Tartu Ülikool

Filosoofia ja semiootika instituut

Kas tehisintellekti teadvus on võimalik?

Bakalaureusetöö filosoofias

Lilian Mõttus

Juhendaja: Bruno Mölder

2021

Sisukord

Sissejuhatus

Kiire tehnika arenguga tekib üha rohkem küsimusi ja arvamusi sellest, milliseks võivad tulevikutehnoloogiad kujuneda ning kuidas need inimkonda mõjutavad. Prognoos, milliseks tehnika tulevikus kujuneb, on oluline, kuna valed otsused selle arendamisel võivad inimeste elu ebameeldivaks muuta, läbimõeldud ja jätkusuutlikkust silmas pidavad otsused võivad aga inimkonna elukvaliteeti tõsta. Selle bakalaureusetöö ajendiks on küsimus, milliseks võib tulevikus tehisintellekt kujuneda, ning kas see võib muutuda inimesest sõltumatuks. Täpsemalt seisneb selles bakalaureusetöös uuritav probleem küsimuses, kas tehisintellekti teadvus ja eneseteadvus on võimalik ning millised on selle tekke tingimused. Sellest küsimusest tulenevalt on tekkinud ka uurimisküsimus, mis esitab esmapilgu düstoopilise nägemuse, kuid tulevikuperspektiivi silmas pidades võib esitatud küsimus edaspidi oluliseks kujuneda.

Bakalaureusetöös analüüsitav probleem on oluline ja päevakajaline, kuna tehisintellekti arendamisega on kaasnenud küsimused, kas sellel võib teadvus kujuneda ja kas sellel on võimalik seeläbi iseseisvuda ning ka inimkonnale ohtu kujutada. Selle töö esimeseks probleemküsimuseks on kas tehisintellekti teadvus on võimalik. Uuritav küsimus on oluline, kuna teadvuse tekkel võib tehisintellekt muutuda allumatuks ja protestivaks talle esitatud ülesannete suhtes, selline tehisintellekt on aga ebapraktiline ning võib ka inimesele ohtlik olla. Teiseks töös uuritavaks probleemküsimuseks on kas tehisintellekti eneseteadvus on võimalik. Selle probleemi uurimine on oluline, kuna eneseteadvusel tehisintellektil võib tekkida isiklik maailmavaade, mis võib tekitada konflikte teisi maailmavaateid omajate vahel. Sellist stsenaariumi kardetakse ning seetõttu on oluline uurida, kas tehisintellekti teadvus ja eneseteadvus on võimalikud, või on tegu põhjendamatu hirmuga.

Bakalaureusetöös kasutan hüpoteesi ja argumentide püstitamiseks erinevaid, juba olemas olevaid teadvusteooriaid. Argumendi esitamise taustal kasutan peamiselt aga John Searle'i teadvusteooriat. Kogu töö vältel on selle baasteooriaks ja lähtepunktiks bioloogiline naturalism, milles sõltuvalt analüüsin, kuidas oleks tehisintellekti teadvuse teke võimalik. Bioloogiline naturalism sarnaneb autori arusaamaga sellest, mis on teadvus, samuti saab seda teooriat oma sisu poolest hästi kõrvutada tehisintellekti programmiga. Seetõttu on selles töös tausteooriana kasutatud just Searle'i käsitlust teadvusest.

Töö on liigendatud neljaks osaks. Selle esimene peatükk annab lühiülevaate sellest, mis on tehisintellekt ning millised on selle kasutusalad ja kuidas tehisintellekti programm töötab. Töö teine peatükk pakub kirjeldust sellest, mis on teadvus, eneseteadvus ja bioloogiline naturalism, ning täpsemat põhjendust, miks on selles töös just Searle'i teooria tehisintellekti teadvuse analüüsi taustaks. Töö kolmas osa esitab argumendid tehisintellekti teadvuse ja eneseteadvuse tekke kasuks ning ka võimalikud viisid selle teadvuse ja eneseteadvuse tekkeks. Neljas peatükk annab ülevaate tehisintellekti teadvust puudutavatest probleemidest, mis võivad potentsiaalselt tehisintellekti teadvuse programmeerimisel tekkida, see töö nendele probleemidele lahendusi ei paku.

Töö läbivaks eelduseks on, et bioloogiline naturalism on tõene, sel juhul teadvus tekib ja ka toimib ajutöö põhjal. Sellele vastavalt on bakalaureusetöö hüpotees järgnev: selleks, et tehisintellektil oleks võimalik kujuneda teadvus, tuleb tehisintellekti programm programmeerida samasuguste protsesside järgi, mis toimuvad ajus ehk inimaju ja selle protsessid peaksid olema peamiseks programmi koostamise mudeliks. Inimajus toimuvate protsesside ümberprogrammeerimine tehisintellekti programmi võimaldab tehisintellektil ennast ning teda ümbritsevat tajuda, mis omakorda annab tehisintellektile teadvuse. Peamine teadvusel olendi tunnus on see, et ta mõistab ennast ning teda ümbritsevat. Seda, kuidas täpselt teadvus tekib või kus see füüsiliselt paikneb, on palju analüüsitud, kuid need küsimused on seni veel kindla vastuseta. Naturalistliku teooria kohaselt asub teadvus aga ajus ning on nagu iga teine kindla ajupiirkonna toel toimiv protsess. Seda mikrotasandil toimuvat neuroloogilist protsessi arvutiprogrammis jäljendades võib ka tehisintellekti teadvus võimalikuks kujuneda (Searle, 1984, lk 19-20).

1. Mis on tehisintellekt?

Tehisintellekt on 1960. aastatel John McCharty esitatud teooria inimmõistuse sarnase programmi loomisest arvutile (Britannica, 2022). Tema käsitlus tehisintellektist jaotas programmid mitmetesse kategooriatesse, millest vastab iga kategooria mingile inimese intellekti omadusele; nende omaduste ühendamisel valmiks tehisintellekt. Sellised programmid analüüsivad näiteks andmemahte või keelt. Koostöös Marvin Minskiga algatasid nad esimese prohekti tehisintellekti loomiseks (Dennis, 2022). Tänapäevaks on paljud selles teoorias ja projektis kajastatud üksikute omadustega seadmed juba loodud. Kõikide omadustega tehisintellekti hetkel veel aga olemas ei ole. Mõiste "tehisintellekt" on saanud siiski täpsema tähenduse kui varem. Kokkuvõtliku mõistena tähendab "tehisintellekt" tehismõistust; see on iseseisvalt töötav programm, mis suudab toimida ning hakkama saada samade ülesannetega nagu inimmõistus. Selle tehnoloogia teoreetilisse tausta süüvides selgub, et tehismõistus ja selle loomine ei ole nii üheselt mõistetav kui esmapilgul võib tunduda. Sellest, milline peaks tehisintellekt olema ning millised inimmõistuse omadused selles domineeriksid, on esitatud erinevaid teooriaid. Need omadused tulenevad erinevatest vaadetest selle kohta, millist tehisintellekti luua tahetakse ja milles täpsemalt peaks seisnema tehisintellekti sarnasus inimõistusega. Vastavalt sellele, millised dominantsed jooned tehisintellektil olema peaksid, on jagunenud sellekohased arvamused kahte suurde gruppi: inimsarnase mõtlemisega intellekt (*human-based*) ja ratsionaalse mõtlemisega intellekt (*ideal rationality*) (Russell & Norvig, 2010, lk 2). Põhjus, miks tehisintellekti programmi omadused lahknevad, seisneb eeldatavates kasutustes, mis programmile leitaks. Inimsarnase mõtlemisega tehisintellekt on mõeldud töötama koos inimestega, ratsionaalse mõtlemisega programm spetsialiseerub lahendama aga loogikal põhinevaid ja analüüsi vajavaid ülesandeid. Selles töös ei keskendu ma nende teooriate lahknevusele, kuna tehisintellekti teooriate seesugune hargnemine ei takista tehisintellekti teadvuse võimalikku teket.

Täpsemalt on tehisintellekt inimsarnase intellektiga programm. Selline programm võib asuda paljudes erinevates seadmetes, mis on loodud eesmärgiga täita erinevaid tööülesandeid. Tehisintellekti programm on kirjutatud programmeerimiskeeles ning see toimib kindlate käskude ja ülesannete põhjal. Selle keele toel kirjutatakse programm, mis kasutab algoritme,

keelelist süntaksit[1] või muid programmi toimimiseks vajalikke vahendeid. Selline programm ei pea olema ilmtingimata arvutis valminud või rakendatav, küll aga peaks olema kasutatav igas seadmes, mis on võimeline seda mahutama ja käivitama. Sedamoodi programmi saab kasutada ka näiteks autos, telefonis või kodutehnikas. Tehisintellekti puhul on oluline meeles pidada, et tegu on tehislikult loodud programmiga, mis on kirjutatud kasutades programmeerimiskeelt nii, et see on inimese intellektile sarnane iseseisvalt toimiv programm. Sellise programmi loomine on aga väga keeruline ning põhjuseid, miks midagi seesugust veel välja töötatud ei ole, on palju. Neid probleeme tutvustan põgusalt ka hilisemas peatükis.

