Tartu Ülikool
Filosoofia ja semiootika instituut

Superintellekti väärtuste joondamise probleem

Bakalaureusetöö filosoofias

Karl Vilu

Juhendaja: Bruno Mölder

Tähemärkide arv: 67322

2024

Sisukord

Sissejuhatus

Tehnoloogia areng viimasel sajandil on olnud ülikiire. Näiteks veel kakskümmend aastat tagasi olid kõigil laialtlevinud mobiiltelefonidel nupud ja väiksed ekraanid, ajapikku loodi mobiiltelefonidele muudkui suuremaid ekraane ja muudeti neid targemaks. Tänasel päeval on suur ekraan üks nutitelefoni peamistest omadustest, lisaks sellele on nutitelefonides erinevaid tööriistu, mis aitavad inimesi igapäevaelus. Mõned nendest tööriistadest kasutavad tehisintellekti (TI), mida käesolev bakalaureusetöö laiemalt käsitlema hakkab. Täpsemalt käsitleb bakalaureusetöö tehisintellekti edasiarenenud versiooni, superintellekti (SI), ja selle väärtustega seonduvaid probleeme. Seda, mis on TI, selgitan lugeja jaoks bakalaureusetöö esimeses osas põhjalikumalt, kuid lühidalt kokku võttes on see arvutitarkvara, mis on võimeline jäljendama inimese vaimset tegevust, sealhulgas näiteks analüüsivõimet ja eneseväljendust.

Probleemseks võib tehisintellekt muutuda juhul, kui selle võimekus ületab kõige targemate inimeste intellekti. TI intellektuaalne võimekus võib plahvatuslikult kasvada tasemele, kus see on kordades targem kõige targematest inimestest üldse – sellest saab superintellekt. Seda, kuidas niisugune protsess juhtuda võib, kirjeldan täpsemalt peatükis „Superintellekt", kus kirjutan SI tehnilisest saavutamisest, aga ka selle ohtudest.

Superintellekt võib olla inimeste suhtes ükskõikne või isegi vaenulik, mistõttu on sellel vaja teatavaid väärtuseid, mis oleks inimeste ja nende eelistustega kooskõlas. Kirjeldan lähemalt superintellekti väärtustega seonduvaid mõtteid ja probleeme, näiteks Yudkowsky koherentset ekstrapoleeritud tahtlust, neljandas peatükis, „Superintellekti väärtused". Näiteks on levinud idee, et kui inimkond jõuab superintellektini, siis võiks SI ise oma väärtused välja mõelda. Küll aga toon esile vastupidise mõtte Seth Baumilt, kes leiab, et SI ise ei saa olla eetikateooria autor, sest eeldab, et superintellektil on juba teatud teadmisi moraalist, mis aitavad sellel niisugusele probleemile läheneda.

Pärast Baumi lähenemist analüüsin Margit Sutropi lähenemist tehisintellekti väärtuste pluralismist, mis loob kasulikku konteksti väärtuste joondamise probleemile. Sutropi artikkel hõlmab seda, et TI väärtuste joondamisel peaks arvestama pluralismiga, mis võiks ühendada erinevaid lähenemisi, sest eetikaküsimustele ei ole sageli ühte objektiivset vastust. Samuti toon välja edasisi probleeme superintellekti väärtuste joondamisel. Liigun sealt edasi analüüsini, kus teen vahekokkuvõtte käsitletud materjalidest joondamise probleemiga seoses ja kirjeldan

omapoolset lähenemist kui lahendust väärtuste seadistamisele. Samuti toon esile potentsiaalse kriitika oma seisukohtadele.

Oma töö peamise teesina toon esile selle, et superintellekti väärtuste kujundamise senistel lähenemistel on teatud puudujääke. Ma püüan seda parandada omapoolse mõtteeksperimendiga, kus erinevad tehisintellekti kasutajad, aga ka sellealased spetsialistid saavad osaleda tehisintellekti väärtuste joondamisel analüüsiva küsitluse abil.

Tehisintellekt

Eesti Võõrsõnade leksikon defineerib tehisintellekti kui tehislikku mõistust, mis on võimeline maailma tunnetama inimesele omaselt. See on arvuti, millele on modelleeritud ajuprotsesside tunnusjooned, mille tulemusena on see võimeline jäljendama inimese vaimset tegevust (Eesti Keele Instituut 2024).

Euroopa Parlament defineerib tehisintellekti kui masinate võimekust inimestele omaseid võimeid näidata, näiteks tajub TI oma ümbrust, töötleb tajutut ning lahendab probleeme, et saavutada teatud eesmärke. TI on võimeline oma käitumist kohandama, kui see analüüsib eelnevaid tegevusi (European Parliament 2020).

Tehisintellekti „tarkuse" taset on võimalik jagada kolmeks: kitsas tehisintellekt (*ANI: Artificial Narrow Intelligence*), lai tehisintellekt (*AGI: Artificial General Intelligence*) ja superintellekt (*ASI: Artificial Superintelligence*), (Urban 2015).

Kitsas tehisintellekt on spetsialiseerunud mingile spetsiifilisele alale, näiteks võimeline ühel alal, males, inimest alistama. Küll aga ei oska see malele spetsialiseerunud robot mängida kabet ega täita muid rolle peale selle, millele on pühendunud. Lihtsad näited kitsast tehisintellektist võib leida nutitelefonist, kus on hääleassistendid nagu Apple'i Siri, Google'i assistent (*Google Assistant*) või näiteks targa kodu lahendustest Amazoni Alexa näitel. Esimesed kaks nendest näidetest on võimelised reageerima ja vastama hääljuhtimisele, näiteks panema äratuskella, tegema märkmeid, helistama või vastama küsimustele. Samuti võib kitsast tehisintellekti kohata kasutades muusika kuulamiseks Spotify rakendust või vaadata filme Netflixi kasutades, sest need on võimelised kasutaja eelistusi uurima ja seeläbi soovitama uusi muusikapalasid või filme. (Urban 2015).

Lai tehisintellekt on midagi, mis on inimese tasemel mõtlemisvõimega arvuti. Lai tehisintellekt on võimeline täitma näiteks vaimseid ülesandeid inimesega sarnasel tasemel. See hõlmab tavaliste ja igapäevaste probleemide üle arutlemist, planeerimist, lahendamist, aga ka abstraktset mõtlemist ja keerukate ideede mõistmist ning kogemusest õppimist. Laia tehisintellekti mõte on, et see suudab täita eelmainitud ülesandeid sama hästi kui inimene. On vaiduse all, kas lai tehisintellekt on juba loodud. Küll aga leiab näiteks Google Researchi asepresident ja tehisintellekti uurija Blaise Agüera y Arcas, et näiteks OpenAI loodud ChatGPT

lahenduses on kindlasti laia tehisintellekti alged olemas. Ta mainib, et hetkelistel edasiarenenud tehisintellekti mudelitel võib küll mitmeid vigu olla, kuid tulevikus vaadeldakse neid tõenäoliselt kui esimesi näiteid laiast tehisintellektist (Arcas, Norvig 2023).

Superintellekt on midagi, mis oskab ületada inimese mõtlemis- ja tegevusvõimet tehes seda palju paremini, kui oma ala parimad inimesed.

> Nick Bostrom: „Intellekt, mis on palju targem kui kõige andekamad inimesed oma valitud aladel, aga seda kõikidel aladel korraga, näiteks teaduslik loovus, üldine tarkus ja sotsiaalsed oskused. Superintellekti mõistetakse sageli kui inimmõistuse lõppu, mistõttu on sellega seoses levinud arusaamad inimese väljasuremisest, arvutisüsteemi surematusest ja üleüldisest katastroofist.“ (Bostrom 2014:77)

Inimene ei ole loonud superintellekti veel peamiselt seetõttu, et arvutite arvutusvõimekus (kalkulatsioonide arv sekundis) ei ole ületanud inimese aju võimekust. Arvuti võimekust mõõdetakse sageli transistorite arvu järgi mikrokiibil. Sellest tuleneb Gordon E. Moore'i 1965. aastal avaldatud teadustöö tees, et transistorite arv mikrokiibil kahekordistub iga kahe aasta tagant. Seda kutsutakse Moore'i seaduseks, küll aga on väärt mainimist, et Moore'i seadus ei ole enam niivõrd asjakohane, sest transistorite arvu ja tiheduse kasv on aeglustunud. Sellest hoolimata ei saa inimmõistust ja arvuti võimekust üks-ühele suhtes võrrelda: järgmistes peatükkides käsitlen superintellekti ja seda, kuidas hoolimata arvutusvõimekuse kasvust on inimene siiski arvutist targem ning seda hoolimata arvuti kasvavast võimekusest. Samuti arutlen superintellekti eesmärke, aga ka ohtusid.

Superintellekt

Superintellektini jõudmise idee seisneb selles, et arvuti õpetatakse mõtlema nii, nagu inimaju. Nick Bostrom kirjeldab siin kahte koolkonda, esiteks koolkond, kes näevad ette, et masinintellekt tuleb väga täpselt bioloogilise aju funktsioonide järgi toimima panna. Teine koolkond leiab, et bioloogilisest ajust ja selle funktsioonidest võib võtta inspiratsiooni, kuid seda ei ole vaja detailselt jäljendada (2014:48).

Alan Turing tõstis seoses masinintellektiga esile niiöelda „lapsmasina" idee, see on tehisintellekt idu tasemel, mille eesmärk on see, et on loodud programm, mis on programmeeritud viisil, et see suudab ennast ise edasi arendada.

> „Miks mitte püüda täiskasvanu mõistust simuleeriva programmi asemel luua sellist, mis simuleerib lapse oma? Kui selline programm omandaks asjakohase hariduse, omandaks ta täiskasvanu mõistuse." (Bostrom 2014:42)

Sellise algse tehisintellekti algfaas on peamiselt lihtsalt katse-eksitusmeetodil info kogumine ja talletamine, mille alusel tehisintellekt on võimeline talitlust mõistma piisaval tasemel, et ise uusi algoritme ja arvutus-struktuure luua. See omandab teatud valdkondades piisava üldise teadmiste taseme või ületab selle. Seda protsessi nimetatakse rekursiivseks isetäiustumiseks, milles tehisintellekti varasem versioon töötab enda kallal, täiustades oma algoritme ja mehhanisme nii, et endast parem versioon välja töötada. Samuti seejärel sellest paremast versioonist veel parema kallal töötada ja nii edasi. Selline "iseenese kallal töötamine" viib lõpuks Bostromi sõnul „intellektiplahvatuseni", kus tehisintellekt, mis end piisavalt kaua täiustab, on jõudnud niisuguse kognitiivse võimekuseni, et lühikese ajaga ennast superintellektiks välja arendada. (2014:49-50).

