MITÄ EROA ON IHMISELLÄ JA TIETOKONEELLA? – LAAJENNETTU TULKINTA HUBERT L. DREYFUSIN ARGUMENTISTA

Maija Raasakka
Kandidaatintutkielma
Filosofia
Yhteiskuntatieteiden ja filosofian laitos
Jyväskylän yliopisto
Syksy 2023

JYVÄSKYLÄN YLIOPISTO

Tiedekunta Humanistis-yhteiskuntatieteellinen	Laitos Yhteiskuntatieteiden ja filosofian laitos
Tekijä Maija Raasakka	
Työn nimi Mitä eroa on ihmisellä ja tietokoneella? – Laajennettu tulkinta Hubert L. Dreyfusin argumentista	
Oppiaine filosofia	Työn laji kandidaatintutkielma
Aika 2023	Sivumäärä 31

Tiivistelmä

Tutkielman tarkoituksena on tarkastella ihmisen ja tietokoneen yhtäläisyyttä ja eroa jäsentävää argumentaatiota, sekä erityisesti informaation roolia tässä argumentaatiossa. Tarkastelu perustuu Hubert. L. Dreyfusin esittämään kriittiseen jäsennykseen siitä, mitä viime kädessä ontologisia ja epistemologisia oletuksia on tehtävä, mikäli oletetaan ihmisen kaltaisen tekoälyn olevan mahdollinen. Pidän Dreyfusin näkemystä varteenotettavana ja otan sen lähtökohdakseni, mutta täydennän sitä keskittyen erityisesti informaation käsitteeseen. Tutkimuskysymykseni on, millä tavalla käsitys informaation luonteesta vaikuttaa käsitykseen ihmisen ja tietokoneen välisestä samankaltaisuudesta.

Muotoilen argumentin rungon, jossa on ensin käsitettävä ihmiset ja koneet samankaltaisiksi ja tämän jälkeen määriteltävä niin informaatio kuin ihminenkin siten, että keinotekoinen representaatio tekoälyn avulla on mahdollista. Johtopäätöksenäni esitän seuraavan ajatusrakennelman: Ensiksi on oletettava informaation luonne materiasta irrallisena todellisuuden tasona, mikä mahdollistaa analogian tekemisen kaikkien olioiden välille. Tämän jälkeen on oletettava, että sekä ihmiset ja tietokoneet prosessoivat informaatiota formaalien sääntöjen mukaisesti diskreettinä datana, joka sellaisenaan representoi todellisuutta. Tämä vastaa sitä, mitä Dreyfus esitti tekoälyn taustalla oleviksi epistemologisiksi ja ontologisiksi olettamiksi. Jotta voidaan vielä ajatella ihmisen kaltaisen tekoälyn olevan mahdollinen, on lisäksi oletettava äly universaaliksi ja kontekstista irralliseksi tarkkarajaiseksi kokonaisuudeksi.

Keskeistä tässä kokonaisuudessa on sen kriittinen potentiaali. Dreyfus on osoittanut, miksi mainitut ihmistä koskevat oletukset eivät pidä paikkaansa. Osoitan lisäksi, kuinka mainitut oletukset informaation luonteesta sekä älykkyydestä voivat olla ongelmallisia. Väitän siten, ettei ihmistä ja tietokonetta voi mielekkäällä tavalla rinnastaa eikä ihmisen kaltaista tekoälyä rakentaa.

Asiasanat: analogiat, Hubert L. Dreyfus, ihminen-konejärjestelmät, ihmisyys, informaatio, tekoäly, tietokoneet

Säilytyspaikka: Jyväskylän yliopisto

Muita tietoja

SISÄLLYS

1 JOHDANTO

Ihmisen mieli prosessoi informaatiota. Tietokone prosessoi informaatiota. Siinä missä tietokone on metalliosista tehty kone, on ihminen orgaanisesta materiaalista tehty kone, joka on lopulta pohjimmiltaan samanlainen tietokoneen kanssa. Nämä ovat ajatuksia, joita esitettiin tietokoneiden kehityksen alkuaikoina 1950-luvulla, sekä edelleen tätä seuraavina vuosikymmeninä.

