olisi toivotun kaltainen. Optimismi sen suhteen, että keinotekoinen älykkyys ollessaan sillä asteella, jolla se kykenee kehittämään itse itsestään aina parempia versioita (esimerkiksi lähdekoodiaan ohjelmoimalla), valitsisi kaikissa tilanteissa pitää eettisen ohjelmansa ihmiskunnalle suopeana, tukeutuu parhaimmillaan oletukseen sen ”tahdon” koherenssista. Ajatellaan, ettei tekoäly milloinkaan valitsisi parannelluksi versioksi itsestään sellaista muunnelmaa, jonka eettinen ohjelma olisi vastoin sille sitä ennen syötettyä normia.

Huolettomuus ja naiivius ilmiön edessä, joka on eettiseltä olemukseltaan näin monitahoinen ja kontingentti, on huolestuttava asennoitumistapa. Tämän ja muiden vastaavien asenteiden siirtyminen jonain päivänä käynnistettävien supertekoälyhankkeiden aikakauteen on omiaan tuottamaan epätoivottavia tulevaisuudennäkymiä. Voi kuitenkin olla, että tavanomaisen tekoälyn teknologiset edistysaskeleet tuottavat aikanaan hyödyllisiä herätteitä.

Kilpailu vie turvallisuuden pohjalle

Toinen epätoivottava jatkumo, jonka varassa supertekoälyn kehittämisen ei soisi toteutuvan, on kilpailuparadigma. Tällä tarkoitan ihmisten välisen yhteistoiminnan mallia, jossa tavoitteeseen pyritään omin panostuksin ja palkkiona muita paremmasta menestymisestä on yksityinen voitto. Uskomuksena on, että kilpailulla on suotuisia vaikutuksia tehokkuuteen ja resurssien allokaatioon. Talouden alueella sen institutionaalinen muoto on markkinat, ja teknologisten innovaatioiden alueella kilpailullisuuden tavallisesti ajatellaan parantavan tuotteiden laatua ja nopeuttavan niiden tuottamista. Kilpailullisuuden olennainen merkitys on houkutella teknologisten kehityshankkeiden taakse riittävästi riskiä sietävää rahoitusta.

Kilpailuparadigma ei muodosta ainoastaan liike-elämän toiminnan institutionaalisia puitteita, vaan myös valtiot ymmärtävät itsensä toinen toistensa kilpailijoiksi. Kylmän sodan jälkeen kansallisvaltioiden välinen ideologinen kilpailu hellitti, mutta optimismi on saanut uudella vuosituhannella usein väistyä pelkojen tieltä. Viime vuosina monen mielessä on käynyt epämiellyttävä aavistus siitä, että liberaalit demokratiat ovat tulevaisuudessa kenties nykyistä vähäväkisempi joukko maailmanyhteisössä. Yhtä kaikki valtioiden välinen luottamus tuskin riittää voittamaan valtioiden tai valtioryhmien keskinäistä epäluuloa nykymaailmassa.

Olipa kyse yritysten tai valtioiden välisistä suhteista, ja olivatpa ne institutionaalisesti asetettuja (markkinat) tai poliittisia/kulttuurisia (kansainväliset suhteet), voidaan kilpailua pitää vallitsevana paradigmana yhteistoiminnan järjestämi-

sessä. Sellaisenaan se nauttii myös syvään juurtunutta kulttuurista hyväksyntää. Supertekoälyn tapauksessa kilpailullinen paradigma on kuitenkin syvästi ongelmallinen. Turvallisuusnäkökohdat tulisi huomioida laajalti ja pettämättömästi, mutta kilpailun osapuolten näkökulmasta voi syntyä tilanteita, joissa niiden kannattaa siirtää turvallisuusinvestoiteihin suunnattuja resursseja tutkimustoimintansa nopeuttamiseen. Tämä seuraa siitä peliteoreettisesta havainnosta, että supertekoälyteknologian kehittämiskilvan osapuolten voi olla rationaalista vähentää investointeja turvallisuuteen saavuttaakseen itselleen etuja kilvassa.[22]

Kilpailullinen paradigma supertekoälyhankkeen toteuttamisen alustana on siis hyvin riskialtis, mikä on vahva peruste sen torjumiselle. Riskien valtavasta mittakaavasta huolimatta supertekoälyhankkeen käynnistyminen kilpailuna yritysten ja valtioiden välillä voi olla todennäköistä. Se johtuu siitä, että supertekoälyhanke luultavasti ja luonnollisesti on alkava tavanomaisen tekoälyn tuottamisen jatkumona, sillä tuottajien näkökulmasta näiden kahden teknologian välinen ero ei välttämättä ole merkityksellinen.