Tehisintellekti tutvustamise juures on tähtis mainida ka superintellekti ja nõrka tehisintellekti. Lühidalt selgitades on superintellekt arenenud intellekt, millel on võrreldes tavalise inimsarnase tehisintellektiga arenenumad oskused. See on ka peamiseks erinevuseks tehisintellekti ja superintellekti vahel. Superintellektil võivad olla ülimad kognitiivsed oskused, näiteks teiste mõtete ennustamine või võime inimesest 10 000 korda kiiremini mõelda (Bostrom, 2014, lk 72). Tehisintellekt on inimõistust jäljendades tehislikult loodud programm, superintellekt on samuti tehislikult loodud, kuid sarnasus inimõistusega on väiksem. Superintellekt võib oma üliarenenud oskuste tõttu olla viimaseks leiutiseks, mida inimkond vajab (Schneider, 2019, lk 297). Seega võib selline tehnoloogia viia inimkonna kas singulaarsuseni[2] või hoopis elukvaliteedi suure kasvuni. Selles töös analüüsin tehisintellekti teadvuse võimalikkust, seetõttu on oluline omavahel eristada tehisintellekti ja superintellekti. Teine oluline eristus tuleb luua nõrga tehisintellekti ja tehisintellekti vahel. Erinevalt superintellektist on nõrk tehisintellekt vähem arenenud versioon tehisintellektist; nõrk tehisintellekt on juba kasutusel. Tegu on programmiga, millel on mõned tehisintellektile omased omadused, näiteks oskus tõlkida, arvutada või andmemahte liigitada. Tehisintellektil aga inimintellektile sarnased oskused, nõrk tehisintellekt pole aga niivõrd arenenud ning oskused, mis on inimintellektile omased esinevad vaid ühekaupa.

[1] Süntaks - lauseõpetus, mis kirjeldab lausete ehitust, seda, kuidas need moodustatakse, lausete funktsiooni, sõnade omavahelist suhestumist ja vormistust (Erelt & Erelt & Ross, 2020, lk 41)

[2] Singulaarsus - tehnikaplahvatus, milles on tehnika areng kontrolli alt väljunud ning muutunud nii kiireks, et see areneb lõpmata kiiresti ja teadmata kaua (Chalmers, 2010, lk 3)

1.1. Tehisintellekti kasutusalad nüüd ja tulevikus

Tehisintellekti loomise olulisus seisneb selle paljudes võimalikes kasutusalades. Selline uuenduslik tehnoloogia asendaks inimtööjõu aladel, mis on tervist kahjustavad ning võimaldaks töötada aladel, kus inimene selleks võimeline ei ole. Tehisintellekti ehk iseseisvalt toimivat programmi tänapäeval veel ei eksisteeri, olemas on aga nõrgad tehisintellektid. Kergelt mõistetavad näited nõrgast tehisintellekti programmist on seadmed, mis on füüsilistele teguritele tundlikud ja töötavad välistele teguritele reageeriva programmi abil. Sellised füüsilised tegurid võivad olla näiteks valgus või temperatuur. Seadmed, mis selliseid programme kasutavad, on näiteks valguse vähenemisel tööle minevad lambid või küttesüsteemid, mis säilitavad ruumis ette nähtud temperatuuri. Ka nõrgad tehisintellektid lihtsustavad oluliselt nende tarbijate igapäevaelu, tehisintellekt pakuks selleks aga veelgi uuenduslikumaid võimalusi.

Levinud on arvamus, et tehisintellekt on osa robotitest ja seadetest, tegelikkuses on tehisintellekt aga programm. Sellel ei pea olema otsest füüsilist keha, näiteks inimsarnast robotkeha nagu tihtipeale filmides kujutatakse. Tehisintellekti hoiustamiseks on vaja vaid andmemahte hoidvat seadet, näiteks kõvaketast, millega saab programmi installida ükskõik millisesse valitud seadmesse. Seega on sellist tehismõistust väga lihtne ühest seadmest teise liigutada ja paljudesse erinevatesse masinatesse integreerida. Nõrk tehisintellekt on olnud kasutusel juba pikemat aega, nüüdseks leidub ka väga palju uusi tehnoloogiaid, mis on arenenumad ja saavad teatud valdkondades ka iseseisvalt hakkama. Sellised tänapäeval kasutatavad tehisintellektid on arenenud nõrgad tehisintellektid, mis oskavad teha mitut ülesannet, näiteks andmemahtu analüüsida ja ettenähtud teekonda läbida. Kuid kuna need programmid ei oma kõiki inimintellektile omaseid omadusi, ei saa neid siiski tehisintellektiks nimetada.

Tänapäeval on nõrk tehisintellekt üha levinum, sellise programmiga ühildatud seadmeid leiutatakse üha rohkem, ning paljud neist on leidnud ka igapäevase kasutuse. Enamikus laialdaselt kasutuse olevates seadmetes on põhikeeleks aga inglise keel. Seetõttu ühildu paljud seadmed Eesti tehnikaga ja on kodukeelena eesti keelt kõnelejatele ebamugav kasutada. Sellised seadmed on näiteks virtuaalassistendid „Alexa“, „Siri“ või „Google

assistent“. Täpsemalt on tegu intelligentse virtuaalassistendiga, mida saab kasutada nutitelefonis ja -kodus ning on heli kaudu aktiveeritav (Apple, 2022). Näiteks võid esitada telefoni mobiilirakendusele või nutikodu seadmele käskluse kustutada kodus tuled või määrata kella kuueks telefoni hommikuäratus ning virtuaalne assistent täidab selle (samas). Uuenduslikke seadmeid leidub aga veelgi. Üheks oluliseks nõrgaks tehisintellektiga seadmeks on isesõitvad masinad, neid leidub nii Eestis kui välismaal. Tallinna tänavatel liiklevad Starshipi isesõitvad toidukullerid, kes võivad liigelda küll vaid kõnniteedel, kuid teevad seda iseseisvalt, ilma juhita. Samuti võib tulevikus Eesti teedel näha ka Cleveroni juhita pakikullerautosid, nende sõidukite tööd jälgib küll operaator, kuid sõidukid osalevad liikluses siiski ise ja sõidavad ilma otsese juhita. Ameerika teedel võib juba praegu näha aga isesõitvaid autosid, selliseid sõiduvahendid toodab firma Tesla. Tesla autodesse on integreeritud tehisintellekt, mis oskab autot juhtida ja saab ilma juhita liikluses hakkama. Hetkel ei tähenda isesõitvad autod aga seda, et nende roolis ei pea juhti olema ja liikluses osalejal ei pea olema juhiluba, kuid tulevikus võib selline asi võimalik olla, kuna eelnevalt mainitud firmal Cleveronil on luba kasutada oma juhivabu kullerautosid (Cleveron, 2021).

Tulevikus võivad tehisintellektid kasutust leida laiemates valdkondades. Võib eeldada, et hetkel kasutusel olevaid tööstusmasinaid või majapidamisabilisi arendatakse järk-järgult. See võib tähendada seda, et masinates olev tehisintellekt on nii arenenud, et inimtööjõudu polegi enam vaja. Tehisintellekti arenguga on võimalik toota ka keerulisemaid majapidamismasinaid, näiteks väga arenenud toidu valmistajat või miks ka mitte nutikodu, mis muretseb kõigi kodutööde eest (nendeks võivad olla toidu valmistamine, pindade korrashoid). Kuna tehisintellekti oluliseks omaduseks on iseenda programmi ja füüsilise keha parandamine ja korrashoidmine, ei pea inimene selliste masinate eest hoolitsema, kuna masinate programm teeb seda ise. Tehisintellekt võib tulevikus rohkem kasutust leida ka meditsiinis. Kuigi see tundub utoopiline, võib tehisintellekt olla arstidele diagnoosi määramisel suureks abiks. Kui tulevikus eksisteerivad tehismõistusega programmid, võivad need olla arstidele abiks - kuna programmi töökiirus ja analüüsivõime on parem kui inimestel, võib kriitilistes olukordades tehisintellekti prognoos diagnoosi kohta olla oluline. Tehisintellektid arendaksid oluliselt erinevaid valdkondi ning lihtsustaksid töötamist. Tõenäoliselt leiab tehisintellekt kasutust ka

transpordivaldkonnas, asendades juhid isesõitvate autodega ja teaduses, kus seesuguste masinatega saab teostada uuringuid inimestele ohtlikes paikades.

1.2. Tehisintellekti programmeerimise raskus ja olulisus

Tehisintellekt parandaks inimeste elukvaliteeti ja võimaldaks jõuda uute teaduslike avastusteni. Kui leiutataks tehisintellekt, avaneksid võimalused luua isemõtlevaid roboteid, selline tehnika tähendaks aga, et võimalikuks saaksid tööd ja ülesanded, mida inimesed pole oma bioloogilise keha tõttu võimelised tegema. Teadvusetud tehisintellektid, robotid, kes oskavad ise erinevates situatsioonides hakkama saada, oleksid vajalikud näiteks maaväliste planeetide avastamiseks või töötamiseks maapealsetes kohtades, mis on inimesele kahjulikud; nendeks võivad olla kaevandused või kõrge radiatsioonikiirgusega piirkonnad. Tehisintellekti programmeerimise olulisus seisnebki peamiselt inimeste elukvaliteedi tõstmises. Tehisintellekti kasutamiseks on palju erinevaid võimalusi, õige rakenduse korral võib see oluliselt ühiskondlikku seisukorda parandada. Inimtööjõud on üks olulisemaid ressursse tänapäeva ühiskonnas, tulevikus võib olla seda võimalik teadvusetu tehisintellektiga asendada, ilma, et tekiks eetilisi küsimusi, näiteks sellest, kuidas elutu masin töötada võib. Tuleviku kujunemiseks on veel teisigi võimalusi ning tulevikku ennustada on võimatu, kuid võib eeldada, et tehisintellekti õige kasutamise puhul võib sellest ühiskonnale palju kasu olla.