Ei ole mitään sellaista inhimillisen toiminnan osa-aluetta, johon tietokoneet eivät pystyisi, argumentoi muun muassa John McCarthy (kts. Weizenbaum 1976, 207). Tietokoneet ajattelevat, oppivat, ratkaisevat ongelmia ja luovat uutta, iloitsivat tekoälytutkijat jo 1950-luvulla (Weizenbaum 1976, 138). Pian koneet pystyisivät kaikkeen siihen mitä ihmisetkin ja voisivat korvata esimerkiksi psykoterapeutit, joiden työ voitiin käsittää pohjimmiltaan informaation prosessointina ja siihen perustuvien päätösten tekona, kuten jokainen muukin inhimillisen toiminnan osa-alue. Ihmisen vakuuttavaan imitointiin kykenevä chatbotti kehitettiin jo 1960-luvulla, jolloin se aiheutti vahvoja emotionaalisia reaktioita, ja pian sen ennakoitiin korvaavan laajasti terapeutit (Weizenbaum 1976, 2–8).

Kaikki eivät kuitenkaan suhtautuneet teknologiaan ja sen kehitykseen yhtä optimistisesti. Yksi tunnetuimmista tekoälyn sekä laajemmin tietokoneisiin liitettyjen mahdollisuuksien kriitikko oli yhdysvaltalainen filosofi Hubert L. Dreyfus (1929–2017). Vuonna 1972 ilmestyneessä teoksessaan *What computers can't do* Dreyfus argumentoi perusteellisesti, kuinka tuolloin esitetyt käsitykset tietokoneiden mahdollisuuksista ja tekoälyn kehittämisestä ovat virheellisiä. Dreyfus pysyi kannassaan, vaikka teknologia kehittyi, ja julkaisi vuonna 1993 päivitetyn version teoksestaan nimellä *What computers still can't do.*

Dreyfusin argumentin ytimessä on tekoälykäsitysten taustalla vallitsevan filosofisen ajatusrakennelman osoittaminen. Dreyfus (1993) jäsentää seikkaperäisesti, millaisista viime kädessä ontologisista ja epistemologisista olettamista muodostuvat ne käsitykset, jotka oikeuttavat tekemään väitteitä ihmisten ja tietokoneiden

viimekätisestä samankaltaisuudesta. Dreyfus (1993) osoittaa myös, miksi ne ovat väärässä. Muun muassa fenomenologiasta ammentavien argumenttien avulla hän osoittaa, kuinka ihmisen olemistapa on perustavanlaatuisen erityinen niin, ettei se ole mitenkään jäljiteltävissä tietokoneiden avulla.

Vaikka Dreyfusin kirjoituksista on kulunut vuosikymmeniä, ovat ne edelleen relevantteja ja mielenkiintoisia esitellessään johdonmukaisen mallin useiden tekoälynäkemysten taustalla olevasta ajatusrakennelmasta. Dreyfus tarjoaa paitsi listan niistä oletuksista, joita meidän on tehtävä, mikäli ajattelemme ihmisen mielen olevan keinotekoisesti jäljiteltävissä ja simuloitavissa, myös useita relevantteja käsitteitä näiden ajatusten jäsentämiseksi.