22 Bostrom 2014, 288-90, Box 13 A; Armstrong 2013.

SUPERTEKOÄLYÄ EI SAA JÄTTÄÄ VAIN ONNEN, KILPAILUN TAI INSINÖÖRIEN VARAAN.

Vastustamattomat välineet

Kolmantena haitallisena jatkumona pidän sellaista kulttuuris-ideologista asennoitumista teknologiaan, jota nimitetään teknokraattiseksi. Inhimillisesti on kyse ihastumisesta teknologisten välineiden mahdollisuuksiin, ideologisesti on kyse suhtautumisesta teknologiaan välineenä ja (arvo)päämäärien sulkeminen rationaalisen harkinnan ulkopuolelle. Asennoituminen ilmenee syvässä luottamuksessa insinööritieteisiin sekä siinä tosiasiassa, ettei insinööreille – tekoälyn tapauksessa keskeisimmille eettisille toimijoille – juurikaan opeteta etiikkaa.

Supertekoälyn kohdalla teknokraattinen rationaalisuus pukeutuu usein kahteen ajattelutapaan. Ensinnäkin eettinen harkinta kutistetaan siinä usein pelkän riskin punnitsemiseksi, jolloin ihmistä korkeamman älyn historiaan astumisen muista seurauksista, seurauksista ihmisyydelle tai ihmisen vapaudelle, ei kanneta enempää huolta. Toiseksi teknokraattinen rationaalisuus huomioidessaan teknologisten tuotosten käytön välineenä eristää itsensä niistä päämääristä tai arvoista ja intresseistä, joita tuohon välineeseen on kirjoitettu sisään. Tällöin mikä tahansa teknologinen laite, joka kyetään tekemään olevaksi, täytyy rakentaa, uhraamatta ajatusta sille, tarvitaanko sitä tai haluaako maailma nähdä sen syntyvän, – tai minkälainen yhteiskunnallinen tai kulttuurinen todellisuus tarvittaisiin, jotta ihmiskunta olisi valmis siirtymään supertekoälyn aikakauteen.[23]

Supertekoälyä ei siis saa jättää vain onnen, kilpailun tai insinöörien varaan. Seuraavassa luvussa vastaan näiden haitallisten jatkumoiden haasteeseen ja muotoilen lähtökohtia edistykselliselle supertekoälypolitiikalle.

23 Erilaisista rationaliteettikäsityksistä sekä teknokraattisen rationaliteetin kriisistä ks. Feenberg 2017.

EDISTYKSELLISEN SUPERTEKOÄLYPOLITIIKAN ÄÄRIVIIVOJA

Onko supertekoäly pakko kehittää?

ON VÄITETTY, ETTÄ ihmiskunta siirtyy supertekoälyn aikakaudelle joka tapauksessa jossain vaiheessa. Joku kehittää superteknologian vääjäämättä, joten tarvitaan edistyksellisten voimien supertekoälyhanke ehättämään tekemään se ensin. Ajatus ajolähdöstä saa tiettyä tukea supertekoälyn erityisluonteesta, siihen liittyvistä riskeistä ja ennen muuta strategisesta edusta, jonka teknologian ensimmäinen tuottaja saisi. On perusteita väittää, että kaikki tärkeät saavutettavissa olevat teknologiset keksinnöt toteutetaan joskus, eikä ole olemassa keinoa tämän estämiseen, ellei siksi lueta kaiken tieteellisen ja teknologisen kehityksen pysäyttämistä. Supertekoälyteknologian potentiaalia voidaan myös käyttää muiden ihmiskuntaa uhkaavien riskien, esimerkiksi kehittyneen nanoteknologian haitallisen tai pahantahtoisen käytön torjumisessa.[24]

24 Bostrom 2014, 278-9.

Näin perusteltua imperatiivia supertekoälyteknologian kehittämiselle ei tule sekoittaa teknokraattiseen pakkomielteeseen tehdä olevaksi kaikki teknologinen, mikä tulee mahdollisen piiriin. Jos supertekoälyhanke käynnistettäisiin edistyksellisin periaattein ja aikein, se voisi tapahtua vain supertekoälyn tuottaman ihmisen hyvän vuoksi ja pahan torjumiseksi. Edistyksellisen supertekoälyteknologian on siis ensimmäisenä pystyttävä perustelemaan itselleen, mitä päämääriä sillä halutaan ajaa ja minkälaista yhteiskunnallista ja kulttuurista kehitystä tukea. Supertekoälyhankkeeseen ei tule lähteä ilman, että se perustellaan oikeilla päämäärillä.