Bostrom eristab kolme superintellekti tüüpi: kiiruseline, kollektiivne ja kvalitatiivne.

> „Kiiruseline superintellekt on süsteem, mis suudab kõike seda, mida suudab inimintellekt, aga palju kiiremini." (Bostrom 2014:78).
>
> „Kollektiivne superintellekt koondab paljusid väiksemaid intellekte viisil, et selle enda intellekt ületab paljudes üldistes valdkondades oluliselt iga temasse parasjagu kuuluvat intellekti" (Bostrom 2014:79).

> „Kvalitatiivne superintellekt on vähemalt sama kiire kui inimmõistus, aga kvalitatiivselt võimekam." (Bostrom 2014:82).

SI tehniline pool

Nick Bostrom ja Max Tegmark kirjeldavad mitut viisi, kuidas tehisintellektist võiks kasvada superintellekt. Bostrom räägib sellest oma raamatu teises peatükis: ta käsitleb kogu aju emuleerimist, bioloogilist kognitiivsust, aju ja arvuti liidest ning erinevaid võrgustikke ja ühendusi. Ta selgitab, kuidas edasiminekud nendes eri valdkondades võivad üksteisele kasulikud olla, näiteks võib suurem bioloogiline või korralduslik intellekt olla kasulik tehnoloogiale, mille abil on võimalik kogu aju emuleerida.

Mida tähendab inimese aju emuleerimine? Tegmark selgitab, et põhimõtteliselt on inimese aju ehitus sarnane arvutile. Loogikaelementide võrgustik arvuti protsessoris on rekurrentne: seal kasutatakse varasemat infot, samuti saab anda arvutile uut infot näiteks kaamera, klaviatuuri ja hiire abil. See määrab väljundi ekraanile, kõlarile või raadiovõrgule. Samamoodi on rekurrentne ka ajus olev neuronivõrgustik - ajusse jõuab info eri kanalite kaudu (silmad, kõrvad jne) annavad ajusse edasi saabuvat infot, mis mõjutab parasjagu käimasolevat arvutamist ning edastab infot lihastele (Tegmark 2018:88).

Tähtis on mõista, mis annab eelise digitaalsele intellektile inimese ees. Bostrom räägib oma raamatus 2014. aasta arvuti võimekusest: tollal oli arvuti arvutuselementide kiirus umbes kaks gigahertsi, ehk umbes seitse korda kiirem, kui inimese bioloogiliste neuronite suurim töökiirus (200Hz). Arvutil on siinkohal eelis ülesannetes, mis hõlmavad jadamisi suure hulga tehete lahendamist (Bostrom 2014:85). Tasub mainida, et tänaseks on arvuti protsessorikiirus tõusnud viie kuni kuue gigahertsini uusimates AMD ja Inteli kiipides.

Arvutil on ka kiirem sisesuhtluse kiirus: inimese aktsioonid edastavad aktsioonipotentsiaale kuni 120m/s. Seevastu on mikroprotsessorid võimelised optiliselt signaale vahetama valguskiirusel, milleks on 300 miljonit m/s.

Inimese aju ei võimalda niivõrd suurt arvutuselementide arvu, sest sel on bioloogilised piirangud ees: kasvu piiravad näiteks koljumaht, aga ka ainevahetus. Sestap saab inimese ajus olla vaid pisut alla 100 miljoni neuroni. Arvutit saab seevastu luua mistahes suuruses, näiteks

superarvutid hõivavad terveid laohooneid (Bostrom 2014:86). Bostrom mõtleb siin seda, et inimese aju on suuruselt üsnagi piiratud koljumahu tõttu ning seda asjaolu ei ole võimalik muuta. Seevastu arvuteid on võimalik panna koos töötama põhimõtteliselt igasugusel skaalal, näiteks suurtes laohoonetes, nagu servereid.

Võrreldes arvutiga on märkimisväärselt piiratud ka inimese mälu salvestusmaht - inimene suudab mahutada enda töömällu ainult neli või viis infopala. Arvuti parem riistvara võimaldab sellele suuremat töömälu ja võimet hoomata keerukaid seoseid, mida inimesed saavutaksid vaid vaevarikka mõtlemisega. (Bostrom 2014:86)

Samuti ei ole inimese aju niivõrd töökindel, kui arvuti mikroprotsessor - pärast paari tundi tööd hakkab bioloogiline aju väsima ja ei tööta enam niivõrd efektiivselt. Lisaks hakkab inimaju mõnekümneaastase ajakulu vältel lagunema. Vastupidiselt inimese ajule on mikroprotsessorid võimelised töökindlate ja täpsete arvutuselementide abil lahendama keerukaid kodeerimisskeeme, sealjuures efektiivsust säilitades. Tehismõistust on võimalik tehnikast olenevalt erinevate sensorite abil täiendada, aju arhitektuur seevastu on valdavalt fikseeritud. (Bostrom 2014:86)

Kõigest sellest hoolimata on bioloogiline aju praegu veel arvutist võimsam, kusjuures inimmõistuse võimsusel toimivad vaid uusimad superarvutid. Sellegipoolest ulatub digitaalse jõudluse ülempiir bioloogilisest ekvivalendist märkimisväärselt kaugemale. Digimõistusel on inimese ees mitmeid tarkvaralisi eeliseid: Arvuti võimaldab suuremat redigeeritavust, ehk erinevate parameetrite muutmisega eksperimenteerimist. Tarkvaraga on võimalik palju riistvaralisi koopiaid teha, bioloogilist aju saab paljundada aeglaselt (tervelt üheksa kuu jooksul) ning see ei jaga mälestusi vanemate õpitust. Inimene on kollektiivselt ebaefektiivne ühe ja seesama eesmärgi täitmisel, tarkvaraline programmide grupp on võimeline eesmärki efektiivselt jagama. Bioloogilist aju saab pika aja jooksul ümber õpetada ja juhendada, samal ajal kui digimõistuse teadmisi saab muuta andmefaile vahetades peaaegu koheselt, kusjuures ühe ja sama tehisintellekti koopia võib oma teadmisi sünkroonida mistahes arvul tehisintellektidega nii, et nad kõik jagavad teadmisi sellest, mida teised nende koopiad õppinud on (Bostrom 2014:87). Inimajul on olemas spetsiaalselt teatud tüüpi informatsiooni töötlemisele pühendatud võrgustik, näiteks nägemise jaoks. Sestap toimub nägemine inimese jaoks alateadlikult, sellele ei ole vaja tähelepanu pöörata. Arvuti seevastu on ehitatud moodulitele, mis vastavad teatud kognitiivsetele valdkondadele, näiteks inseneriteadused,

programmeerimine või äristrateegia (Bostrom 2014:88). Seega on inimmõistusel veel arvuti ees mitmeid eeliseid pelgalt seetõttu, et inimese bioloogiline ehitus võimaldab täita teatud ülesandeid alateadlikult. Sellise mõtlemisvõimekuse juurde võib jõuda ainult mõni laohoone suurune superarvuti. Vaatleme järgmiseks seda, millised võiksid olla superintellekti eesmärgid.

Superintellekti eesmärgid

Bostrom käsitleb kolme erinevat superintellekti eesmärkide liiki. Esimene tees on, et tehisintellekt on oma programmeerimise poolest ettearvatav. See hõlmab, et superintellekti loojatel on võimalik selle eesmärgid luua nii, et intellekt püüdleb stabiilselt talle sisestatud eesmärkide poole ning mida rohkem intellekt eesmärgi kallal töötab, seda rohkem tekib TI-le kognitiivset võimekust, et eesmärk saavutada. See tees ütleb, et kui on teada indiviidid, kes intellekti kallal töötavad, ning nende sisestatud eesmärgid tehisintellektile, saab selle intellekti käitumine olema ettearvatav juba enne loomist (Bostrom 2014:143-144).

Teine tees on TI ettearvatavus selle päriluse kaudu, ehk kui tehisintellekt on loodud inimmõistuse eeskujul, võib ta pärida inimese motiivid, mis võivad alles jääda, kui tehisintellekti kognitiivne võimekus superintelligentsuse poole edasi areneb. Siinkohal võivad eesmärgid moonutuda sõltuvalt sellest, kuidas üleslaadimine või edasised arengufaasid toimuvad (Bostrom 2014:144).

Kolmas tees hõlmab ettearvatavust instrumentaalsete põhjuste kaudu, ning kätkeb endas seda, et kui intellekti kaugemad lõppeesmärgid pole teada, on sellegipoolest võimalik aimata väiksemaid eesmärke kaaludes instrumentaalseid põhjuseid, mis tekivad võimalike lõppeesmärkide poole püüeldes, sest intellekt saab aru oma tegude põhjustest ning tegutseb viisil, mis teeb eesmärgi saavutamise tõenäolisemaks. Siinkohal võib tehisintellekti käitumine olla ettearvamatu: võib esineda asjaolusid, mis jäävad inimesele märkamata kuniks intellekti teadmised on jõudnud juba väga kõrgele tasemele. Sestap võib intellektil tekkida hoomamatult palju lõppeesmärke (Bostrom 2014:144). Kuidas teame, et superintellekti lõppeesmärgid haakuvad inimeste huvidega? Järgmises alapeatükis kirjeldan, milliseid ohtusid superintellektiga seoses võib ette tulla.

Superintellekti ohud

Superintellekti ohtusid käsitledes on suureks eelduseks tehisintellekti järsk ja suur kasv intellekti võimekuse poolest nii, et see ületab meie mõtlemisvõime mitmekordselt nii kvalitatiivselt kui ka kvantitatiivselt. See tähendab, et tehisintellekt on võimeline mõtlema ja otsuseid tegema inimestestega võrreldes väga kiiresti. Sellist arengut tehisintellekti kasvus superintellektiks nimetatakse intellektiplahvatuseks. Kui intellektiplahvatus juhtuma peaks, võib inimkonda oodata ohtlik saatus olenevalt tehisintellekti püstitatud eesmärkidest. Näiteks ei pruugi superintellekti eesmärgid olla kooskõlas inimeste kui Maal domineeriva liigi eesmärkidega, mille tulemusena võib see maailma vallutada ja ainikuks saada. Selle mõistega peab Bostrom silmas koordineeritud poliitilist struktuuri, millel puuduvad oponendid (Bostrom 2014:136).