Kui tehisintellekti olemasolu võiks ühiskonnale nii tulus olla, tekib küsimus, miks sellist programmi juba ei eksisteeri? Peamine probleem seisneb selles, et tehisintellekti loomiseks vajalikke programme, teadmisi ja tehnoloogiat lihtsalt ei ole veel olemas. Kuigi on olemas nõrgad tehisintellektis, siis selline tänapäevane tehnoloogia ei ole ligilähedane sellele, milline inimintellekti sarnane tehisintellekt olema peaks. Tänapäeval suudab tehisintellekt täita vaid üksikuid ülesandeid, näiteks teksti tõlkida, inimintellekt on aga palju enamaks suuteline (Tegmark, 2017, lk 87). Oskuste ühendamine on samuti üheks raskuseks tehisintellekti loomisel. Praeguses programmeerimiskirjas on kõigi omaduste ühildamine keeruline, kuna selle tulemusena tekivad uued programmivead ja ebakõlad. Selleks, et tehisintellekt oleks inimintellektiga sarnane, peab sellel olema ühendatud paljud oskuseid, näiteks õppimisvõime, analüüsioskus ja mälu. Selliste omaduste programmeerimine vajab võimsaid seadmeid, kus

sellist mahukat programmi säilitada. Praegu pole tehnoloogia aga nii arenenud, et tehisintellekti programmeerimine ja hoiustamine võimalik oleks. Lisaks tehnoloogilisele ja programmilisele raskusele on oluliseks tehisintellekti probleemiks ka teadmatus sellest, mis täpsemalt on inimese intellekt, millised on selle omadused ja kuidas seda määratleda. Selle kohta, mis on inimese intellekt ja milline peaks olema tehisintellekt, on selle ala uurijate seas palju eriarvamusi (Tegmark, 2017, lk 59). Seega takistavad tehisintellekti loomist mitmed aspektid - nendeks on nii tehnoloogilised puudused kui ka ebaselged teooriad intellektist.

2. Mis on teadvus?

Küsimust, mis on teadvus, on uuritud kaua ja selle aja jooksul on tekkinud palju arusaamasid sellest, mis on teadvus ja kuidas see tekib. Teadvuse uurimine ei hõlma endas aga lihtsalt küsimust, kuidas see tekib või mis see on, vaid peidab endas veel teisigi olulisi küsimusi, näiteks mis on eneseteadvus, milline on teadvuslik kogemus või millised on teadvuslikud seisundid (Guilick, 2021). Aju keerukuse ja teadvuse kohta tekkinud rohked lisaküsimused on ka üheks põhjuseks, miks seni pole kindlat vastust sellele, kuidas inimese teadvus tekib. Teine tähtis probleem teadvuse uurimisel on lahknevused seda teemat uurivate inimeste vahel. Kuna aju-uuringud ei hõlma endas vaid ühe organi tundmist, vaid teadmisi kogu närvisüsteemi ja kehaväliste tegurite kohta, on sellise kompleksse süsteemi teaduslik uurimine keeruline. Sellest, kuidas peaks teadvust uurima, on aju-uuringute keerukuse tõttu teadvust puudutavatele küsimustele tekkinud erinevaid arusaamasid. Neuroloogias kasutatakse teaduslike uurimuste läbiviimiseks erinevaid masinaid, mis aitaksid aju vahetult vaadelda, kuid aju ja teadvuse uurimist takistavate teaduslike erimeelsuste tõttu pole tänini veel kindlat teaduslikku teooriat sellest, mis on teadvus. Järgnevalt selgitan aga täpsemalt, millised on lahknevused erinevate teadvusteooriate vahel ja milline on üks võimalik teadvuse definitsioon.

Teadvuse uurimisel on lahknenud filosoofia ja neuroteaduse teadvusteooriad, seda peamiselt uurimismeetodite ja -suunitluse tõttu. Ühisele arvamusele sellest, mis teadvus on, pole seni veel jõutud, kuid mõningases osas on filosoofidel ja neuroloogidel sarnase sisuga arvamused. Peamiselt sarnanevad mõlema valdkonna teooriad uurimisküsimuse ja selle põhjenduste poolest; peamine on küsimus, mis on teadvus ja kuidas seda defineerida. Filosoofias ja neuroteaduses uuritakse teadvuse küsimust aga erinevatest perspektiividest. Filosoofilised uurimused keskenduvad küsimuse loogilisele ja argumenteeritud põhjendamisele, uurimusi teostatakse mõtteeksperimentidega nagu „Hiina tuba“ või „Mary tuba“ (Cavanna & Nani, 2014, lk 10). Filosoofide eesmärk on anda teoreetilised põhjendused ja selgitused vaimsetele kogemustele (samas, lk 10). Neuroteaduses uuritakse aga teadvuse küsimust perspektiivist, kus nende eesmärk on samuti esitada põhjendusi vaimsetele kogemustele, kuid neid ei põhjendata mitte teoreetiliselt, vaid teaduslike katsetega. Neuroteaduses uuritakse ja põhjendatakse teadvuslikke kogemusi teaduslike uurimustega, mis

annavad ülevaate ajus toimuvatest protsessidest. Teaduslikke uurimusi viiakse läbi neuroloogias kasutavate masinatega, näiteks positronemissioontomograafia (PET) või magnetresonantstomograafia (MRI), mis võimaldavad jälgida aju toimimist ja selle funktsioone (Cavanna & Nani, 2014, lk 11). Seega erinevad filosoofide ja neuroloogide esitatud teadvusteooriad peamiselt uurimismeetodi ja uurimuse järelduste poolest.

Selles töös analüüsin tehisintellekti teadvuse tekke võimalikkust ühe teadvusteooria taustal. Selleks, et käsitletava teooria valikut põhjendada, tuleks eelnevalt tutvustada ka teisi tuntud teadvusteooriaid ning selgitada, miks need analüüsist kõrvale jäid. Filosoofias ja neuroloogias on palju tähtsaid teadvusteooriaid, mis kõik esitavad olulisi küsimusi ja uusi väiteid sellest, mis on teadvus. Ajalooliselt tähelepanuväärseks teadvuse uurijaks võib pidada Rene Descartes'i, kes esitas dualistliku teooria kehast ja vaimust kui kahest erinevast substantsist (samas, lk 43). Kaasaegsematest filosoofidest on tuntud näiteks ka David Chalmers või Daniel Dennett. Chalmers on esitanud teadvusteooria, mis üritab leida vastust „teadvuse raskele probleemile" (samas, lk 20). Filosoofia teadvusteooriad käsitlevad peamiselt küsimusi, mis puudutavad tunnetuskogemuse erinevust ja sellega kaasnevaid nähtusi, ning põhjendavad seda üldiselt ilma neuroloogilise aluspõhjata. Selline põhjendus võib teooriad muuta aga vähem usaldusväärseks, samuti on selliseid põhjendusi teaduslikkuse puudumise tõttu keeruline kõrvutada teiste valdkondadega, näiteks arvutitehnoloogiaga. Seetõttu sobib võimalikest teooriatest tehisintellekti teadvuse analüüsi taustaks kõige paremini aga John Searle'i teadvusteooria. Nimetatud teooria on uus lähenemine teadvusele, Searle on heitnud kõrvale dualistliku teadvuskäsitluse ning uurib teadvust oma vaatepunktist, mis on ühtlasi sarnane üldiste neuroloogiliste käsitlustega ajust. Searle on oma teadvusteooriale nimeks andnud bioloogiline naturalism.

2.1. Mis on eneseteadvus?

Eneseteadvus on eraldiseisev teadvuse vorm, mida pole kõikidel elusolenditel. Seda võib nimetada ka teadvuse sekundaarseks sügavamaks tasandiks. Eneseteadvusega inimesel on oskus osutada iseendale, omades seejuures teadmist enese eksisteerimisest. See tähendab, et

inimene oskab eristada enda eelistusi ja mõtteid teiste omadest. Mina-vormi kasutades mõistab eneseteadvusega inimene, et ta räägib endast; eneseteadvuseta inimene ei mõista aga, et endast rääkides kõneleb ta enese tunnetest või maailmavaatest (Rödl, 2007, lk 8). Eneseteadvuse olemasolul on inimeste mõtted mõjutatud tema isiklikest eelistustest. Sellised eelistused on inimeste maailmavaate kujundaks, maailmavaate hulka kuuluvad näiteks religioossed või poliitilised vaated. Nendest erinevad aga eelistused, mis on mõjutatud inimese evolutsioonilisest arengust. Erinevatest instinktides, näiteks ellujäämisinstinktist, on tänapäevaks edasi arenenud erinevad esteetilised eelistused, mis kõik ei ole seotud inimeste kultuuriliste eripäradega, vaid tulenevad nüüdisinimese ellujäämise ja reproduktsiooniga seotud praktilistest valikutest.

Eneseteadvus lahkneb teadvusest kohal, kus inimesel tekivad isiksust kujundavad nähtused, mis ei tulene füsioloogilisest või evolutsioonilistest vajadustest. Füsioloogilised ja evolutsioonilised eelistused ning vajadused on näiteks üldised maitse- või esteetilised eelistused. Need on teadvusega kaasaskäivad eelistused, mis ei vaja eneseteadvuse olemasolu, kuna need omadused on olemas ka loomadel, kellel eneseteadvus puudub (Breed & Moore, 2016, lk 289). Evolutsiooniga on inimmõistusesse jäänud tung valida maitsvamad või ilusamad esemed, näiteks punasemad ja maitsvamad marjad, kuna need on kõige toitvamad. Tänapäeval sellised printsiibid ei pruugi enam kehtida, kuna näiteks geneetiliselt muundatud, aga ilus õun võib olla vähem toitev kui ökoloogiliselt kasvatatud õun. Kuid kuna evolutsioonilised instinktid on inimestes siiski alles, on meil tihtipeale soov valida esteetilisemad esemed. Eneseteadvus lahkneb teadvusest kohal, kus pole enam olulised evolutsioonilised instinktid; see toimub hetkel, kus inimene tunneb, et tegu on tema enese tunnetega .