Edistyksellisen supertekoälypolitiikan hallittava aikaa

Tavanomaisen tekoälyn kehittämisen aikamuoto alkaa olla preesens, mutta supertekoälyn aikamuoto pysynee vielä pitkään futuurina. Edistyksellisen politiikan on hallittava sekä supertekoälyhankkeen alkamista edeltävää aikaa että siirtymistä sen aikakauteen.

Supertekoälyn riskien hallinta ajan hallintana merkitsee sitä, että edistyksellisen politiikan on hankittava keinot mahdollisesti viivästyttää supertekoälyn tuottamista maailmaan. Kontrolliongelmaan on oltava ratkaisu ennen kuin mikään taho saa tuottaa aidosti ihmistä älykkäämmän koneen. Tämä on välttämätön ehto teknologian turvallisuuden takaamiseksi. Jos mikä tahansa hanke on lähestymässä maalia ilman, että kontrolliongelma on ratkaistu, yhteisöllä on oltava keinot jarruttaa tällaisen hankkeen etenemistä ja edesauttaa toivottujen hankkeiden valmistumista.[25]

25 Bostrom 2014, 278-9.

Edistyksellisen supertekoälypolitiikan aikaan kohdistuvaa hallintaa tarvitaan vastavaikuttamaan niihin epätoivottaviin jatkumoihin, joiden varassa supertekoälyteknologian tuottaminen ei ole suotavaa. Edistyksellisen politiikan nimissä ja keinoin on haastettava niitä jatkumoita, jotka jättävät supertekoälyhankkeen hallinnan vain onneen tai kaikkien osapuolten hyväntahtoisuuteen luottamisen varaan, eivät ymmärrä sen normatiivisesta erityisluonteesta nousevia antiteesejä kilpailulliselle paradigmalle tai jotka kehystävät sen kaiken muun teknologian tavoin vain

välineeksi muiden välineellisten arvojen, kuten taloudellisen voiton, tavoitteluun. Edistyksellisen supertekoälypolitiikan on kyettävä tuottamaan katkoksia niihin haitallisiin ei-sosiaalisiin jatkumoihin, joiden varassa supertekoäly on muuten astuva historiaan.

TÄRKEIN TEHTÄVÄ ON VALMISTAA YHTEISKUNNALLISTA TODELLISUUTTA SIIRTYMÄLLE SUPERTEKOÄLYN AIKAKAUTEEN.

Hallitulle siirtymiselle luotava edellytykset

Koska supertekoälyteknologia kehitetään aikanaan sosiaalisessa ja kulttuurisessa kontekstissa, siihen tullaan muiden teknologisten tuotosten tavoin kirjoittamaan sisään aikakautensa edustajien intressit ja arvot. Supertekoälyn tapauksessa tämä on erityisen konkreettista, sillä keinotekoisen intelligenssin ohjelma on sisältävä myös sen moraalin, toisin sanoen apparaatin, joka eettisin perustein suuntaa älyn toimintaa. Tämä tuottaa supertekoälyhankkeeseen erityispiirteen, että siinä joudutaan operoimaan etiikan sisältökysymysten kanssa hyvin soveltavalla tasolla: mitkä arvot keinotekoiselle intelligenssille asetetaan, minkälainen toimija keinoly saa etiikkansa suhteen olla, mitä etiikka pohjimmiltaan on ja onko eettinen harkinta tai kokemus edes mallinnettavissa. Tämän käsitteellisen selvitystyön keskeneräisyydestä huolimatta on selviö, että teknologisen kehittämisen ohella myös eettinen komponentti on otettava olennaisena osana mukaan, kun kehitetään keinotekoista älyä, joka tulevaisuudessa tekee moraalisesti merkitseviä tekoja sosiaalisessa todellisuudessa.