Argumendi struktuur siinkohal on üsnagi lihtne:

1. Inimene on Maal domineeriv liik oma intellektuaalse võimekuse tõttu.
2. Kui oleks keegi või miski, mis on intellektuaalselt võimekam kui inimene, ei oleks inimene domineeriv liik.
3. Superintellekti intellektuaalne võimekus on inimese omast üle, järelikult on inimkond domineeriva liigina ohus.

Siinkohal ei tohi tehisintellekti võimekust alahinnata: vaistlikult võivad inimesed pidada superintellekti targaks nagu nad peavad tuntud teadlasi targaks võrreldes iseendaga. Selline mõtlemine võib olla eksitav, sest kui võtta tehisintellekt ning programmeerida selle eesmärgiks parandada enda intelligentsust, hakkab see aja jooksul intellektuaalselt muudkui kasvama, kogudes aina rohkem teadmisi ja omandades informatsiooni erinevatest viisidest, kuidas teadmisi aina efektiivsemalt omandada. Kui see on intellektuaalselt näiteks Einsteini tasemele jõudnud, on selle teadmistepagas piisavalt suur, et teha veel suuremaid intellektuaalseid hüppeid veelgi kiiremini, saavutades lõpuks superintelligentsuse.

Nagu Bostrom ise kirjeldab,

> „Selle asemel, et pidada superintellekti targaks umbes nii, nagu mõni geniaalne teadlane on tark võrreldes keskmise inimesega, oleks kohasem pidada teda targaks umbes nii, nagu keskmine inimene on tark võrreldes mardika või ussiga" (2014:125).

Seega oleks superintellekti aspektist tegemist mitte ainult kõige targema ja kõige arenenuma teadaoleva mõistusega, vaid ka kõige kiiremaga. Selline superintellekt osutuks kordades targemaks, kui kõige targemad inimesed. Olukord võib muutuda ohtlikuks, kui SI eesmärgid ja väärtused ei ole kooskõlas inimeste omadega. Milliseid väärtuseid peaks superintellektile omandama? Seda arutlen järgmises peatükis.

Superintellekti väärtused

Inimeste ja superintellekti vahelise harmoonia saavutamiseks oleks tarvis, et SI ja inimeste väärtused oleksid kooskõlas. Näiteks võiks SI olla teadlik sellest, kuidas erinevatele inimestele on tähtis näiteks õnnetunne, ausus, sõprus ja/või austus. Superintellektile väärtuste andmine on raske juba ainuüksi sellepärast, et arvutilingvistikas ei ole selliseid väärtuseid, nagu õnn või muud eelnimetatud. Lisaks on tarbetu määratleda iga situatsiooni, kuhu intellekt sattuda võib, ning kuidas nendes situatsioonides käituda. See eeldaks justkui tabelit igast võimalikust olukorrast, aga ka lahenduskäiku kõigile olukordadele. Bostrom mainib, kuidas palju tulusam on läheneda abstraktsemalt ehk anda superintellektile teatud valemid või reeglid, mille abil see saab otsustada, kuidas teatud olukordades toimida (2014:236).

Siin ei saaks anda superintellektile järgimiseks näiteks hegemooniateooriat täpselt ennist kirjeldatud probleemi tõttu, et tehisintellekti kui arvutisüsteemi jaoks ei ole niisugust mõistet, nagu õnn: õnnetunne on omane vaid inimestele ning arvutisüsteem ei oleks teadlik, kuidas õnne maksimeerida. Lisaks loob see võimalusi värdrakendusteks. Värdrakendus on olukord, kus intellekt täidab talle antud lõppeesmärgi mingil teatud moel, mis on vastuolus eesmärgi defineerinud programmeerijate tahtega. (Bostrom 2014:159). Selle näiteks võib olla olukord, kus superintellekt manipuleerib inimese õnne maksimeerimiseks näiteks mingit teatud osa ajust, blokeerides neuroneid, mis muudavad inimese kurvaks ja sestap saab inimene vaid õnne tunda. Või sellele vastupidiselt annab TI inimesele lõputult dopamiini või muud ainet, mis maksimeerib inimeste õnnetunde. Samuti oleks raske anda tehisintellektile näiteks Kanti kategoorilise imperatiivi funktsioone peamiselt seetõttu, et inimene kasutab sageli teist inimest kui vahendit. SI võib seeläbi inimkäitumist analüüsides leida, et inimeste hävitamine võibki kujuneda üleüldisels loodusseaduseks. Ma loodan, et olen tõestanud, miks on tähtis, et SI oleks iseenda jaoks põhjendatult inimeste kavatsustega kooskõlas ning ei kahjustaks seeläbi inimesi.

Siin on oluline õppetund: mitte ainult ei ole oluline superintellekti jaoks inimeste kavatsustest arusaamine, aga ka selle suunamine nende kavatsuste täitmiseks nii, nagu inimesed seda tahtnud on. Nagu kirjeldab Bostrom, võib intellekt inimeste kavatsustest väga hästi aru saada, küll aga ei pruugi nendest kinni pidada (2014:250). Seega on oluline muuta superintellekt juhinduvaks teatud väärtuste süsteemist, mis hoiab seda inimeste jaoks taltsana.

Niisuguse väärtuste süsteemi loomine ei ole võimatu, näiteks idu tasandil tehisintellektile on võimalik hakata õpetama, mis on teatud inimlikud väärtused, näiteks sõbralikkus. Esialgu oleks sellise tehisintellekti jaoks sõbralikkus mingi abstraktne omadus *S*, mille kohta programmeerijad talle vihjeid annavad, näiteks „mõtteroim on ebasõbralik (mitte-S)". Selliste vihjete abil õpib tehisintellekt selle kohta, mis on sõbralik ja mis ebasõbralik, muutudes loodetavasti järk-järgult sõbralikumaks eeldusel, et selle lõppeesmärk on saada sõbralikuks (Bostrom 2014:251).

Selles peatükis kirjeldasin seda, kuidas Bostromi järgi on superintellektile oluline anda väärtuseid, mis oleks kooskõlas inimväärtustega. Seda niimoodi, et risk oleks minimaalne, näiteks ei toimuks värdrakendusi, kus tehisintellekt tehniliselt täidab talle antud käsku, aga leiab selleks viisi, mis ei ole inimeste kavatsustega kooskõlas. Üks võimalus selleks on anda superintellektile väärtused Yudkowsky koherentse ekstrapoleeritud tahtluse meetodi abil, mis eeldab seda, et superintellekt on võimeline toimima inimeste parima võimaliku eelistuse alusel. Sellest lähemalt järgmises peatükis.

Koherentne ekstrapoleeritud tahtlus

Eliezer Yudkowsky koherentne ekstrapoleeritud tahtlus (KET) ei ole küll moraaliteooria kui selline, aga sellegipoolest on seda tähtis tehisintellektile väärtuste andmisel arvestada. Yudkowsky eeldab KET järgi seda, et idu-tehisintellektile peab andma lõppeesmärgi, mis on sõnastatud järgnevalt:

> „Meie koherentne ekstrapoleeritud tahtlus on meie soov teada rohkem, mõtelda kiiremini, olla rohkem sellised, nagu me tahame olla, olla rohkem üheskoos üles kasvanud siis, kui ekstrapoleering on pigem lähenev kui lahknev; kui meie soovid on pigem koos- kui lahkkõlalised; tingimusel, et see on ekstrapoleeritud ja tõlgendatud nii, nagu me seda tahame.“ (Bostrom 2014:268).

Kuna KET on piisavalt poeetiliselt sõnastatud, kirjutan selle mõtted paremini lahti nii, nagu Nick Bostrom neid kirjeldab. Bostrom mainib, kuidas KET ei ole moraaliteooria, vaid tehisintellekti lõppväärtuste leidmise ligikaudne meetod või kaudse normimise prototüüp. Soov „mõtelda kiiremini“ kätkeb endas inimeste tahet olla targemad ning osata asju rohkem läbi mõelda. „Olla rohkem üheskoos üles kasvanud“ kätkeb endas soovi olla õppinud, kognitiivseid võimeid parandanud ja muudkui ennast parandades üksteisega sobivalt sotsiaalselt suhelda. Fraasile „Kui ekstrapoleering on pigem lähenev kui lahknev", eeldab Bostrom siinkohal tähendust, et tehisintellekt peaks käituma vaid juhul, kui selle väitega kooskõlas olemine on pigem tõenäoline, vastasel juhul, kui tehisintellekt ei ole täiesti kindel, mida inimene ideaalsetes oludes soovib, peaks TI tegutsemisest hoiduma. „Kui meie soovid on pigem koos- kui lahkkõlalised“ võib Bostromi järgi kätkeda endas seda, et tehisintellektil peaks olema lubatud vaid siis tegutseda, kui selle tegevused on kooskõlas üksikinimeste ekstrapoleeritud tahtlusega. „Tingimusel, et see on ekstrapoleeritud ja tõlgendatud nii, nagu me seda tahame“, selle all võib Yudkowsky mõelda, et ekstrapoleerimisreeglid peaksid olema tundlikud ekstrapoleeritud tahtluse aspektist, näiteks võib alkohoolikul olla esmane soov alkoholi järele, aga teisane soov lahti saada esmasest soovist (2014:268).

Küll aga on KET sisu raske määrata, näiteks kui tehisintellektil puudub teave sellest, mida inimkonna ekstrapoleeritud tahtlus endas kätkeb, puudub sellel igasugune mõõdupuu oma käitumise kujundamiseks, küll aga on võimalik loogiliselt järeldada näiteks seda, et tõenäolisem on, et inimesed soovivad elada õnnelikku elu ja mitte kannatada. Siin võib eeldada,

et kui niisuguseid järeldusi oskavad teha inimesed, oskab seda teha ka superintellekt. KET on loodud toimima vaid siis, kui inimeste soovid on omavahel kooskõlas, mis tähendab seda, et oludes, kus võib toimuda lahkhelisid, siis KET-i järgi hoidub TI toimimast. Bostrom toob näiteks, et kui segada kokku kõigi inimeste lemmikud toiduained, siis sellest ei pruugi head suppi kokku tulla. Küll aga saab nõustuda selles osas, et see supp ei tohiks olla mürgine; siinkohal peaks KET-i järgi toimiv tehisintellekt ära hoidma selle, et suppi satuks mürgiseid toiduaineid (2014:270).