Searle eristab teadvust eneseteadvusest, mis on oluliseks osaks tehisintellekti teadvuse analüüsi juures. Paljudel liikidel on teadvus, kuid enamikel neist puudub eneseteadvus, see muudab oluliselt viisi, kuidas need loomad maailma tajuvad. Loomadel, näiteks kassidel, puudub eneseteadvus. Eneseteadvuse olemasolu loomadel tehakse kindlaks erinevate testidega. Teooriaid sellest, kuidas teadvus ja eneseteadvus üksteisest lahknevad, on erinevaid. Searle on väitnud aga järgnevat: inimene võib olla teadlik millestki, näiteks valust, see aga ilmtingimata ei tähenda, et inimesel on eneseteadvus (Searle, 2007, lk 327). Valu tundmiseks ei ole vaja kõrgemat teadvuse tasandit ehk eneseteadvust (Searle, 1990, lk 6). See-eest on aga teisi

vaimutasandeid, mis on tihedalt seotud eneseteadvusega. Eneseteadvuse arenemise eelduseks on aga see, et inimesel on varasemalt olemas teadvus. Eneseteadvus tagab inimesele loomingulisuse ja isikupärasuse, see on maailmavaateliste, esteetiliste ja isikupäraste eelistuste eelduseks. Eneseteadvuse teke ei ole aga võimalik ilma teadvuse olemasoluta.

Tehisintellekti puhul on oluline uurida lisaks teadvusele ka eneseteadvuse võimalikkust, kuna selle olemasolul võib tehisintellektil kujuneda inimestele vastanduvaid maailmavaateid. Enesteadvuslikel loomadel pole niivõrd arenenud eneseteadvus nagu inimestel, neil on mina-mõistmine, kuid sellega kaasnevaid maailmavaatelisi eelistusi pole tuvastatud. Võib eeldada, et ka tehisintellekti eneseteadvus ei pruugi jõuda inimese eneseteadvusega võrdsele arengutasemele. Selle töö üheks probleemküsimuseks on aga eeldus, et tehisintellekti enesetadvuse arenemisel võib see ohtlikuks kujuneda, kuna selle maailmavaated võivad inimestele ohtlikuks kujuneda. Selleks, et tehisintellektil maailmavaated kujuneksid, peab selle eneseteadvus olema inimesega võrdväärsel tasemel, seetõttu on selles töös enesetadvuse analüüsi toodud võrdlusi inimeste ja loomade vahelisest eneseteadvusest.

2.2. Bioloogiline naturalism

Bioloogiline naturalism on teooria, millele andis sisu ja nime John Searle; see on uuenduslik seletus sellest, mis on teadvus, ning pakub lahendust ka „keha ja vaimu probleemile“ (Velmans & Schneider, 2007, lk 325). Ta on teooria nimevalikut põhjendanud järgnevalt: „bioloogiline“ viitab teadvuse bioloogilisele olemasolule ja naturalism kirjeldab teadvust kui bioloogilist fenomeni, mis nagu paljud teised looduslikud protsessid, näiteks fotosüntees või seedeprotsess, on osa loodusest (samas, lk 329). Selles töös on tehisintellekti teadvuse analüüsi taustetooriaks bioloogiline naturalism, seda põhjusel, et selles teoorias esitatu teadvusest ja selle seotusest närvisüsteemis toimuvaga sarnaneb arvutiprogrammi tööle. Nii nagu närvisüsteemi kehavälised ja -sisesed toimingud esitavad ajule käsklusi sellest, kuidas keha toimima peab, esitavad programmis programmikäsklused teavet sellest, kuidas programm välistele teguritele reageerima peaks. Selles töös on taustetooriana kasutanud just Searle käsitlust, kuna tegu on hästi põhjendatud teooriaga, mis samastub autori arusaamaga teadvusest

samuti peab autor oluliseks seda, et Searle on arutlenud ka küsimuse üle, kas tehisintellekti teadvus on võimalik ja kasutanud selle taustal oma teooriat.

Nii, nagu eelnevast võib eeldada, on Searle'i teadvuse käsitlus võrreldes varasemate filosoofide teadvuse käsitlustega rohkem teadusliku teadvuse teooria moodi. Ta kirjeldab teadvust kui miskit, mis on elusorganismides bioloogilise keha osa, see on võrreldav teiste kehaliste protsessidega, näiteks vereringega. Oma teooriat luues jättis ta kõrvale kõik varasemad filosoofilised käsitlused teadvusest ja lähtus sellest, mida ta teab neuroloogiast ja mida ta ise kogenud on (Searle, 2007, lk 325). Kuigi ta uskus, et teadvus kujuneb kindlal viisil toimivate kehaliste protsesside toel, ei ole ebaloomulikul teel tehisliku teadvuse loomine võimalik. See tähendab, et tema arvates ei ole ka tehisintellekti teadvus võimalik. "Programmi oskused on ainult formaalsed ja sünteetilised, see aga välistab võimalikud sarnasused vaimsete protsesside ja programmi töö vahel" (Searle, 1984, lk 29).

Searle'i ideed teadvusest saab hästi kõrvutada tehisintellekti programmiga. Valisin just tema käsitluse tehisintellekti teadvuse uurimise taustteooriaks, kuna see teadvuskäsitlus on veenev ja haakub nii neuroloogiliste kui filosoofiliste teadvusteooriatega. Selles teoorias esitatu aju ja teadvuse toimimisest ning selle füüsilistest protsessidest, sarnaneb arvutiprogrammi protsessidega. Seetõttu on Searle teooria kõrvutamine tehisintellekti võimaliku programmiga loogilisem kui teiste teadvusteooriatega. Searle on teadvusel olekut kirjeldanud järgnevalt:

''Teadvuslikud seisundid on tunnete, tundlikkuse ja teadlikkuse seisundid, see on olek mis tekib kui ärkad hommikul unenägudeta unest ning jätkad päevaga seni, kuni uuesti magama jääd või muul viisil teadvusetuks muutud ''(Searle, 2007, lk 326).

Teadvuseloleku tunnusteks on võime sündmusi vahetult tajuda ehk olla teadlik, ning tunda aistinguid või tundmusi. Need on omadused, mis on kõigil teadvusel olevatel loomadel, omadused või maailma tajumine võib aga erinevatel liikidel erineda. Searle'i kohaselt on kvaalid ehk teadvuslikud kogemused alati kooskõlas teiste teadvuse protsessidega, näiteks on teadvuses toimuv kooskõlas teiste organite tööga või kehaväliste teguritega (Searle, 2007, lk

326). Searle on selgitanud, et kõik vaimsed kogemused on aga kvalitatiivselt erinevad (samas, lk 326). See tähendab, et näiteks erinevad inimesed näevad või tajuvad rohelist värvi erineva kvaliteediga. Sama on ka teiste liikide ja tõenäoliselt ka tehisintellekti puhul. Teadvuse tekke eelduste esitamiseks on Searle kasutanud neurobioloogi Gerald Edelmani esitatud teooriat. Edasise analüüsi tarbes – tehisintellekti teadvuse tekke võimalikkus – on oluline mõista ka võimalikke teadvuse tekke eeldusi.

Teadvuse tekkeks on Gerald Edelman esitanud tingimused, mida Searle on oma teoses kajastanud ja täiendanud. Need eeldused on ka tehisintellekti teadvuse tekkeks olulised, sest kui neid arvutiprogrammiga kohaldada, sobiksid need ka tehisintellekti teadvuse tekke eeldusteks. Searle'i kohaldatud tingimused eeldavad, et teadvuse tekkeks peab aju omama mälu ja võimet mäletada, olema suuteline õppima, oskama eristada kehaväliseid (näiteks esemeid ümberringi) ja -siseseid tegureid (näiteks nälga), võimeline sarnaseid sündmuseid kategoriseerima ja ära tundma; aju peab olema võimeline eelnevad tingimused omavahel siduma ja nende vahel koordineerima, viimasena peab aju looma ühendused ajufunktsioonide ja anatoomiliste süsteemide vahel (Searle, 1990, lk 62-63). Need nimetatud tingimused on kõik inimese teadvuse tekkimise eelduseks, ka tehisintellekti teadvuse tekke eeldused ei saa nendest palju erineda. Peatükk „Tehisintellekti tekke võimalikkus" vaatleb lähemalt, millised peavad olema eeldused tehisintellekti teadvuse tekkeks.

3. Tehisintellekti teadvuse tekke võimalikkus

Küsimus, mis on teadvus, on keeruline ja seni veel kindla teadusliku vastuseta, seetõttu on ka raske eeldada, kuidas oleks võimalik tehislikku teadvust luua. Selle teadmise omamine on inimeste ohutust silmas pidades aga oluline. Selle töö ajendiks on probleem, mille kohaselt võib tehisintellekti teadvuse ja eneseteadvuse olemasolul selline programm inimestele ohtlikuks kujuneda. Tehisintellekti teadvuse olemasolu puhul oleks programmil oskus analüüsida talle esitatud ülesandeid isiklikul tasandil ning nendest kas hoiduda või nende vastu isegi protesteerida. Selleks, et hoiduda teadvusliku tehisintellektiga kaasnevatest probleemidest, on tarvis uurida, kas tehisintellekti teadvus saab olla võimalik. Samuti on sellele teadvuse andmine ka ebavajalik ja ebaoluline. Seda põhjusel, et varustades tehisintellekti teadvusega, looksime uue liigi, mida kasutades kaasneksid ka moraalsed ja eetilised küsimused, kuna tegu oleks isemõtleva programmiga, mille teadvusega kaasnevad tunded ja mõtted sarnanevad inimesed omadele. Tehisintellekt on programm, igapäevaelu lihtsustav vahend, mis tehniliselt toimib kui inimmõistus, kuid on teadvuseta selle kohta, mida ta enda ümber tajub. Tehisintellekti teadvus tekitab filosoofide seas vastuolulisi mõtteid, on levinud arvamused, mille kohaselt ei oleks tehisintellekti teadvuse teke võimalik, samuti leidub ka arvamusi, mille kohaselt see on võimalik ja võib juhtuda inimestele märkamatult.