Sen varmistaminen, että kaikista kehitettävistä supertekoälyteknologioista tulee moraalinsa puolesta ihmisen kannalta suopeita, on kuitenkin vain osa edistyksellistä supertekoälypolitiikkaa. Tärkein tehtävä on valmistaa yhteiskunnallista todellisuutta valmiimmaksi siirtymälle supertekoälyn aikakauteen. Bostrom on tunnustautunut optimistiksi nimeämällä muutamia avaintrendejä, jotka ovat hänen mielestään lupaavia nimenomaan tuottamaan edellytyksiä sille, että supertekoälyhankkeen lopputulos olisi myönteinen.[26] Voidaan ajatella, että edistyksellisen politiikan tehtävänä olisikin analysoida tarkemmin niitä yhteiskunnallisia ja kulttuurisia ehtoja, joiden voi olettaa luovan yleisiä edellytyksiä supertekoälyn aikakauteen kohdistuvien toiveiden toteutumiselle.

26 Bostrom 2014, 279.

Kun supertekoälyn aikamuoto on futuuri, kenties kaukainen futuuri, on maailmaa valmistettava sen aikakauteen hyvin pitkäjänteisellä, kauas ajassa kurottavalla ja ylisukupolvisella politiikalla. Silti uuden teknologian kehitys- ja käyttöönottotahdin kiihtyessä ja tavanomaisen tekoälyn taitojen jo ällistytellessä ihmisiä tulee supertekoälypolitiikalle alati uusia tilaisuuksia tarkentaa arviota kohteestaan ja laajentaa tavoitteistoaan. Pitkään aikaväliin voi sisällyttää myös suurta optimismia. Kenties tämä aika pystytään käyttämään yhteiskunnallisten konfliktien ja antagonismien perusteelliseen purkamiseen, ja kun samalla ymmärrys supertekoälyteknologian erityisluonteesta etenee, voi ihmiskunta ehkä olla riittävän valveutunut siirtymään seuraavaan todella suureen teknologiseen vallankumoukseen nykyistä valveutuneempana ja yksimielisempänä.

Pehmeiden ideaalien kova merkitys

Sosiaalieettisesti supertekoäly on kiintoisa sen vuoksi, että se näyttää normatiivisen erityisluonteensa vuoksi tuottavan sävynmuutoksen eräisiin yleisiin käsitteisiin. Kun esimerkiksi kilpailuparadigman tuottaman kilpailullisen dynamiikan potentiaalisesti haitallinen ja vaarallinen vaikutus supertekoälyteknologian riskien toteutumiselle

tunnistettiin, näyttäytyi yhteistyöparadigma suositeltavampana tapana supertekoälyhankkeen osapuolten keskinäisten suhteiden järjestämiseen. Yhteistyö muuttuu siis valinnaisesta ideaalista välttämättömäksi keinoksi varmistaa luottamuksen säilyminen hankkeen osapuolten kesken ja vahvistaa sitä. Luottamuksen ideaali muuttuu juhlapuheiden pehmeästä itsestäänselvyydestä ”kovaksi” osaksi supertekoälyteknologian turvallisen tuottamisen ehtoja, jota ilman ihmiskunnalla ei ole varaa realisoida tätä teknologiaa.

Edistyksellisen supertekoälypolitiikan keskiössä on siis luottamuksen ideaalin tuominen osaksi keinotekoisen älykkyyden tuottamishanketta. Bostromin pystyttämä maksiimi kaikille pyrkimyksille tuottaa ihmistä älykkäämpi kone sisälsi velvoituksen tavoitella niillä ainoastaan sitä, mikä koituu kaikkien yhteiseksi hyväksi (the common good)[27]. Tämän yhteyteen luottamuksen ideaali sopeutuu hyvin, säätelemään yhteiskunnallisia suhteita aikana, jona tuo teknologia kenties tuotetaan.

27 Bostrom 2014, 293.

KIRJOITTAJA

MARKKU VALTANEN, TM, HuK, on Kalevi Sorsa -säätiön projektitutkija. Hän valmistelee sosiaalietiikan väitöskirjaa Manuel Castellsin verkostoyhteiskuntateoriasta.

KIRJALLISUUS

Armstrong, S. (2013). Racing to the Precipice: a Model of Artificial Intelligence Development. Technical Report #2013-1, Future of Humanity Institute, Oxford University, 1-8.

Armstrong, S. (s.a.). General Purpose Intelligence: Arguing the Orthogonality Thesis. Future of Humanity Institute, University of Oxford. *https://www.fhi.ox.ac.uk/wp-content/uploads/Orthogonality_Analysis_and_Metaethics-1.pdf* (19.6.2018).