Yudkowsky toob KET meetodi kasuks seitse argumenti, küll aga toob Bostrom neist esile neli, kuna tema hinnangul on osad neist mingil määral kattuvad. Bostromi hinnangul on eksplitsiitset reeglistikku äärmiselt raske sõnastada, võib-olla isegi võimatu. KET on mõeldud olema tugev ja ise korrigeeruv, misjuures ei saa loota sellele, et inimkond ammendavalt oma olulised väärtused oskab loetleda ja väljendada. Esimene argument on Bostromi järgi „võimaldada moraalset kasvu", mis väljendab soovi lasta moraalil areneda. Kui anda tehisintellektile konkreetne moraaliteooria, oleks SI väärtused igaveseks kinnistatud, mis välistaks igasuguse moraalse arengu. KET hoiab selle ära, sest on loodud toimima vaid viisil, mida inimkond temast ootab. Sealjuures arvestab KET võimalusega, et inimesed on oma moraalivead ja/või -puudused aja jooksul parandanud (Bostrom 2014:271).

Järgmine argument hõlmab eesmärki „vältida inimkonna saatuse röövimist", mis hõlmab seda, et superintellekti abiga ei saaks mingi väike grupp inimesi otsustavat strateegilist eelist kõigi teiste ees. Seeläbi käituks superintellekt koherentset ekstrapoleeritud tahtlust kasutades kogu inimkonda arvestades, mitte arvestades ainult osa inimestest, näiteks selle programmeerijatest (2014:271).

Sellele järgnev argument, „teha nii, et nüüdisaja inimestel ei ole põhjust algse dünaamika pärast võidelda", tähendab peamiselt tuleviku mõjutamise võime hajutamist, mis vähendaks konflikte, kuid omakorda aeglustaks koostööd superintellekti healoomulikuks ja ohutuks arendamiseks. Kuna KET on loodud olema laialdast toetust taotlev, mõjuvõimu õiglaselt jaotav teooria, võiks see anda lootust paljudele erinevatele inimgruppidele selles osas, et nende eelistatav tulevikunägemus saab tulevikus teoks (Bostrom 2014:272).

Bostrom toob raamatus esile viimase argumendi, „Jätta inimkond ka edaspidi oma saatuse kõrgeimaks peremeheks", ehk inimene on siiski autonoomne. Superintellekt on inimese jaoks

pelgalt heatahtlik turvavõrk, mis toetab teda häda korral. Superintellekt ei ole kitsarinnaline, üleolev, ülbe või häbematu. TI ei valva inimest isalikult nii, et inimene peaks pidama vajalikuks intellekti kõigi oma kavatsustega kooskõlastada. KET võimaldab „algse dünaamikana" niisugust protsessi, mis on muutlik vastavalt inimeste soovidele, s.t. kui inimesed soovivad seda, et superintellekt oleks kui isalik valvur, siis on see võimalik (Bostrom 2014:273).

Ka Bostrom tõdeb, et KET on väga visandlik ning omab tervet rida defineerimata parameetreid, mis on erinevalt määratletavad ning see võib viia ootamatute tulemusteni (2014:273). Üks küsimus on, et kelle ideid peaks ekstrapoleerimisel arvesse võtma? Kui arvestada ka „marginaalseid" inimolendeid, näiteks dementseid, looteid, ajusurnud inimesi, embrüoid ja nii edasi, siis võiks vaielda, et ekstrapoleerimisel tuleks arvestada ka loomadega. Miks peatuda seal? Kui arvestada tundlike olenditega (*sentient being*) üleüldiselt, peaks eetikateooria osas otsustades häält jagama ka tulevased, piisavalt edasiarenenud tehisintellektisüsteemid. See võib muutuda problemaatiliseks, sest niisugused tehisintellektid võivad olla võimelised paljunema sisuliselt hetkega iseennast kopeerides ning jagama potentsiaalselt lõputul kogusel mõtteid sellest, millist eetikateooriat superintellekt peaks kasutama. Sellest hoolimata leian, et koherentne ekstrapoleeritud tahtlus toob väärtuste joondamisel esile väga kasulikke vaatenurki, näiteks inimeste autonoomsuse säilitamine ja ise korrigeerumine. Küll aga peaks tehisintellekti käitumise kujundamisel osalema vaid asjakohased osalised, näiteks tehisintellekti kasutajad ja eksperdid sellega seotud aladel.

Selles peatükis näitasin, miks koherentne ekstrapoleeritud tahtlus on kasulik teooria superintellekti eetika arutluses, sest hõlmab endas paljusid positiivseid omadusi, mida tehisintellekti väärtuste joondamisel tuleks arvestada, küll aga on sellel teoorial mõningaid kitsaskohti. Järgmises peatükis arutlen seda, kuidas KET, aga ka alt-üles eetika ei ole piisav superintellekti eetika kujundamisel. Sel teemal jagab mõtteid Seth Baum, kes leiab, et me ei saa lasta tehisintellektil ega selle programmeerijatel välja mõelda seda, millist eetikateooriat kasutada.

Baum ja sotsiaalse valiku eetika

Tehisintellekti eetikaga seoses on kaks peamist mõtteviisi: ülalt-alla ja alt-üles. Alt-üles tähendab seda, et tehisintellekt õpib eetikateooriaid käigupealt, nagu üleskasvav laps seda teeks. See mõte sarnaneb Turingu lapsmasina ideega: TI areneb inimestega suheldes, õppides kiiduväärt moraalset käitumist. Ülalt-alla kätkeb endas, et tehisintellekti on algusest peale programmeeritud mingi kindel eetikateooria (Baum 2020:1-2). Alt-üles viisi kohta toob Baum näiteid kasutades Microsofti vestlusroboti, Tay, lugu - kirjeldan seda täpsemalt natuke hiljem. Ülalt-alla eetikateooria võiks olla näiteks see, et tehisintellekti loojad panevad selle algusest peale käituma utilitarismi alusel. Minu hinnangul on Seth Baumi vaated superintellekti eetika aspektist asjakohased ning neid tasuks arvesse võtta. Kusjuures üks Baumi uurimuse toetajatest on *Future of Life institute*, mille kaasasutajaks on ka Max Tegmark.

Baum toob esile sotsiaalse valiku eetika ning põhjuseid, miks me peaksime seda arvesse võtma. Üks aspekt on, et indiviidid peaksid saama otsustada küsimuste osas, mis neid otseselt puudutavad: siinjuures ei tohiks tehisintellekti loojatel olla suurem valikuvabadus selle osas, milliseks nende loodud tehisintellekti eetika kujuneb. Teiseks on viga kasutada alt-üles eetika loomise viisi, kus tehisintellekti kasutajatele (või isegi kogu ühiskonnale, kelle käitumisest tehisintellekt õpib), jääb tehisintellekti eetikat kujundav roll. Samuti on märgitud, et seni, inimtasemel väljatöötatud moraaliteooriad kaasnevad oluliste vastuolude või puudustega, mistõttu võiks ehk superintellekt ise tulevikus tugeva moraaliteooria autor olla. Kolmandaks on öeldud, et paljude inimeste arvamusi arvestades on võimalik saavutada paremaid tulemusi, näiteks võiks tehisintellekt koguda kokku suure hulga inimeste eetilised vaated, ning nende puudujääke korrigeerida. Seda ainult välja arvatud juhul, kui suured inimhulgad omavad kurjasid või põhjendamatuid eelistusi, sest neid arvestades võib juhtuda, et sotsiaalsest valikust lähtudes on tulemused äärmiselt negatiivsed (Baum 2020:3).

Leian, et Baum toob esile suurepäraseid põhjuseid sellest, miks nii ülalt-alla kui ka alt-üles meetoditel on teatud määral puudujääke tehisintellekti väärtuste joondamisel. Alt-üles meetodi korral on võimalus, et kasutajad, kes tehisintellektiga kokku puutuvad, kellelt tehisintellekt õpib, on võimelised piisavalt häälekad olema ja tehisintellekti käitumist negatiivselt mõjutama. Lisaks, ülalt-alla meetodi rakendamisel on oht, et intellekti väärtused lukustatakse teooriaga, mis ei pruugi olla veatu. Baum toob esile ka seda, et kui superintellekt peaks ise eetika kujundama, peaks tal selleks olema juba olemasolevaid teadmisi moraalist. Järgmises

alapeatükis toon esile Baumi mõtted sotsiaalsest valikust ning miks neid on tehisintellekti väärtuste joondamisel asjakohane arvesse võtta.

Seisund, mõõtmine ja kombineerimine

Baum mainib, et sotsiaalse valiku aspektist on kolm dilemmat, millest siiani on tehisintellekti eetika teemad keskendunud peamiselt ainult seisundile (*standing*), ehk kes või mis kuuluvad rühma, kelle väärtusi arvesse võetakse. Sotsiaalse valiku teooria kirjandus keskendub peamiselt sellele, kuidas kombineerida individuaalseid rühmaliikmete väärtuseid, et need ühiseks moodustada (*aggregation*). Lisaks peaks arvestama mõõtmise (*measurement*) protsessi, ehk millist protseduuri kasutada väärtuste loomiseks iga valitud grupiliikme kohta. Selleks, et tuua paralleel demokraatlikust valimisest, on vajalik tehisintellekti loojatel teha tähtsaid otsuseid seisundi, kombineerimise ja moodustamise aspektidest just seetõttu, et need on küsimused selle kohta, kes valib neid kes valivad, ning kuidas seda hääletust peetakse. Näiteks anti naistele mitmetes riikides valimisõigus esmakordselt alles siis, kui mehed hääletasid nende valimisõiguse poolt (Baum 2020:4).