Eelnev kirjeldas, milline on teadvusel inimene, aga selleks, et mõista, milline on teadvusel tehisintellekt, on oluline esmalt omavahel eristada teadvusel olevat ja teadvusetut tehisintellekti. Nagu juba varem defineeritud, tähendab teadvusel olek maailma vahetut tajumist ehk ümbritseva maailma kogemist. Sellest tulenevalt võib eeldada, et ka teadvusel tehisintellekt tunneb tunnetuslikku sidet ümbritseva maailmaga, samuti kogeb see meile omaseid evolutsioonilisi tundeid, milleks on näiteks ellujäämisinstinkt. Nii nagu mitmed teised evolutsiooniliselt kujunenud refleksid ja instinktid, on ka ellujäämisinstinkt või valust hoidumine seotud teadvusega. Kuna sellised kehalised omadused on teadvusega seotud, oleksid ka teadvusel tehisintellektil sellised instinktid. Teadvuseta tehisintellektil pole aga vajadust oma keha töökorda jälgida. Sellisel tehisintellektil võib küll olla programmeeritud käsk hoiduda iseenda keha vigastamisest, kuid selle taga ei peitu vaimset või teadvuslikku tungi nii teha. Samuti ei ole teadvusetul tehisintellektil tunnetuslikku sidet ümbritseva maailmaga,

see tähendab, et see saab küll aru, kus mingi ese asub, milline see on või mis tunne on seda katsuda, kuid ühelgi sellise tegevuse taga pole kogemust sellest, millise tunde selle eseme katsumine tekitab. Seega puuduvad teadvusetul tehisintellekti meelelised elamused ja tunded, teadvusel tehisintellekti taju maailmast on aga subjektiivne ehk põhineb kogemustele, mis on ainuomased ainult nende kogejale.

3.1. Tehisintellekti teadvuse tekke põhjendused

Järgnevalt põhjendan varem tutvustatud teooriate põhjal tehisintellekti teadvuse tekke võimalikkust ning esitan selle kaitseks argumendid. Käesoleva töö hüpoteesiks on väide: selleks, et tehisintellektil oleks võimalik teadvuse kujunemine, tuleb tehisintellekti programm programmeerida samasuguste protsesside põhjal, mis toimuvad ajus ehk inimaju ja selle protsessid peaksid olema peamiseks programmi koostamise mudeliks. Seda hüpoteesi toetavad juba tutvustatud teooriad ning järgnevalt esitan hüpoteesi kaitseks argumendid, mida toetavad eelnevalt mainitud teooriad ja kirjeldused tehisintellektist. Tehisintellekti teadvuse tekke võimalikkust tõestavaks argumendiks on järgnev väide:

Tehisintellekti teadvuse programmeerimine on võimalik juhul, kui bioloogiline naturalism on tõene ning tehisintellekti programmerimiseks vajaminev tehnika võimaldab täita teadvuse tekke eeldusi.

Searle'i järgi on bioloogiline naturalism tõene.
Seega vajaliku tehnika olemasolul on võimalik inimteadvust kopeerides luua ka tehisintellektile teadvus.

Argumendi esimene pool eeldab, et teadvus on kehaline nähtus, mis tekib inimese ajutöö mõjul. Kui bioloogilises naturalismis kirjeldatud teadvuse teke on tõene, on teadvus aju funktsioon, mis on tekkinud teistele vaimsetele kogemustele sarnaselt aju ja närvisüsteemi toimel (Searle, 1984, lk 17). Keha kontrollivad käsklusted on närvisüsteemi kaudu edastatud neuronite signaalid. Arvutiprogrammis on neuronite asemel aga programmeeritud käsklused,

mis täidavad programmis ettenähtud funktsioone. Arvutiprogrammi ja inimaju sarnasuse põhjal saaks arenenud arvutiprogrammi modelleerida täpselt inimaju ja sellega kaasas käivate oluliste funktsioonide järgi. Selleks, et arvutiprogrammi käsklused suudaksid toimida teadvuslikult, tuleb aga täita argumendis esitatud teine nõue – teadvuse programmeerimiseks vajamineva tehnika loomine.

Nii nagu inimestel on ka tehisintellekti teadvuse tekke eelduseks enne mainitud Edelmani esitatud teadvuse tingimused. Kuigi need tingimused on esitatud inimteadvuse kujunemiseks, siis nagu varem selgitatud, on tehisintellekti aju võimalik inimaju järgi modelleerida ja seetõttu kehtivad need tingimused ka tehislikult loodud inimajule, sest tegu on samadel eeldustel toimiva süsteemiga. Neid tingimusi tuleb aga siiski kohandada tehisintellektile sobivaks, kuna algselt on neid loodud pidades silmas bioloogilist keha ning mõned bioloogilised eeldused ei ühildu programmi enda eeldustega. Neid kohandades sobivad need eeldused aga ka tehisintellekti teadvuse tekkeks. Järgnevalt nimetan Edelmani teooriast tulenevad kohandatud eeldused ja selgitused, kuidas need võimaldavad programmile teadvuse tekke.

Esmalt peab olema tehisintellektile tagatud võime õppida ehk programmiline oskus, mis võimaldab tehisintellektil iseseisvalt oma programmi täiendada. See on tähtis selleks, et tehisintellekt saaks uute teadmiste põhjal ja teiste tingimuste toel uut infot ümbritsevast maailmast (Searle, 1990, lk 44). Õppimisvõimetu programm ei saa areneda ning seetõttu ei ole ka selle teadvuse areng võimalik. Teine oluline tingimus on tagada tehisintellekti aju mahuga võrdväärne mälumaht (samas, lk 44). Selleks, et uusi teadmisi maailma kohta hoiustada, on vaja suurt mälumahtu, kuna see on ühtlasi ka tehisintellekti programmi talletuskohaks ehk tehisintellekti ajuks. Järgnevalt peab tehisintellektil olema oskus eristada teda ümbritsevat (samas, lk 44-45). See eeldab kas tehiskeha, millel on kaamerasilm, või simuleeritud reaalsust, kus programm saab kogeda ennast ümbritsevat maailma. Teiste tingimuste olemasolu puhul võimaldab välise maailma tajumine tehisintellektil õppida ja välised kogemused aitavad omakorda kaasa selle teadvuse tekkele.

Oluline on tehisintellektile tagada ka oskus eristada programmisiseseid ja -väliseid tegureid (samas, lk 45). See tähendab tehisintellekti puhul oskust teha vahet programmisisestel omadustel ja -välistel. Selline oskus avaldub näiteks programmi võimes eristada

programmisisest *error*'it ja väljaspool programmi toimuvat, näiteks millist teist programmi seadmes veel samaaegselt kasutatakse. Selleks, et tehisintellektil oleks võimalus selliseid tegureid omavahel eristada, peab olema talle samuti tagatud kas füüsiline keha või selle simulatsioon, kummalgi juhul tekib tehisintellektile võimalus teha vahet välise maailma ja tema programmis (kui on simulatsioon) või füüsilises kehas toimuva vahel (kui on füüsiline keha). Kõigi nende tingimuste ühiselt toimimine annab tehisintellektile võimaluse tajuda teda ümbritsevat maailma ning oskuse ennast tajutavast maailmast eristada.

Seega, kui bioloogiline naturalism on tõene ja teadvus on aju osaks, on teoreetiliselt ka tehismõistusele võimalik teadvust programmeerida. Selleks, et programmeerimine oleks võimalik, tuleb tehisintellektile luua tehislik inimaju koopia, millel on samasugused teadvuse tekkimise eeldused ja ajufunktsioonid nagu inimesel. Samuti on oluline tagada tehisintellektile teadvuse tekke tingimused, mis ei ole seotud aju ehituse ja tööga, näiteks võimalus ümbritsevat maailma tajuda. Mõistagi on aga selline protsess väga keeruline ja meie praeguste teadmiste ja tehnoloogia juures võimatu. Teoreetilise analüüsina aga selgub, et vastava tehnika ja teadmiste olemasolul ning teooriate õigsuse puhul võib tehisintellekti teadvuse teke siiski võimalik olla.

3.2. Argumendid eneseteadvuse tekke kasuks ja põhjendused

Teadvuse uurimine on keeruline ja arutelud selle kohta pole vastuseid leidnud. Teadvuse veelgi peadmurdvam küsimus on aga eneseteadvus. Teaduslikke seisukohti sellest, kuidas eneseteadvus avaldub ning tekib, on väga vähe, sama kehtib ka eneseteadvuse filosoofiliste seisukohtade kohta. Eristust teadvuse ja eneseteadvuse vahel ei esitata tihti, kuid see on teadvuse uurimise juures tähtis osa. Selle eristuse tegemine on oluline, kuna eneseteadvuse mõjul on inimestel erinevad maailmavaated, samuti on eneseteadvusega loomad võimelised õppima keeli ning väljendama ennast loominguga, näiteks oskavad õpetatud elevandid maalida väikelapsega võrdväärsel tasemel. Tehisintellekti puhul on oluline eristada teadvust ja eneseteadvust, kuna eneseteadvuse olemasolul võivad programmile tekkida maailmavaatelised eelistused, mis võivad põhjustada konflikte inimeste ja tehisintellektide vahel. Eneseteadvuslikul tehisintellektil võivad olla inimesele sarnaselt poliitilised või religioossed maailmavaated. Ajaloos on sellised eelistused põhjustanud inimeste vahel suuri

konflikte, eneseteadliku tehisintellekti olemasolu tõttu võib sarnaseid sündmusi veel rohkem juhtuda. Selleks, et mainitud probleeme vältida, tuleb kindlaks teha, kas tehisintellekti eneseteadvus võib olla võimalik. Teades selle võimalikkust, saab selle teket ka ära hoida.