Tässä tutkielmassa olen kiinnostunut ihmisen ja tietokoneen yhtäläisyyttä ja eroa jäsentävästä argumentaatiosta. Olen kiinnostunut erityisesti siitä, kuinka käsitys informaatiosta ja sen luonteesta vaikuttaa käsitykseen ihmisestä ja koneesta ylipäätänsä rinnasteisina olioina. Dreyfus osoitti, kuinka käsitykset ihmisen kaltaisen tekoälyn mahdollisuudesta pitävät viime kädessä sisällään tietynlaisia ontologisia ja epistemologisia käsityksiä informaation, tiedon ja todellisuuden luonteesta. Näiden käsitysten suhde on se, johon tutkielmassani keskityn. Pidän Dreyfusin näkemystä varteenotettavana ja otan sen lähtökohdakseni, mutta tarkoitukseni on täydentää sitä keskittyen erityisesti informaation käsitteeseen, jota Dreyfus itse ei käytä johdonmukaisesti. Informaation käsite on kuitenkin hyvin relevantti, joten tämän käsitteen avulla voidaan Dreyfusin esittämiin teorioihin on löytää uusia näkökulmia.

Tutkimuskysymykseni on: *millä tavalla käsitys informaation luonteesta vaikuttaa käsitykseen ihmisen ja tietokoneen välisestä samankaltaisuudesta?*

Kysymykseen vastaaminen edellyttää ensinnäkin, että jäsennämme sen osat: mitä tarkoitamme käsityksellä informaation luonteesta, mitä tarkoitamme ihmisen ja tietokoneen välisellä samankaltaisuudella, sekä mitä tarkoitamme vaikuttamisella tässä kontekstissa. Toiseksi, on tehtävä täsmennys siihen, millä laajuudella ja mistä näkökulmasta kysymystä tarkastellaan.

Informaation luonne. Tarkastelen nimenomaan *käsityksiä* informaation luonteesta. Kysyn siis, mitä ja millaista informaation ajatellaan tietyissä näkemyksissä olevan, mistä se koostuu ja onko sillä materiaalinen perusta. En käy kattavasti läpi erilaisia potentiaalisia informaatiokäsityksiä, vaan keskityn työni rajauksen kannalta olennaisiin teemoihin.

Ihmisen ja tietokoneen välinen samankaltaisuus. Tarkastelen ihmisen ja tietokoneen välistä suhdetta, mutta en tässä yhteydessä viittaa suhteella kahden erillisen olion väliseen relaatioon (relation), vaan näiden väliseen yhtäläisyyteen (association). Tätä suhdetta tarkastellessani hyödynnän Espositon ja Baravallen (2023) mallia siitä, millainen on mielekäs analogia. Erilaisia väitteitä ihmisten tai eläinten sekä koneiden samankaltaisuudesta on esitetty kautta filosofian historian, mutta kysymys siitä, onko

ihminen kuin kone, tai edes onko ihminen kuin tietokone, on liian geneerinen eikä siten filosofisesti relevantti (Esposito & Baravalle 2023). Minäkään en pyri tutkielmassani pohtimaan, oikeuttavatko tietyt informaatiokäsitykset sen, että ihminen ”on kuin tietokone”. Kysyn sen sijaan, millä tavalla käsitykset informaation luonteesta mahdollistavat sen, että ihmisten ja tietokoneiden voidaan sanoa olevan analogisia jossakin tietyssä suhteessa eli jakavan joitakin yhteisiä ominaisuuksia tai aspekteja. Nämä jaetut aspektit ovat sittemmin niiden argumenttien pohjalla, jotka väittävät ihmisen mielen olevan kauttaaltaan keinotekoisesti simuloitavissa.

Lähestyn aihettani seuraavanlaisesti. Ensin esittelen lyhyesti tekoälykeskustelun juuria sekä keskeisiä ajattelijoita ja heidän väitteitään. Tämän jälkeen esittelen Dreyfusin esittämän jäsennyksen siitä, kuinka käsitykset liittyvät toisiinsa. Analysoin mallia tarkemmin ja keskityn informaation konseptin näkökulmaan. Tämän jälkeen siirryn tarkastelemaan perusteellisesti itse kysymystäni: jäsennän Dreyfusin ajatusten pohjalta argumentin mallin, jossa ymmärrys informaatiosta toimii perusteluna sille, kuinka ihmiset ja koneet käsitetään. Osoitan, kuinka kyseessä on olennaisesti kaksitasoinen argumentti: ensin on käsitettävä ihmiset ja koneet samankaltaisiksi, ja tämän jälkeen määriteltävä niin informaatio kuin ihminenkin siten, että keinotekoinen representaatio tekoälyn avulla on konkreettisesti mahdollista. Lopuksi esitän lyhyesti jotakin kritiikkiä sekä vaihtoehtoisia näkemyksiä.