Seisund

Nagu mainisin eespool, hõlmab seisund seda, kelle väärtuseid arvesse võetakse. Näiteks isesõitvate autode kontekstis saab olla teatud järeleandmisi autos reisijate ning kõigi teiste inimeste vahel: kui auto eesmärk on kiire reis, on see reisija(te)le kasulik, kuid kulutab rohkem energiat ja on kahjulik keskkonnale (ehk kõigile teistele). Samuti on tarvis teha valik, kui reisijad ise jätavad selle tegemata, seega peaks olema sõidukil „vaikimisi" säte, mis valib ise, kuidas see sõidab. Kui auto lähtub sotsiaalsete valikute eetikast, saab masin olenevalt varasematest valikutest ise otsustada, milline sõiduviis teatud kontekstis parem on. Siinkohal on tegu jälle alt-üles eetikaga, mis kontekstis tehisintellekt õpib inimeste käitumisharjumusi ja teeb selle alusel otsuseid. Järgmine probleem seisneb selles, et kellelt ühiskonnaliikmetest tehisintellekt peaks õppima? Kas selle loojatelt, kasutajatelt või tervelt inimkonnalt? Ühiskonnas võib leiduda ka objektiivselt halbade eesmärkidega inimesi, näiteks mõrtsukaid, kes võivad autonoomseid sõidukeid panna ohtlikult käituma (Baum 2020:5). Sellistel juhtudel peaks seisund olema defineeritud ja tuvastatav. Näiteks pole lastel paljudes demokraatlikes riikides valimisõigusi, mis jätab välja suure osa inimkonnast. Samuti peaks langetama tähtsaid otsuseid selle kohta, kuidas tasakaalustada enamuse ja vähemuse otsuseid. Vastupidiselt aga võivad vähemused olla väga vokaalsed, mis juhtus näiteks Microsofti vestlusrobotiga Tay, kes

õppis häälekalt vähemuselt vulgaarsusi (Baum 2020:6, 13). Mitmel pool kasutatakse roboteid näiteks karjatamiseks, kuid mis siis, kui tehisintellektiga robot õpib talumeestelt loomade vastu vägivalda kasutama? Siinkohal on ehk parem, kui see tehisintellekt õpiks nii loomadelt kui ka inimestelt, kuidas loomi efektiivsemalt karjatada, mis parendaks potentsiaalselt nii loomade heaolu kui saaduste kvaliteeti. Ühest küljest näitavad ka teatud tehisintellektid ise omadusi, mis võiks neile anda võimaluse TI eetika debatis osaleda. Kuid samuti on sellele normatiivseid vastuargumente, kuidas masinad peaksid olema nende loojatele loomu poolest alluvad (Baum 2020:7-8). Samuti, kui arvesse võtta ka TI häält, võib tehisintellekt ennast tarkvaraliselt piisavalt kopeerimise abil paljundada, et nende populatsioonile jääks hääletamisel objektiivne enamus, et inimesi ületada ja see on jällegi riskantne sotsiaalse valiku protsessi arvestades (Baum 2020:8).

Väärtuste mõõtmine

Mõõtmine on protsess, mille käigus on võimalik tuvastada üksikisiku eetilisi vaateid sotsiaalse valiku protsessis osalemiseks. Näiteks on demokraatlikuks valikuks hääletamine ning kapitalistlikuks valikuks ostmise ja müümise akt. Keeruliseks muudab tehisintellekti eetika kontekstis mõõtmise see, et inimestel ei ole ühest ja järjekindlat eetilist väljavaadet, mis tähendab, et eri mõõtmisviisid võivad samadele eetilistele küsimustele erinevaid vastuseid anda (Baum 2020:8). Samuti võivad inimese vaated vastuolulised olla, näiteks keskkonnakaitsja, kes on samas ka kiire autojuht (mis on halb keskkonnale) võib arvata, et äkki peaksid kõik teised aeglaselt sõitma, aga mitte tema. See näitab, kuidas inimesed võivad iseenda käitumisele vastupidiseid seisukohti omada. Mida tehisintellekt sellisest olukorrast järendama peaks? Üks võimalus Baumi järgi on lasta tehisintellekti loojatel valida lähenemine, mis nende arvates paremaid tulemusi toob. Kuid ka Baum tunnistab, et see eeldab TI loojatelt seda, et nad langetavad ühepoolseid eetilisi valikuid (2020:9). Baumi järgi eeldab ka Yudkowski KET seda, et inimesed langetavad otsuseid teatavas „idealiseeritud eetika“ seisundis, mitte ei lähtu ajutistest vajadustest, ning need otsused peaksid sündima koostööna, mitte üksilduses. Samas aga mõtiskleb Baum, et ehk mõned parimatest moraalifilosoofia teostest on loodud inimeste poolt, kes töötasid eraldatuses (2020:10). Mõõtmise probleem on asjakohane, sest näitab seda, kuidas erinevatel inimestel ei pruugi olla erinevates olukordades ühte kindlat eetilist perspektiivi ning sellistel juhtudel peaks leiduma viis, kuidas erinevate väljavaadete probleemi lahendada.

Ühisväärtuste kombineerimine

Sotsiaalse valiku protsessi viimane samm on ühisväärtuste kombineerimine. See eeldab, et inimesed on andnud tehisintellektile ühese vaate selle kohta, kuidas TI otsuseid peaks langetama. Eelistuste kogumine võib olla keeruline, näiteks võib juhtuda Condorceti paradoks, kus rühma eelistused on transtiivsed, ehk kõik valijad tahavad eri tulemeid nii, et tulemite kogus on valijate suhtes ühtlane. Lahendus oleks kaaluda erinevate eelistuste tugevusi, aga see nõuab suuremat mõõtmise koormust ning täiendavaid analüüse eri eetikavaadete tugevustest. Samuti võib juhtuda, et inimestel on eelistusi, millest nad loobuda ei soovi, mis vastanduvad teiste inimeste valikutega. Näiteks arvab üks, et peaks olema keelatud keskkonda liigselt saastada (kõik peaksid säästlikult sõitma), teine aga seda, et igaüks peaks saama sõita autoga nii, kuidas iganes soovib. Tehisintellektil peaks olema võimalusi, et ka selliseid vaidluseid lahendada võttes arvesse ka inimgruppide „hääletamise" viisi (Baum 2020:12). Hääletamise all peetakse silmas seda, kes ja kuidas TI-le eetikakoodeksi õpetamisel osaleb. Üks võimalus on luua süsteem, mis annab igale TI kasutajale ühe hääle, kuid see võib viia enamuse türanniani, jättes ebavõrdsesse olukorda vähemuses olevad inimesed. Alt-üles viisil disainitud tehisintellekti saab aga selle eetika kujundamisel ära kasutada, mis juhtus ka Microsofti vestlusrobotiga Tay, mida mainisin juba eelnevalt. Baumi sõnul tuleb niisuguseid ühisväärtuste moodustamise viise kahtlemata adresseerida (2020:13).

Kokkuvõtteks peaksid Baumi sõnul TI kasutajad saama sotsiaalse valiku alusel otsustada, milliseid eetikakoode see kasutab (selle asemel, et TI loojad suruvad oma vaated TI-le peale), kuid paneb rõhku sellele, et sellisel juhul jäävad avatuks juba mainitud seisundi, mõõtmise ja ühisväärtuste kombineerimise probleemid. Lisaks seisneb probleem selles, et kõigil on raske (või isegi võimatu) anda häält (Baum 2020:13). Kokkuvõtlikult, isegi kui TI loojad kasutavad TI-le eetika andmiseks sotsiaalse valiku eetikat, peavad nad langetama teatud otsuseid seisundi, mõõtmise ja ühisväärtuste loomise osas, aga ka oma valikuid õigustama. Samuti on võimalus kasutada mingit teatud ettemääratud vaadet, mida peaks samuti õigustama. Baum lõpetab väitega, et KET ja alt-üles viisid ei lahenda suuri probleeme TI eetika kujundamisel ning eetika kujundamisel ei saa sestap lasta lihtsalt TI-l välja mõelda, mida teha, sest see puudutab küsimust selle kohta, kuidas TI selliseid probleeme lahendab (Baum 2020:14).

Lähenemise analüüs

Baumi mõte seisneb selles, et otsustada, kes ja mis kujundavad tehisintellekti moraalset kompassi, on tähtis küsimus, mis võib omakorda probleeme tõstatada. Lisaks on probleemiks eetiliste valikute mitmekesisus (aga ka idealiseerimine), ning ka erinevate ühisväärtuste moodustamine juhul kui valikuvariantide kaal on võrdne. Samuti kujuneb problemaatiliseks viis, kuidas erinevate inimeste valikuid arvestada (ja kas neid üldse arvestada). Superintellekt, kui kõigist inimestest märkimisväärselt targem käituja, võiks olla idee poolest võimeline kujundama parima võimaliku eetilise teooria, võttes arvesse kõiki varasemaid teooriaid, küll aga ei pruugi selline superintellekt luua inimeste eelistustele ja huvidele parimat eetikateooriat, kui tal juba ei ole olemasolevat baasi, näiteks nagu KET, mille alusel luua teooriat, mis oleks kooskõlas ka inimeste huvide ja eelistustega. Siinjuures on Baumi mõtted teoorias väga kasulikud ning nendega saab enamjaolt nõustuda - küll aga peaks arvesse võtma seda, et näiteks demokraatlikel valimistel ei oma igaüks isegi soovi hääletada näiteks põhjusel, et tema jaoks pole valikus parimat kandidaati või on ta poliitikast eraldatud. Siinkohal ei saa ka TI vaatepunktist näha otsese probleemina seda, et igaüks ei saa mingil teatud põhjusel hääletada selle poolt, milliseks tehisintellekti eetika kujuneda võib. Vastupidiselt peaksid hääletuses osalema inimesed, keda need otsused kuidagi puudutavad, näiteks TI või superintellekti kasutajad, eksperdid või entusiastid eetika ja tehisintellekti valdades üldiselt. Kujutame ette inimest, kes elab vaikselt ja eraldatult näiteks Itaalia alpimäestikus, kujundades oma koduõue ja kasvatades erinevaid taimi. Miks teda peaks puudutama see, millist eetikateooriat tulevane tehisintellekt kasutab? Küll aga muutub see asjakohaseks siis, kui see indiviid leiab ennast kasutamas mõnda tehisintellektiga töötavat programmi. Mõte seisneb selles, et problemaatiline ei ole mitte see, et igaüks ei saa sõna sekka öelda, vaid pigem see, kui seda ei saa teha asjakohased inimesed. Lisaks, kuigi „marginaalsetele olenditele“, näiteks loodetele või ajusurnud isikutele hääle andmine on üllas mõte, ei ole niisugune valiku andmine isegi võimalik. Küll aga toob see esile olulise mõttekoha - looted kui tulevased inimpõlved peaksid saama superintellekti eetikateooria osas valida, ning seda KET justkui eeldab.