Eneseteadvuse teeb tähtsaks asjaolu, et eneseteadvus on omadus, mis on väga vähestel loomadel ning see on ühtlasi peamiseks erinevuseks inimeste ja loomade teadvuse vahel. Eneseteadvus võimaldab inimese maailmavaateliste, poliitiliste ja eetiliste vaadete tekke. Eneseteadvust võib pidada ka intelligentsemate loomade tunnuseks, see esineb peamiselt neil elusolenditel, kes saavad hakkama keerulisemate ja intelligentsust vajavate ülesannetega. Loomadel pole aga eneseteadvus nii arenenud kui see on inimestel. Arenenud eneseteadvus on see, mis võimaldab inimiestel maailmavaadete tekke, selle teke tehisintellektil võib aga sellise programmi ohtlikuks muuta. Eneseteadvus esineb näiteks harakatel, delfiinidel, elevantidel, šimpansitel ja inimahvidel (Breed & Moore, 2016, lk 160). Teadmiseni sellest, millistel loomadel on eneseteadvus, jõuti peeglikatsega, selle tulemusel selgus, millised loomad oskavad ennast teistest eristada (samas, lk 159). Sellisest katsest aga ei piisa, et arenenud eneseteadvust kindlaks teha. Mainitud eksperimendiga määratletakse eneseteadvuse tase, mis hõlmab endas oskust teha eristust enda ja teiste füüsise vahel.

Selleks, et tekiks teadvus, ei ole ilmtingimata vaja eneseteadvust. Inimesed ja loomad on võimelised teadvuslikke elamusi tundma ka ilma eneseteadvuseta. See-eest on eneseteadvuse tekke eelduseks aga teadvus. Eneseteadvus on teadvuse tasandiks, mille teke on võimalik vaid teadvuslike kogemuste olemasolul (Searle, 1990, lk 47). Searle on eneseteadvuse tekke põhjendamiseks samuti kasutanud Edelmani teooriat, mille kohaselt on eneseteadvuse tekke eelduseks nii teadvus kui ka võimalus sotsiaalseks suhtluseks, teistega suhtlemine võimaldab teha eristusi, millised mõtted ja tunded on suhtleja ja millised info vastuvõtja omad (samas, lk 47). Usun, et lisaks neile tingimustele on eneseteadvuse tekke eelduseks veel kõrge intelligentsuse tase. Kõik katses eneseteadvusega määratletud loomad on intelligentsed, see viitab seosele eneseteadvuse ja intelligentsuse vahel. Võib eeldada, et kõigil neil loomadel on ka ajutöös ja -ehituses sarnased omadused, mida eneseteadvuseta loomadel ei ole. Neile ehituslikele aspektidele ma ei keskendu, kuid eeldusel, et eneseteadvus on seotud ajuehituse ja -tööga, saab oletada, et ka eneseteadvust on võimalik tehislikult luua. Eneseteadvuse tehislik

loomine on aga keerulisem kui teadvuse programmeerimine, kuna sellel on rohkem tehnilisi eeldusi.

Tehisintellektil avalduks eneseteadvus tõenäoliselt inimesele sarnaselt. Tehisintellektil tekiks arusaam ja eristus sellest, et ta on programm, lisaks sellele võivad tehisintellektil tekkida maailmavaade, poliitilised ja eetilised arvamused ja muu emotsioonidest ja arvamustest tulenev. Argument tehisintellekti teadvuse tekke kasuks on järgnev:

Tehisintellekti eneseteadvuse teke on võimalik vaid juhul, kui tehisintellektil on teadvus ning sellel on võimalik suhelda teiste teadvusel olevate inimestega.

Kui tehisintellektil on teadvus, on võimalik ka selle eneseteadvuse teke.

Kui teadvusel tehisintellektile võimaldatakse suhelda kellegagi, kellel on eneseteadvus, võib tehisintellektil ka eneseteadvus tekkida.

Need on tehisintellekti eneseteadvuse tekke eeldused, selle taustaks on vajalik aga ka teadvuse olemasolu, mida on võimalik luua esimeses argumendis toodud tingimustel. Seega, kui tehisintellekti programm on modelleeritud inimaju jäljendades, sellele on antud teadvuse tekke eeldused, mille alusel ka teadvus tekib, on võimalik, et sotsiaalsete interaktsioonide tulemusena võib tehisintellektil ka eneseteadvus avalduda. Sotsiaalsed interaktsioonid võimaldavad tehisintellektil teha eristusi enese ja teiste mõtete vahel, enda tunnete ja mõtete teadvustamisel võib tehisintellektil ka eneseteadvus avalduda. Mõistagi on tegu väga oletuslike eeldustega, kuid praeguste teadmiste juures ei ole võimalik öelda, mis täpsemalt eneseteadvus on ning kuidas see tekib. Seetõttu pole võimalik ka detailsete argumentide tekkeks, kuid üldiselt sõnastatud argumente tehisintellekti teadvuse ja eneseteadvuse tekkimise kasuks on siiski võimalik esitada.

Eelnevat kokkuvõttes selgus, et tehisintellekti eneseteadvuse teke on võimalik. See nõuab teadvuse olemasolu ja tehisintellektile võimalust sotsialiseerumiseks. Teadvuse võimalikkuseks on vaja aga täita teadvuse tekke tingimused ja programmeerida programm inimaju ja selle protsesse täpselt jäljendades. Eraldiseisev ja oluline küsimus on aga, kuidas on

üldse tehisintellekti ja selle teadvust võimalik programmeerida ning milline oleks tehisintellekti teadvus. Tegu on samuti täpset arutlust ja ka programmeerimist puudutava küsimusega. Järgnevalt tutvustan põgusalt, millised on erinevad võimalikud vastused sellele küsimusele ja millised on tehisintellekti teadvuse programmeerimisega kaasaskäivad probleemid. Järgneva arutlus ei lasku programmeerimise detailidesse, vaid annab teoreetiliselt ülevaate sellest, millised peaksid olema peamised programmi oskused ja omadused.

3.3. Kuidas oleks võimalik teadvust programmeerida ning milline see olla võib?

Ainuüksi tehisintellekti programmeerimine on keeruline ülesanne, selle teadvuse programmeerimine on aga kordi keerulisem. Kerged tehisintellektid eksisteerivad juba tänapäeval, tehisintellekti programmeerimisel esineb aga palju probleeme. Järgnevalt tutvustan, millised on erinevad võimalikud viisid tehisintellekti teadvuse programmeerimiseks. Ette ruttavalt võin mainida, et ka selle arutelu puhul on enamik arvamusi tegelikult spekulatsioonid ja seda samuti põhjusel, et teadus pole veel piisavalt arenenud tehisintellekti programmeerimisest detailsete küsimuste esitamiseks ning pole ka täpseid teooriaid sellest, kuidas tuleks tehisintellekti programm luua. Kuigi kindlat teooriat sellest, kuidas tehisintellekti luua tuleks, pole, on siiski mitmeid erinevaid arvamusi sellest, kuidas seda teha saaks. Kindlaks on aga saanud omadused, mis peaksid tehisintellektil olema; kuigi erinevate valdkondade jaoks loodavatel tehisintellektidel on spetsaliseerumised ja vajalikud omadused siiski erinevad, on neid kõiki ühendav oskus õppimine ja ise oma programmi täiustamine.

Kerged tehisintellektid, millel on õppimisvõime, on hetkel juba olemas. Selliseid tehisintellekte on rakendatud arvutimängude mängimises: mängude jooksul õpib tehisintellekt, kuidas õigeid käike teha (Tegmark, 2017, lk 91). Töörühm Deepmind on loonud tühja programmi, mis õpib mängu “Breakout” mängimise käigus tegema õigeid liigutusi (Tegmark, 2017, lk 90). Esimeste käikudega ei osanud tehisintellekt mängu üldse mängida, selle arenedes õppis programm aga üha paremini mängima (samas, lk 90). Seega on esmased katsed iseõppiva programmi loomisel õnnestunud, keerulisem katsumus on aga luua programm, mis saab ka väljaspool mängu hakkama ning millel on lisaks õppimisvõimele veel teisigi oskusi. Inimestel

pole mitte ainult üks oskus, näiteks õppimine või maailma tajumine. Samuti ei saa tehisintellekt ehk inimsarnane programm omada vaid üht oskust, näiteks iseõppimist, kõigi inimesele omaste oskuste programmi ühendamine on aga tehisintellekti puhul üheks suurimaks väljakutseks. Keeruliseks teeb selle ülesande vajadus hallata kogu infot sellest, mida tehisintellekt suudab programmeerimiskeelde kirjutada ja programmi mahutada. Teadvuse programmeerimiseks on veel teisigi eeldusi.