Boden, M. A. (2016). AI: Its nature and future. Oxford: Oxford University Press.

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. Future of Humanity Institute, Oxford University. *https://nickbostrom.com/superintelligentwill.pdf* (19.6.2018).

Bostrom, N. (2014). Superintelligence. Paths, Dangers, Strategies. Oxford: Oxford University Press.

Eden, A. H. (2013). Singularity Hypotheses: A Scientific and Philosophical Assessment. Berlin, Heidelberg: Springer-Verlag.

Feenberg, A. (2017). Technosystem. The Social Life of Reason. Cambridge, Massachusetts & London: Harvard University Press.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. Teoksessa F. L. Alt & M. Rubinoff, Advances in Computers. Vol. 6, s. 31-88. Academic Press.

Grace, K. & Salvatier, J. & Dafoe, A. et al. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. *https://arxiv.org/pdf/1705.08807.pdf* (19. 6 2018).

Helsingin sanomat 19.6.2018. Tekoäly pelottaa tulevaisuuden tutkijaa. Talous A 23-24.

Kaplan, J. (2016). Artificial Intelligence: What Everyone Needs to Know. Oxford: Oxford University Press.

Norvig, P. & Russell, S. J. (2016). Artificial Intelligence. A Modern Approach. Third Edition. Essex: Pearson Education Limited.

Sandberg, A. (2013). Feasibility of Whole Brain Emulation. Teoksessa V. C. Müller, Theory and Philosophy of Artificial Intelligence (s. 251-64). Berlin: Springer. *https://pdfs.semanticscholar.org/375b/1b4540121f6be993591d0e5ba43cb84708f1.pdf* (19.6.2018)

Yudkowsky, E. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. San Francisco, CA: The Singularity Institute. *https://intelligence.org/files/CFAI.pdf* (19.6.2018).

Yudkowsky, E. (2008). Artificial Intelligence as a positive and a negative factor in global risk. Teoksessa N. Bostrom & M. M. Cirkovic, Global Catastrophic Risks (s. 303-339). Oxford: Oxford University Press.

Impulsseja-sarjan uusimpia julkaisuja

NIEMI, Matti: 'EU:n tuleva rahoituskehys: Nettomaksajan kirous' Kesäkuu 2018

BORDI, Laura & OKKONEN, Jussi: 'Informaatioergonomian näkökulma tietotyöhön' Toukokuu 2018 (23 s.)

OJANEN, Hanna: 'Suomen EU-puheenjohtajuus: mitä kolmas kerta sanoo?' Huhtikuu 2018 (23 s.)

ELOMÄKI, Anna: 'Feministisempää talouspolitiikkaa. Seitsemän askelta kohti tasa-arvoa edistävää ja syrjimätöntä talouspolitiikkaa' Maaliskuu 2018 (39 s.)

SAPP, Will: 'Onko Justin Trudeau edistyksellisten arvojen globaali valopilkku?' Maaliskuu 2018 (11 s.)

HOLMGREN, Markus: 'Tietön taival. Mikä on Kiinan Uusi silkkitie, mitä sillä tavoitellaan ja mitä sen toteutuminen edellyttää?' Helmikuu 2018 (37 s.)

MATTILA, Maija: 'Alustatalouden haasteet työntekijälle' Tammikuu 2018 (35 s.)

HUUPPONEN, Mari: 'Ruotsin feministinen ulkopolitiikka' Joulukuu 2017 (12 s.)

HONKANEN, Petri: 'Lohkoketjuteknologia – Luottamuksen koodi hajautuneessa yhteiskunnassa' Lokakuu 2017 (31 s.)

MEYER, Henning: 'Poliittisia vastauksia digitaalisen vallankumouksen haasteisiin' Syyskuu 2017 (12 s.)

MUSTOSMÄKI, Armi: 'Pohjoismainen työmarkkinamalli digipaniikin aikakaudella' Kesäkuu 2017 (27 s.)

WINGBORG, Mats: 'Ruotsin terveydenhuollon uudistukset ja niiden vaikutukset' Maaliskuu 2017 (14 s.)

BLÅFIELD, Ville: 'Uusi työ – uudet duunarit. Keskusteluja työn muutoksesta' Helmikuu 2017 (52 s.)

KALEVI SORSA SÄÄTIÖ