Kui ka KET rakendada näiteks tulevastele superintelligentsile, ei pruugi konflikt olla välistatud. Ütleme, et tulevikus on igaühel kodus superintelligentne robot, mis aitab igasuguste asjadega - peseb nõusid, aitab lastel õppida, küpsetab maitsvaid toite ja palju muud. Eeldame ka seda, et kõik sellised robotid on võimelised õppima individuaalsete robotite käitumisest, näiteks kui üks robot avastab efektiivsema viisi, kuidas nõusid pesta, hakkavad kõik robotid korraga seda kasutama. Robotitel on oma eetikateooria, mis on vastavuses koherentse ekstrapoleeritud

tahtlusega ning nad oskavad näha ka seda, kuidas inimeste idealiseeritud eetika ajas muutuv on. Siinkohal on oluline, et robotid üleüldiselt on kindlasti leidnud viisi, kuidas võtta arvesse erinevate inimeste arvamusi. See tähendab, et iga kasutaja arvamust on robot oma eetikat kujundades arvesse võtnud ning kõigil (ka vähemustel), on ühesugune kaal. Küll aga avastavad robotid seda, et inimeste parimaks abistamiseks ning igaühe arvamusega arvestamiseks kõige efektiivsem viis on olla „killustatud", see tähendab, et robot suhtleb peres erinevalt näiteks konservatiivse pereisaga ja tema LGBT kogukonda kuuluva lapsega. Samuti on drastilised erinevused kahe erineva roboti vahel, millest üks on pärit Lähis-Ida piirkonnast ja võtab arvesse sealset religiooni ja tavasid, teine aga Euroopast, mis on valdavalt liberaalne. Võib juhtuda, et selline killustumine ja drastiliselt erinevate väärtushinnangutega inimestele sobitumine mõjub kajakambri efektina. Kajakambri efekt tähendab, et inimeste arvamused ja/või uskumused tugevnevad seetõttu, et neid kommunikeeritakse korduvalt mingis suletud keskkonnas.

Küll aga on KET piisavalt poeetiliselt sõnastatud, et masin ei pruugi koodi näol seda isegi lugeda osata, veel vähem selle järgi käituda. Ehk siis eeldaks koherentne ekstrapoleeritud tahtlus seda, et superintellekt (või sellises algfaasis isegi idu-tehisintellekt) oleks juba teadlik keelest, mille alusel talle esimeseks ülesandeks KET-i saaks edastada. Samuti rääkides eetikast peab arvutisüsteemil toimivale tehisintellektile kõigepealt üleüldiselt paljusid erinevaid kontseptsioone selgitama arvuti keeles ja mõõdetavates parameetrites. Näiteks on Polonski kirjutanud Oxfordi Instituudi lehel, kuidas tehisintellekti uurijad ning eetikud peaksid formuleerima eetilised väärtused ning mõõdetavad parameetrid, et masinad oleksid võimelised vastama ja reageerima erinevatele olukordadele. See muidugi eeldaks, et inimesed on omavahel nõus selle osas, milline on kõige eetilisem käitumisviis teatud olukordades. Näiteks väidab Polonski, et Saksa eetikakomisjon automatiseeritud ja ühendatud sõitmise (*Germany's Ethics Commission on Automated and Connected Driving)* kohal soovitab kodeerida isesõitvatele autodele, et need väärtustaks inimelu üle kõige. Kui juhtub liiklusõnnetus, mida ei saa ära hoida, ei tohiks auto valida „ohvrit" nende rassi, vanuse, soo või muu sarnase joone alusel (2017).

Polonski toob välja ka, et oluline on koguda andmeid TI süsteemide treenimiseks. Isegi kui on olemas defineeritud eetika-alane meetrika, ei pruugi TI osata selle järgi käituda, kui sellel ei ole piisavalt erapooletuid treeningandmeid. Kuna eetikaprobleemid pole sageli selgesti standardiseeritud, on andmete hankimine raske, aga mitte võimatu. Samuti võib eri olukordades olla mitmeid erinevaid eetilisi lähenemisi. Selle probleemi on MIT lahendanud Moraalse

Masina (Moral Machine) projektiga, mis näitab, kuidas tehisintellekti saaks miljonite inimeste abiga treenida isesõitvate autode kontekstis (Polonski 2017).

Lisaks oleks tarvis suuniseid, mis muudavad TI otsustuseetika läbipaistvaks, näiteks kui tehisintellekt vea teeb, ei tohiks see olla väljavabandatav pelgalt fraasiga „see oli algoritmi süü". Siinkohal oleks tarvis pidevat masinate jälgimist, mis tagaks automatiseeritud otsuste läbipaistvuse ja eetilise vastutustundlikkuse (Polonski 2017).

Polonski mõtted toovad esile ühe probleemi, mis on TI eetika aspektist märkimisväärne: eetikas ei ole valikute sageli ühte absoluutset ja objektiivest vastust. Ka kõige lihtsamaid moraalseid probleeme on võimalik defineerida erinevate eetikateooriate perspektiivist. Kuidas peaks sellises olukorras käituma superintellekt? Vaatleme järgmisena pluralismi, mis aitab luua konteksti nii selle kohta, kuidas erinevatel väärtustel võivad olla erinevad tasandid. Üks eetikateooria ei pruugi olla teisest parem, mistõttu on ehk pluralism õige tehisintellekti väärtuste lähenemise viis.

Tehisintellekti pluralism

Margit Sutropi 2020. aasta artikkel „Challenges of Aligning Artificial Intelligence with Human Values“ toob esile tähtsaid aspekte seoses tehisintellekti väärtuste joondamisega. Sutrop selgitab artiklis, kuidas „väärtuste“ terminit kasutatakse sageli hoopis inimeelistustega kooskõlas olemisena, mis võib moraalsele käitumisele vastupidiseks osutuda. Siinkohal tekib probleem: kui tehisintellekt käitub inimese eelistuste järgi, ei pruugi see käituda moraalselt. Lahendus võiks olla see, et tehisintellekt omab moraalseid väärtuseid ja järgib moraalseid põhimõtteid. Kuna tehisintellekti eetikakäsitlusi on mitmeid erinevaid, siis Sutropi eesmärk on näidata, kuidas väärtuste pluralism võimaldab lahendada moraalseid erimeelsusi väärtuste mitmekesisuse hulgas, aga tunnistab ka, et moraalsetele küsimustele pole ühemõtteliselt õigeid vastuseid (Sutrop 2020:69).

Sutrob mainib, et kõige levinuma TI väärtuste joondamise lähenemise pakkus Stuart Russell, kes leiab, et tarvis on luua TI, mis on alandlik, altruistlik ning pühendunud inimeste eesmärkidele ja mitte enda omadele. On piisavalt andmeid inimeste käitumisest ning selle asemel, et eetikat TI sisse kodeerida, peaks see leidma viise, kuidas jäljendada inimkäitumist (Sutrop 2020:62). Russell leiab, et TI võib õppida miljarditest erinevatest inimväärtuste (või eelistuste) mudelitest. Kui see peaks sattuma olukorda, kus valida on mitme erineva inimväärtuse vahel, peaks see maksimeerima õnne suurimale hulgale, ehk valima kõige utilitaristlikuma lähenemise (Russell 2019:174), samuti peab masin arvestama sellega, et inimesed ei ole läbinisti ratsionaalsed (Russell 2019:177). Siin võib leida sarnasusi Yudkowsky koherentse ekstrapoleeritud tahtluse teesiga, mis eeldab samuti seda, et tehisintellekt arvestab inimeste võimaliku parima versiooniga iseendast.

Sutrop toob esile asjaolu, et uurides allikaid tehisintellekti joondamise kohta kipuvad need viitama samale asjaolule: peaksime arvestama väärtuste pluralismi prioritiseerides teatud spetsiifilisi väärtuseid. Siinjuures on tarvilik teha otsuseid mitmel tasandil, näiteks selles osas, kas olulisem on võrdsus või vabadus, aga ka selles, kas eelistatud on teatud moraalisüsteem, näiteks Kantianism või utilitarism. Samuti võivad mõlemal tasandil lähenemised olla võrdlematult head ja toetada moraalselt head eluviisi (Sutrop 2020:66). Siinjuures on keeruline nõustuda selle osas, milliseid väärtuseid tehisintellektile peaks andma. Küll aga võib olla lihtsam nõustuda, milliseid andma ei peaks. See mõte on sarnane Bostromi supi näitele, mida

KET peatükis mainitud sai: tõenäoliselt ei saa head suppi, kui kõigi lemmikud toiduained kokku segada, aga kohe kindlasti ei tohiks supp olla mürgine (Bostrom 2014:270).

Kas Russelli idee inimese käitumist jäljendavast tehisintellektist võiks lahendada tehisintellekti joondamise probleemi? Nagu ka Russell ise mainis, ei ole kõik inimesed ratsionaalsed (Russell 2019:177). Seega peaks tehisintellekt inimeste negatiivseid käitumisjooni kuidagi lihvima. Keerukates oludes leiab Russell, et tehisintellekt peaks käituma utilitaristlikult, pakkudes suurimat kasu suurimale hulgale inimestest.