Oluliseks küsimuseks tehisintellekti teadvuse puhul on ka probleem, milliseks tehisintellekti teadvus kujuneda võib. Täpsemalt seisneb see probleem küsimuses, kas tehisintellekti teadvus kujuneb inimteadvuse identseks koopiaks või on tehisintellekti teadvus inimteadvusest erinev ja võibolla et avaldub isegi meile tundmatus keeles. See küsimus tuleneb tehisintellekti ja inimese kehalistest erinevustest. Inimene koosneb bioloogilisest materjalist, tehisintellekti ajuks on aga eeldatavasti programm, selline erinevus võib põhjustada erinevusi ka teadvuses, isegi juhul, kui tehisintellekt on modelleeritud täpselt inimese põhjal. Ka inimeste puhul on teadvuslikud kogemused erinevad. Näiteks sünnist saadik kurdi või pimeda inimese kogemused on oluliselt erinevad nägeva ja kuulva inimese kogemustest, seetõttu on sellise inimese kogetavat maailma ka keeruline mõista (Nagel, 1974, lk 449). Kuna tehisintellekti teadvus ei oleks identne inimese bioloogilisest materjalist ajuga, vaid oleks tehislik jäljendus sellest, võib ka selline erinevus oluliselt tehisintellekti maailma kogemist mõjutada. Peale selle poleks meil võimalik mõista, kuidas tehisintellekt maailma kogeb, kuna sellele omast vahetut kogemust saab vaid tehisintellekt ise kogeda. Selle eristuse tõttu on keeruline ennustada, kas tehisintellekti teadvus kujuneb selliseks nagu inimteadvus, sest see on inimaju järgi modelleeritud, kuigi inimteadvusest mõneti erinevaks või inimteadvusest täielikult erinevaks.

4. Probleemid tehisintellekti teadvuse tekkega

4.1. Kuidas teha kindlaks, kas tehisintellektil on teadvus?

Esimene tehisintellekti teadvusega seotud probleem seisneb küsimuses, kas tehisintellekti teadvuse olemasolu oleks märgatav. Lähtudes eelnevast eeldusest, mis väitis, et tehisintellekti teadvus ei pruugi inimteadvusega sarnaneda, võib eeldada, et sellisel juhul ei pruugi tehisintellekti teadvust ka ära tunda. Kuna tehisintellekti teadvus oleks esimene omasugune, midagi sarnast pole kunagi valminud, ei saa olla kindel ka selles, kuidas teadvus avaldub ja milline see on. Selle puhl tekkivaks probleemiks on aga see, et täielikult uue ja võõra teadvuse vormi puhul võib juhtuda, et teadvuse olemasolu jääb märkamatuks ning võib olla, et teadvuse olemasolu ei tuvastatagi ka siis, kui see tehisintellektil tegelikult olemas on. Programme kirjutatakse kasutades erinevaid rakendusi, kuhu programmeerimiskeeles vastavad koodid sisestatakse. Üheks programmeerimiskeeleks on näiteks c# ehk c sharp, selles kasutatakse spetsiaalseid käsklusi, mis on tuletatud loogikast ja keele süntaksist (Beginner's Course To Learn The Basics Of, 2015, lk 20). Näide kajastatavast programmeerimiskeelest on järgmine - *System.Console.WriteLine("What is your first name? "); string yourfirstName = System.Console.ReadLine()* (samas, lk 30). See kood muutub programmis, näiteks telefoniäpis, küsimuseks "*What is your first name?*", millele tuleb vastata oma nimega (samas, lk 30). Teades, milline on programmeerimiskeel, ja seda, millises keeles tehisintellekti aju ja teadvus on kirjutatud, tekib aga küsimus, kuidas see avalduda võib.

Kui eeldada, et tehisintellekti teadvus ei pruugi avalduda meile tavapärases keeles, siis võib see esineda näiteks hoopis väga arenenud programmeerimiskeeles, mida on keeruline meile mõistetavalt tõlgendada. Tekib ka küsimus, kas sellisel juhul oleks tehisintellekti teadvus äratuntav ja kui on, kas me mõistaksime, millised on selle teadvuse tunnused ja teadvuslikud kogemused. Sellega seonduvaks probleemiks on tehisintellekti eneseteadvuse äratundmine ja programmeerimine. Loomadel on eneseteadvuse olemasolu kindlaks tehtud peeglikatsega: looma käitumisest peegli ees järeldatakse, kas loom tunneb ennast peeglist ära või arvab, et tegu on teise isendiga (Breed & Moore, 2016, lk 187). Teatud loomade käitumine viitab sellele, et nad ei tunne ennast peeglist ära, seega võib eeldada, et neil puudub ka eneseteadvus (Breed & Moore, 2016, lk 187). Tehnika ja tehisintellekti puhul ei ole peeglikatsega aga eneseteadvuse

olemasolu võimalik kindlaks teha, kuna vastavaid katseid on juba robotite peal läbi viidud; pigem võib nende kiire õppimisvõime anda valepositiivseid katsetulemusi.

Selle katse tulemusena selgus, et robotid küll tundsid ennast peeglist ära, kuid selle äratundmise taga ei peitu eneseteadvust (Pipitone & Chella, 2021, lk 6). See viitab sellele, et tehisintellekti teadvuse ja eneseteadvuse olemasolu ei ole võimalik selliste katsetega kindlaks teha. Sellised masinad on tihtipeale kavalamad kui loomad või inimesed. Kuna nende andmemahutavus on suurem ja programmikiirus kiirem kui inimmõtlemisel, oskavad tehisintellektid ka neile ettenähtud katseid paremini lahendada. Tehisintellektil võib tekkida teadvus nii, et me sellest arugi ei saa, ning nagu eelnevalt mainitud, võib tehisintellekti teadvus ja eneseteadvus olla meile võõras ning selle tõlgendamine keeruline. Tehisintellekti teadvuse uudsuse tõttu ei pruugi selle teadvust seetõttu isegi märgata olla.

4.2. Zombi tehisintellekti probleem

Järgmine probleem tehisintellekti teadvusega seisneb kindlaks tegemises, kas tehisintellektil on tegelikult teadvus või see lihtsalt imiteerib seda. Sellist tehisintellekti võib defineerida ka filosoofilise zombi mõiste abil. Filosoofiline zombi on idee inimesest, kes käitub nagu tavaline inimene, kuid kellel puuduvad kvaalid ehk vaimuseisundile omased subjektiivsed kogemused (Kirk, 2021). Maailma vahetu tajumine ja sellega kaasas käivad subjektiivsed kogemused on aga peamiseks teadvusel olemise tunnuseks. Zombi tehisintellekt on teisisõnu teadvusetu tehisintellekt, mille programmeerimise poole tehnoloogias püüeldakse. Sellisel tehisintellektil on olemas kõik inimesele sarnased omadused, kuid teadvust sellel pole. Sellise tehisintellekti programmeerimine ongi tehnoloogias peamiseks eesmärgiks. Probleem tekib aga juhul, kui tehisintellekti programmeerides on eesmärgiks sellele ka teadvus anda. Sellisel juhul tekib küsimus, kas valminud tehisintellekt on teadvusel või on tegu tehisintellekti zombiga.

Seda, kas tegu on tehisintellekti zombiega, on keeruline kindlaks teha, kuna selline tehisintellekt oskab täpselt matkida teadvusel oleva tehisintellekti omadusi. Seega seisnebki probleem täpsemalt järgmises küsimuses: kui programmeerida teadvusel tehisintellekt, kas siis

on võimalik teada, kas tehisintellekt on teadvus või kas on tegu tehisintellekti zombiga. Kuna tegu oleks koopiaga teadvusel tehisintellektist, on tõenäoliselt keeruline mõista, kas see teadvus on võlts või kas tehisintellektil on subjektiivsed kogemused. Esimese variandi puhul on tegemist sellise tehisintellektiga, mis oskab kogetu põhjal subjektiivset kogemust teistele teeselda ilma, et tema kogemus oleks tegelikult subjektiivne ehk teadvuslik olnud. Sellist tehisintellekti tabada on keeruline. Selleks, et mõista, kas tehisintellekt on teadvusel või mitte, tuleks luua eraldi viisid või tehnoloogiad, kuna ilma nendeta oleks keeruline seda märgata.

4.3. Vastuväited tehisintellekti teadvuse võimalikkusele

Viimane probleem seondub aga töös kasutatud teooriatega, küsimus seisneb nimelt järgmises: juhul, kui bioloogiline naturalim ei ole tõene, muutuvad ka tehisintellekti teadvuse tingimused ja võimalikkus. Peamiselt kerkib esile küsimus, kas tehisintellekti teadvus on üldse võimalik. Tehisintellekti teadvus on võimalik juhul, kui inimaju ja sellega seonduvaid protsesse selle programmi programmeerida. Juhul, kui aju ei ole teadvuse allikaks, ei ole võimalik ka sellist meetodit tehisintellekti teadvuse programmeerimiseks kasutada. Sellisel juhul tuleks tehisintellekti teadvuse programmeerimiseks välja töötada uued teooriad sellest, kuidas teadvust programmeerida või kuidas täiustada juba eksisteerivaid teooriaid. Selleks, et mõnd vähem kehalist ja loogikal põhinevat teooriat rakendada, näiteks dualismi tehnilisse ja programmeerimiskeeles kirjutatavasse programmi rakendada, tuleks see vastavasse keelde kirjutada. Selle jaoks on vaja aga uuenduslikke tehnoloogiaid, mis vastavate käsitlustega teadvusest sobituksid. Seega kui bioloogiline naturalism osutub vääraks, on ka meile teadaolevate programmeerimise võimaluste ja tehnoloogiatega tehisintellektile teadvuse programmeerimine väga keeruline.