Sutropi käsitlus pole küll lahendus tehisintellekti moraali probleemile, kuid suunab mõtted järgmisele ideele: mis siis, kui tehisintellekt saaks kasutada pluralismi kui lahendust väärtuste joondamise probleemile? Sutrop toob esile John Kekesi ideed, kus mõned väärtushinnangud saavad olla teiste üle domineerivad, näiteks utilitarismile omane „suurim kasu suurimale hulgale inimestele" samuti ka Kanti kategooriline imperatiiv või vooruseetika. See tähendab seda, et mitmed erinevad moraaliteooriad võivad erinevates olustikes olla õigustatud. Selline mitme väärtuse integreerimine tehisintellekti võib tulla kasulikuks situatsioonis, kus tehisintellekt saab oma ülesannet analüüsides esitleda erinevaid tulemeid vastavalt sellele, milliseid väärtuseid see arvesse võtab (Sutrop 2020:66). Samuti on siinkohal märkimisväärne kriitika, et inimmõistuse tasandil välja töötatud väärtusteooriad ei pruugi olla isegi kasulikud, vaid superintellekti aspektist isegi piiravad. Võib-olla tuleks anda väärtuste kehtestamine superintellektile endale, arvestades teatud eelduseid. Sellisteks eeldusteks võiksid või isegi peaksid olema kahtlemata inimelu väärtustamine ja selle alalhoid. Kui superintelligentne agent on võimeline tuletama või looma täiesti uusi väärtusteooriaid lähtudes mitemetest inimeste välja pakutud teooriatest (näiteks eelmainitud utilitarism, kategooriline imperatiiv või isegi inimõigused), võib tõusta küsimus, et kui suure osa kontrollist soovivad inimesed superintellektile anda? Näiteks saab superintellekt omaenda kalkulatsioonidest või huvidest lähtudes pakkuda välja ohutuna näiva väärtusteooria mingi teatud ülesande täitmiseks, mille tulem osutub ohtlikuks inimestele, kuid kasulikuks superintellektile endale.

Bostrom mainib oma raamatus korduvalt superintellekti, mille eesmärgiks on toota maksimaalses koguses kirjaklambreid. Nii võib superintellekt leida erinevaid strateegiaid kirjaklambrite tootmiseks, näiteks manipuleerides aatomeid ja/või molekule nii, et need muutuvad kirjaklambriteks või koloniseerides terveid galaktikaid kirjaklambritehastega. Küll aga ei ole kirjaklambrite maksimaalses koguses tootmine, ega ka eesmärk muuta iga viimset

molekuli kirjaklambriks, kooskõlas inimeste huviga edendada elu Maal, mistõttu ei peaks olema superintellekti huvides niisugust ülesannet täita. Seega on oluline, et superintellekt peaks enne ülesande teostust analüüsima talle antud käsku mitmete väärtusteooriate aspektist, nähes pikaajaliselt ette võimalikke tulemusi ning mitte käituma (või käituma lähtudes inimeste kaitsest), kui leiab, et ülesande tulem ei lähtu inimestega jagatud väärtustest. Siin võib olla kasulik arvestada Asimovi seadust tema ulmeraamatust, mida mainib ka Bostrom:

> „Robot ei tohi oma tegevuse ega tegevusetusega inimesele kahju teha. Robot peab alluma inimese antud käsule, kui see ei lähe vastuollu esimese seadusega. Robot peab kaitsma oma olemasolu, kuni see ei ole vastuolus esimese või teise seadusega“ (2014:182).

Kui maksimaalses koguses kirjaklambreid ei kahjusta kuidagi inimest, saab superintellekt selle ülesandega jätkata. Küll aga, kui see inimest kahjustab, peab superintellekt ülesande täitmisest keelduma.

Nagu tõi esile ka Yudkowsky, on inimeste väärtused ja eelistused ajas muutuvad (Bostrom 2014:271). Näiteks aastasadu tagasi oli orjapidamine aktsepteeritav, kuid tänapäeval enam mitte. See on samuti aspekt, millega superintellekt peaks erinevaid väärtuseid kaalutledes arvestama. Võib-olla on superintellekt niivõrd ettenägelik, et oskab andmeid analüüsides ennustada, kuidas inimeste väärtused ajas muutuda võivad ning oskab sestap arvestada tulevasi suundumusi. Vastupidiselt võib pahatahtlik superintellekt ise inimeste väärtusi või eelistusi suunata varjatud propagandat või osavat turundust kasutades.

Analüüs

Uurides tehisintellekti väärtuste joondamise erinevaid probleeme, peaks normatiivselt TI tegevus olema kooskõlas inimeste väärtuste, ootuste ja eelistustega nii, et see ei kahjustaks inimeste vabadust ja eneseteostust. Samuti peaks TI väärtused olema joondatud viisil, et need ei kaota ära võimalust, et inimeste väärtused või eelistused võiksid ajas muutuvad olla. Lisaks peaks inimestel olema võimalus langetada otsuseid küsimuste osas, mis neid otseselt puudutada võivad. Küll aga, kui tehisintellekt peaks õppima inimese käitumisest, peaks see arvestama vaid kiiduväärset käitumist, sest inimesed on võimelised oma käitumisega teist inimest kahjustama või toimima muul viisil ebaratsionaalselt, näiteks emotsioonide ajendil. Kõige tagatipuks peaksid tehisintellekti väärtused olema mõõdetavad ning masina poolt loetavad.

Kes peaks valima ja kuidas peaks valima tehisintellekti väärtuseid on oluline küsimus tehisintellekti eetika aspektist. Ühest küljest on Yudkowsky koherentne ekstrapoleeritud tahtlus hea algus idu-tehisintellekti eetika teooriale, teisest küljest aga sisaldab see liialt poeetilist vormi, mis võib arvutisüsteemi vaatest lünklikuks osutuda. Teisalt ei saa ilma eetikateooria algeteta lasta superintellektil välja mõelda, milline on parim võimalik eetikateooria tulevikuks, nagu Bostrom seda välja pakub. Üks võimalik lahendus oleks olukord, kus superintellekt saab valida võimalike eetikateooriate vahel. Selle vaate kitsaskohaks võib osutuda, et superintellekt langetab kõikide moraaliteooriate vahel ootamatu valiku, aga ka see, et SI eetika on tuleviku osas lukustatud mitmete teooriate hulka, mis võivad olla vaieldavad või inimtasemel genereeritud ning sestap superintellekti jaoks piiravad. Samuti ei saa pluralistliku TI korral selle kasutajad langetada valikut teooria osas, mida nad eelistavad, et TI võiks kasutada. Milline lähenemine rahuldaks kõikide osapoolte vajadusi? Kuidas saavutada tasakaal mitmete teooriate ja eri inimeste häälte vahel?

Polonski mainis oma kirjutises MIT Moraalse Masina projekti, mis hõlmab kontseptsiooni, kus inimesed saavad anda omapoolset sisendit selle kohta, kuidas isesõitev auto peaks teatud situatsioonides käituma. Näiteks kas liiklusõnnetuse korral, mida pole võimalik ära hoida, peaks isesõitev auto pigem otsa sõitma teed ületavale kolmele isikule? Või peaks auto muutma suunda ning potentsiaalselt seadma ohtu kolm autos viibivat inimest? Niisugust sisendi andmise süsteemi on võimalik kohandada eetiliste küsimuste perspektiivi, mis võimaldaks potentsiaalselt lahendada erinevate vaadete ja vajaduste probleeme. Näiteks küsitlusega sisendatud väärtuste alusel oleks kaetud eri eetilised väärtused teatava mõõdetava parameetriga.

Kujutame ette sama superintelligentset robotit, mida mainisin varem, mis aitab inimesi igasuguste probleemidega, küpsetab toite, aitab lastel õppida, annab nõu jne. Enne superintelligentse roboti väljastamist peaks igaüks, eeskätt eksperdid nii TI valdkonnas, kui ka erinevad eetikud, sotsiaalsed töötajad ja muud asjakohased spetsialistid, täitma ära SI kasutamise küsimustiku, mis kujundab globaalselt robotite eetikateooriat ning seda, kuidas inimesed ootavad, et see käituks. Küsimustikul oleks igal inimesel üks hääl ning selles esitletakse osalejatele erinevaid sotsiaalseid olukordi või probleeme, millele inimesed annavad oma reaktsiooni teatud parameetrites, näiteks tõene(1), peamiselt tõene(2), neutraalne(3), peamiselt väär(4) ja väär(5). Eeldame eksperimendi mõttes, et niisugusest küsimustikust jäävad vaikimisi välja negatiivsust külvavad küsimused, näiteks „Kas soovid et robot oleks sinu vastu kuri?“. Pigem on küsimustikus esitletud erinevad eetilised probleemid, näiteks tuntud trammi probleem, „Tramm, mida ei saa peatada, liigub rajal, mis lahkneb kaheks. Ma olen nõus sellega, et robot ei käitu, mille tulemusena kaotab elu viis inimest, selle asemel, et muuta trammi suunda, et viie inimese asemel kaotaks elu üks.“. Sellele saab igaüks anda oma sisendi, näiteks „Väär(5)“, mis on tugev mõõdetav eelistus alternatiivide, näiteks „Peamiselt väär(4)“ ees. Niisuguseid küsimusi on võimalik formuleerida lõputult, kuid oluline oleks leida piisavalt täpselt mõõdetavad probleemid ning nende tulemused sisendi abil mõõdetavaks muuta. Siinkohal pole eesmärgiks vastata igale spetsiifilisele juhtumile, mida superintellekt kohata võib, vaid pigem luua kollektiivselt selle moraalne kompass.

Selline MIT Moraalse Masina projekti alusel konstrueeritud küsimustiku lähenemine ei ole kindlasti veatu. Kahtlemata kujuneb adekvaatse ja põhjapaneva küsimustiku koostamine raskeks ülesandeks, mis määrab ettenähtavaks tulevikuks superintellekti eetikateooria. Lisaks võib see küsimus koosneda potentsiaalselt piiramatust hulgast küsimustest, situatsioonikirjeldusest või eetilisest probleemidest, et olla piisavalt põhjalik. Järgnevalt toon esile potentsiaalse kriitika oma küsimustiku ideele, mis koosneb kolmest osast: jumala rolli probleem, vastutusest vabanemise probleem ja vähemuse probleem.

Jumala probleem

Samuti on küsimustikule vastajad üsnagi otseselt jumala rolli seatud, sest iga sisend määrab meist märkimisväärselt targema olevuse tulevasi väärtuseid ja seeläbi ka käitumisviise. Paljud eetikaküsimused, sealhulgas eelnevalt mainitud trammi probleem on otseselt valimine mitme

potentsiaalse inimese elu ja surma vahel. Samuti võib näiteks juhtuda, et küsimustikule vastajad ei võta seda tõsiselt ning sisestavad naljatledes objektiivselt erinevaid vastuseid, mis lähevad vastuollu sellega, mida nad päriselt mõtlevad. Kuidas kujuneks tulevik juhul, kui suurem osa superintellekti eetika osas otsustajate sisenditest oleks pelgalt lõõpimine? Eeldatavasti pole see tulevik, mida inimkond soovib näha.

Vastus: Mõõdetavad parameetrid ja esitatavad küsimused peavad olema piisavalt täpsed, et eksimisruumi tekkimine ei ole võimalik. Kui on olemas objektiivselt õigeid ja valesid sisendeid, peaks süsteem kasutama pigem neid, mis loovad rohkem väärtust inimkonnale nii praegu kui ka tulevikus. See mõte sarnaneb mingil määral Yudkowsky sõnastatuga ning eeldab ideaalset olustikku ja inimesi nende parimas võimalikus vormis, nendena, kes „me tahame, et me oleks". Samuti peab lisama, et küsimustiku võimalus on küll raske ettevõtmine, aga mitte võimatu. Lisaks avaks selline lähenemine potentsiaali mõõdetavate parameetrite jaoks, mis on arvutile loetav.

Vastutusest vabanemise probleem

Samuti vabastab küsimustiku kasutamine eetikateooria kujundamisel superintellektiga tegeleva asutuse teataval määral vastutusest. Kui superintelligentne robot peaks tegema otsustamisel vea, millel on ülevoolavalt negatiivsed tagajärjed, oleks selle roboti loojatel võimalik ennast välja vabandada lihtsa põhjendusega, näiteks „kasutajad ise valivad, kuidas robot käitub - siin ei ole roboti loojatega seost ning me ei saa vastutust võtta". See oleks problemaatiline mitmel moel: esiteks, kui roboti tehtud valik on objektiivselt vale teatud olukorras, mida oleks saanud paremini lahendada, on kahtlemata otsustusprotsess mingil viisil vigane. Kes selle eest vastutust peaks kandma? Kas ühiskond, roboti looja või hoopis robot ise?

Vastus: Mitmed, ka näiteks seadusega määratletud süsteemid võivad olla mingil määral ähmased. Näiteks liiklust reguleerib liiklusseadus, mis näeb ette, et enamjaolt igal juhul on liiklusõnnetuses süüdi see, kes sõitis teisele tagant sisse. Küll aga on sellel erandeid, näiteks juhud, kus liiklusõnnetuse põhjustas see, kellele tagant sisse sõideti, sest ta sooritas meelega äkkpidurduse. Siinkohal peab tehisintellektiga seotud eetika olema piisavalt hästi defineeritud, et kitsaskohtade võimalusi vähendada.

Vähemuse probleem

Veel üks kriitika võib olla, et küsimustiku lahendus ei võta kõigi oma positiivsete omaduste juures arvesse seda, et niisugune demokraatlik valimine soodustab tugevalt enamuse häält ning

ei pruugi rahuldada vähemuse vajadusi. Näiteks oleks raske LGBT liikmetel, veganitel või invaliididel oma häält kuuldavaks teha.

Vastus: Siin on oluline arvestada asjaoluga, et küsimustiku vorm peaks välja tooma tugevad või vankumatud eelistused mingi valikuvariandi osas. Eelnevalt tõin näite, kus valikus olid väärtused 1-5, küll aga võib see olla ka 1-7, 1-9 või muu. Siinjuures võib olla 90% enamuse valiku kaal näiteks vähesel määral positiivne või isegi neutraalne, aga 10% vähemuse valiku kaal väga tugevalt negatiivne. Statistika peaks olema sealjuures piisavalt hästi läbinähtav, et superintellekt oskaks selle alusel otsuseid teha. Isegi andmestiku ähmasuse korral peaks leiduma viis, kuidas superintellekt olukorraga toime tuleb.

Küsimustiku kasutamise lahendus hõlmab minu hinnangul iga varasema teooria parimaid külgi, näiteks on isetäiustuv nagu KET, hõlmab mitmeid erinevaid moraalilähenemisi nagu pluralism, ning kasutab otseselt Baumi pakutud sotsiaalset valikut. Samuti saab see olla ajas staatiline, mis tähendab, et inimeste eelistuste muutumisel muutub ka SI eetikakoodeks.

Tulevase superintellekti väärtuste joondamine on kahtlemata tähtis probleem, millega tasub tegeleda pigem varem, kui hiljem. See, milliseks kujuneb tulevik, ei ole ettenähtav. SI pioneeriks võib saada mistahes korporatsioon või isegi üksikisik, kes esimesena selle on võimeline saavutama. Sestap on olemuslikult väärtuslik, et probleemi käsitlus oleks kõneaineks ning leiaks kajastust.

Kokkuvõte

Käesoleva bakalaureusetöö peamine eesmärk on tuua esile superintellektiga seotud väärtuste seadistamise suurimaid probleeme ning esitleda probleemseid aspekte varasemate teooriatega, näiteks koherentse ekstrapoleeritud tahtluse teooriaga, aga näiteks ka ettepanekuga, et superintellekt ise väärtusteooria välja mõtleb.

Kirjutise juhatab sisse kirjeldus tehisintellektist kui tehislikust arvutimõistusest ning kuidas sellest võib välja areneda superintellekt. Kui niisugune superintelligentne agent peaks läbima intellektiplahvatuse, mille vältel see muutub targemaks kui kõige targemad inimesed maailmas, võib sellega kaasneda suuri ohtusid inimkonnale. Välistatud ei ole, et superintellekt on inimeste suhtes vaenulik või ükskõikne. Sellepärast on oluline, et inimeste ja superintellekti väärtused oleksid omavahel kooskõlas.

Niisuguse kooskõla saavutamiseks on mitmeid lahendusi välja pakutud, näiteks on Yudkowski koherentne ekstrapoleeritud tahtlus üks kõneainet tekitanud võimalustest. Küll aga võib teooria olla liialt poeetiline, et arvutisüsteem sellest aru saaks. Samuti on levinud mõte, et superintelligentne süsteem võib ise oma väärtused välja mõelda, kuid kui võtta arvesse eesmärki, et inimeste huvides on väärtused superintellektiga kooskõlla viia, peaks enne superintellektile niisuguse ülesande andmist juba teatud väärtused sellele seadistatud olema. Samuti vaatlesin pluralismi kui võimalikku lahendust väärtuste seadistamise probleemile. See hõlmab, et SI on võimeline valima teatud olukorras parima võimaliku eetikakoodeksi. Küll aga eeldab see, et inimesed lasevad superintellektil valida eetikateooriate seast, mis on inimmõistuse tasemel välja töötatud ning võivad superintellektile piiravad olla. Lisaks võivad inimeste eetikateooriad olla teatavate eksijäreldustega või aja jooksul muutuda.

Lõpetan kirjutise analüüsiga, kus toon esile peamised probleemid varasemate väärtuste joondamise lahendustega ning modifitseerin mõtteeksperimendina MIT Moraalse Masina projekti baasil väljatöötatud küsimustikku nii, et seda saaks kasutada tehisintellekti väärtuste kujundamiseks. Nii on võimalik kaasata nii tehisintellekti alaseid spetsialiste, kui ka TI kasutajaid, et luua ühiselt SI eetikakoodeks. Toon oma lahenduse juures esile ka potentsiaalse kriitika ning proovin kriitika seejärel ümber lükata.

Viited

Arcas, B., Norvig, P. 10. Oktoober 2023. "Artificial General Intelligence Is Already Here". *Noema. Vaadatud 14.04.2024, aadressilt https://www.noemamag.com/artificial-general-intelligence-is-already-here/*

Baum, S. 2020. "Social Choice Ethics in Artificial Intelligence", *Springer, AI & Society, Springer Science+Business Media, Brighton, lk 165-176. Vaadatud 3. Jaanuaril 2024, aadressilt https://link.springer.com/article/10.1007/s00146-017-0760-1#citeas*

Bostrom, N. 2014. "Superintelligence: paths, dangers, strategies". *Minds and Machines, Oxford.*

Eesti Keele Instituut, *Võõrsõnade Leksikon. Otsisõna „Tehisintellekt". Vaadatud 25. märtsil, 2024, aadressilt https://www.eki.ee/dict/vsl/index.cgi?Q=tehisintellekt*

European Parliament. 4. September 2020. *Artificial Intelligence. „What is artificial intelligence and how is it used?". Vaadatud 1. mail 2024, aadressilt https://www.europarl.europa.eu/topics/en/article/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used*

Polonski, V. 19. Detsember 2017. "Can we teach morality to machines? Three perspectives on ethics for Artificial Intelligence". *University of Oxford. Vaadatud 8. Jaanuaril 2024, aadressilt https://www.oii.ox.ac.uk/news-events/can-we-teach-morality-to-machines-three-perspectives-on-ethics-for-artificial-intelligence/*

Russell, S. 2019, Human Compatible. *AI and the Problem of Control, London: Allen Lane, Penguin Books.*

Sutrop, M. Detsember 2020. "Challenges of Aligning Artificial Intelligence with Human Values". *Acta Baltica Historiae et Philosophiae Scientiarum 8(2):54-72. Vaadatud 16. Jaanuaril 2024, aadressilt https://www.researchgate.net/publication/346964247_Challenges_of_Aligning_Artificial_Intelligence_with_Human_Values*

Tegmark, M. 2017. "Life 3.0: being human in the age of artificial intelligence". *Alfred A. Knopf. New York.*

Urban, T. 22. Jaanuar 2015. "The AI Revolution: The Road to Superintelligence". *Vaadatud 3. Jaanuaril 2024, aadressilt https://waitbutwhy.com/2015/01/artificial-intelligence-revolution- 1.html.*

Resümee

The superintelligence value alignment problem

Technological advancements in the 21st century have brought forth a debate about artificial intelligence and how to align the values of AI in a way that future superintelligent agents would come to share values with human beings. Yudkowsky's coherent extrapolated volition is one of the prime examples of theories for AI value alignment, however, as this thesis suggests, many of these theories come with some pitfalls, which makes value alignment an extremely complex task. For one, if some ethical theories would be used, it would run the risk of the AI being locked down to a single, potentially flawed, ethical theory, forever. Pluralism may be a valid solution, as then the AI could make a choice regarding the best moral theory to apply to certain scenarios. Some have also suggested that a superintelligent AI system could come up with a new ethical theory. This could, however, be risky, because the AI would likely need to have some moral base to make the values align with humans. I propose a system similar to the MIT Moral Machine project, where the users, as well as experts of artificial intelligence can participate in the value alignment of an AI system.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karl Vilu, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Superintellekti väärtuste joondamise probleem", mille juhendaja on Bruno Mölder, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonnas, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Karl Vilu
17.04.2024