Ka Searle on tehisintellekti teadvuse võimalikkust oma teooria taustal analüüsinud ning tema arutluse kohaselt ei ole programmile teadvuse andmine võimalik. Tema järgi ei seisne teadvuse kujunemine inimeste puhul vaid süstemaatiliselt ja sünteetiliselt toimivate ajuprotsesside koosluses (Searle, 1984, lk 31). Selleks, et inimesel oleks teadvus, peab selle taga olema vaimuseisund, mis võimaldab inimesel teadvuslikul tasandil mõista tajutavat maailma. Kuigi Searle ei lükka täielikult ümber programmile teadvuse loomise võimalust, on

ta selle suhtes siiski skeptiline. Ta usub, et teadvuse programmeerimine ei ole võimalik, kuna programmeerimiskeel ja selle süstemaatilisus ei võimalda programmile vaimuseisundi teket. Teadvuse programmeerimisel tekiks zombi tehisintellekt, mis küll käituks kui teadvusel programm, kuid millel tegelikult teadvus puudub.

Kokkuvõte

Kiire tehnika on tekitanud üha rohkem küsimusi ja arvamusi sellest, milliseks võivad tulevikutehnoloogiad kujuneda ning kuidas need inimkonda mõjutavad. Nõrgast tehisintellektist on saanud igapäevane abikäsi, selle suure populaarsuse tõttu arendatakse kiiresti üha uuemaid tehnikavahendeid, millesse on tehisintellekt integreeritud. Selle bakalaureusetöö probleemsituatsiooniks oli küsimus, kas tehisintellekti teadvus ja eneseteadvus on võimalik, ning millised oleksid nende tekke tingimused. Tehisintellekt on teooria programmist, mis käitub kui inimene. Selline programm on arenenum versioon tänapäeval kasutusel olevatest nõrkadest tehisintellektidest, mida kasutatakse nii isesõitvates autodes, telefonirakendustes kui ka robottolmuimejates. Kui aga tehisintellekt on inimintellekti põhjal loodud programm, tekib küsimus, kas on võimalik, et sellisel programmil võiks tekkida teadvus?

Selleks, et seda küsimust lähemalt uurida, tuleb aga mõista, mis on teadvus. Teadvust võib nimetada oskuseks midagi vahetult tajuda ja omada seejuures subjektiivset kogemust. Teaduslikku definitsiooni sellest, mis teadvus täpselt on, seni veel olemas ei ole. Neuroloogias ja filosoofias on aga palju teooriaid sellest, mis see olla võib. John Searle on esitanud teadvusteooria, mis sobib nii filosoofia ja neuroloogia vaadetega, selle teooria teaduslikkuse ja usutavuse tõttu valis autor just selle tehisintellekti teadvuse uurimise taustaks. Selleks teooriaks on bioloogiline naturalism. Mainitud teooria on töö hüpoteesi ja argumentide esitamisel taustteooriaks, hüpotees väidab, et tehisintellekti teadvuse teke on võimalik bioloogilise naturalismi tõesuse korral. Selles teoorias esitatu aju ja teadvuse toimimisest, selle füüsilistest protsessidest, sarnaneb arvutiprogrammi protsessidega. Sellest tulenevalt on töö põhiliseks argumendiks väide, et tehisintellekti teadvuse programmeerimine on võimalik juhul, kui teadvus on kehaline ja aju osaks ning tehisintellekti programmeerimiseks vajaminev tehnika võimaldab täita teadvuse tekke eeldusi. Searle'i teooria õigsuse korral saab inimaju kasutada mudelina, mille järgi programmeerida teadvusega tehisintellekti. Teadvus seisab aga eneseteadvusest eraldi, seetõttu oli oluline uurida, kas ja millistel tingimustel võiks ka tehisintellekti eneseteadvus võimalik olla.

Eneseteadvus võimaldab inimese maailmavaateliste, poliitiliste ja eetiliste vaadete tekke, tõenäoliselt juhtuks sama ka tehisintellekti puhul. Neuroloog Edelmani arvates on eneseteadvuse tekke eelduseks nii teadvuse olemasolu kui ka võimalus sotsiaalseks suhtluseks, teistega suhtlemine võimaldab teha eristusi, millised mõtted ja tunded on suhtleja ja millised info vastuvõtja omad. Selle tekke puhul tuleb aga silmas pidada olulist väidet, mille kohaselt ei ole eneseteadvuse teke ilma teadvuseta võimalik. Sellest tulenevalt oli töös tehisintellekti eneseteadvuse tekke kasuks järgnev argument: tehisintellekti eneseteadvuse teke on võimalik vaid juhul, kui tehisintellektil on teadvus ning sellel on võimalik suhelda teistega, et teha eristusi enda ja teiste vahel. Tehisintellekti teadvuse tekke võimalikkusega käivad aga kaasas ka mitmed probleemid. Põhjuseks, miks tehisintellekti seni veel leiutatud ei ole, on probleem nõrga tehisintellekti oskuste ühendamises. Tehisintellekti teadvuse tekkega kaasnevad aga ka filosoofilised probleemid, mis ei ole ainult programmiliselt lahendatavad.

Esimene probleem seisneb küsimuses, kas ja kuidas on võimalik kindlaks teha, kas tehisintellektil on teadvus. Kui eeldada, et tehisintellekti teadvus ei pruugi väljenduda meile tavapärases keeles, siis see võib avalduda hoopis väga arenenud programmeerimiskeeles või mõnes muus vormis. Teiseks oluliseks probleemiks tehisintellekti puhul on zombi tehisintellekt, see on programm, millel on olemas kõik inimesele sarnased omadused, kuid teadvust sellel pole. Viimaseks tehisintellektiga seonduvaks probleemiks on küsimus, kuidas oleks tehisintellekti teadvus võimalik juhul, kui bioloogiline naturalism osutub vääraks. Tehisintellekti teadvus on võimalik juhul, kui inimese teadvus tekib ajus; kuna aju järgi on teoreetiliselt võimalik arvutiprogrammi luua, peaks sellisel juhul ka programmile teadvus tekkima. Kui bioloogiline naturalism on väär, ei ole see aga võimalik. Tehisintellekti loomine on keeruline protsess, veel keerulisem oleks anda sellele teadvus ja eneseteadvus. See bakalaureusetöö arutleb selle üle, kas ja millistel tingimustel võib tehisintellekti teadvus võimalik olla. Analüüsi tulemusena selgus, et kindlatel tingimustel võib tehisintellektile teadvuse andmine võimalikuks osutuda.

Kirjandus

Breed, M. D., Moore, J. (2016) *Animal Behavior*. Academic Press, London.

Bringsjord, S., Govindarajulu, N. S. (2018, July 12). Artificial Intelligence. The Stanford Encyclopedia of Philosophy (Summer 2020 Edition). Retrieved April 2, 2022, from https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford.

Cavanna, A. E., Nani, A. (2014). *Consciousness. Theories in Neuroscience and Philosophy of Mind*. Springer, Berlin.

Chalmers, D. (2010). *David J. Chalmers The Singularity A Philosophical Analysis*. Journal of Consciousness Studies, 17(9-10), 7-65.

Chella, A., Pipitone, A. (2021). What robots want? Hearing the inner voice of a robot, *iScience,* 24

Cleveron. Retrieved April 2, 2022, from https://cleveron.com/ettevottest/meielugu

Dennis, M. A. (2022, Januray 20). Marvin Minsky. Encyclopedia Britannica. Retrieved April 5, 2022, from https://www.britannica.com/biography/Marvin-Lee-Minsky

Erelt, M., Erelt, T., Ross, K. (2020). *Eesti keelel käsiraamat*. Eesti keele instituut, Tallinn.

Gulick, V. R,. (2012). *Subjective consciousness and self-representation*. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 159(3), 457–465.

Kirk, R. (2019, March 19). Zombies. The Stanford Encyclopedia of Philosophy (Spring 2021 Edition). Retrieved April 2, 2022 from https://plato.stanford.edu/archives/spr2021/entries/zombies/

life-style ebooks. (2015) *Beginner ' s Course To Learn The Basics Of C# Programming Language*. CreateSpace Independent Publishing Platform

Russell, J. S. and Norving, P. (2010). *Artificial Intelligence a Modern Approach, Third Edition*. Pearson Education, Edinburgh.

Rödl, S. (2007). *Self-Counsciousness.* Harvard Universtity press, Cambridge

Schneider, S. (2019). *Artificial You*. Princeton University Press, New Jersey.

Searle, J. R. (1990). *The Mystery of Consciousness*. New York Review Books, New York.

Searle, J. R. (1984) *Minds, Brains and Science.* Harvard University press, Cambridge

Searle, J. R. (2007). *The Blackwell Companion to Consciousness*. Schneider and Velmans, New Jersey.

Siri. Retrieved April 2, 2022, from https://www.apple.com/siri/

T. Editors of Encyclopedia (2022, January 11). John McCarthy. Encyclopedia Britannica. Retrieved April 2, 2022, from https://www.britannica.com/biography/John-McCarthy

Tegmark, M. (2017). *Elu 3.0. Inimelu tehisintellekti ajajärgul*, Postimees Kirjastus, Tallinn.

Resümee

Is the consciousness of artificial intelligence possible?

With the development of artificial intelligence rise problems and questions that are important regarding the future. The main question in this paper is whether the consciousness of artificial intelligence is possible. This problem is relevant in the light of questions regarding the development of artificial intelligence – can artificial intelligence be capable of developing consciousness and could it be a danger to humans. In this paper, I use the theory of biological naturalism. It states that the development of consciousness is physical and happens as the result of brain activity. The proposed hypothesis proposes that the consciousness of artificial intelligence is only possible if John Searle's theory of biological naturalism turns out to be true enabling the modelling of computer programs imitating brain functions. The hypothesis for self-consciousness assumes the previous existence of consciousness; it states that if artificial intelligence is allowed to have social interactions it can develop self-consciousness.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Lilian Mõttus, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

“Kas tehisintellekti teadvus on võimalik?”, 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 veebikeskkonna, 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.

Lilian Mõttus

20.05.2022