Huomautan vielä joistakin aiheeseen liittyvistä rajauksista. Tekoäly on sekä järjestelmänä että tieteenalana kehittynyt vuosikymmenien saatossa merkittävästi, ja vaikka Dreyfusin kritiikkiä voidaan pitää ajankohtaisena, voidaan myös todeta, kuinka se perustuu pitkälti niin sanottuun klassiseen tekoälyyn. Uudempi konnektionistinen lähestymistapa muuttaa asetelman ja haastaa Dreyfusin argumentin. (Niiniluoto 2021, 125–129; Raatikainen 2021, 10–13.) En kuitenkaan tarkastele tässä tarkemmin konnektionismia, vaan tarkastelen Dreyfusin ajattelua aikansa kontekstissa. Sivuan kuitenkin aihetta luvussa 5. En myöskään tarkastele tekoälyn kehityksen alkuaikojen tekstejä sellaisenaan, vaan tarkastelen sen sijaan tekoälyn kehittäjien väitteitä Dreyfusin sekä muutamien muiden myöhempien kirjoittajien referoimina. Pidän tätä kuitenkin tarkastelunäkökulmalleni riittävänä, sillä en tässä tutkielmassa ole kiinnostunut varsinaisesti väitteistä itsessään, vaan niistä ja niihin liittyvästä filosofiasta tehdystä tulkinnasta.

Työni sivuaa useita filosofisia keskusteluja, jotka linkittyvät myös toisiinsa. Ensinnäkin se tarkastelee ja jäsentää yhtä klassisimmista tekoälyyn liittyvistä filosofisista kysymyksistä: kysymystä siitä, missä määrin keinotekoinen älykkyys on ylipäätänsä mahdollista. Lisäksi se osallistuu laajempaan keskusteluun rationalismista, formalisoinnin rajoista sekä todellisuuden abstrahoitavuudesta, mikä voidaan Dreyfusin mukaan jäljittää Platonin idealismiin asti (Dreyfus & Dreyfus 1986). Tämän ohella työ toimii pohjana informaation konseptin filosofiselle tarkastelulle, jota

työn pohjalta on mahdollista syventää. Työ toimii myös ylipäätänsä jo mahdollisesti joiltakin osin vanhentuneiden ajatusrakennelmien päivittäjänä ja esittää niistä uusia näkökulmia.

Tähän kytkeytyy myös työn yhteiskunnallinen relevanssi. Teknologinen kehitys etenee edelleen, ja tekoälyjärjestelmät pystyvät jatkuvasti sellaiseen, minkä niille ajateltiin olevan määritelmällisesti mahdotonta vielä muutama vuosikymmen tai jopa vuosi sitten. Tämän vuoksi on kriittistä jäsentää, paitsi mistä itse teknologioissa on kyse, ennen kaikkea, mitä ajatusrakennelmia ja oletuksia niiden edistämiseen ja implementointiin liittyy. Kysymys ihmisen ja tietokoneen välisestä assosiaatiosta on keskeinen, kun teknologiat vaikuttavat pystyvän yhä ihmismäisempiin toimintoihin. Olennaista on jäsentää nimenomaan laajempia teoreettisia näkemyksiä ja osoittaa ajatuskulkuja, joiden voidaan katsoa olevan kokonaisten sosioteknisten rakennelmien taustalla. On myös löydettävä relevantteja käsitteellistyksiä, joiden avulla voidaan hahmottaa, millaisista asioista tässä sekä teknologisen että yhteiskunnallisen kehityksen vaiheessa on kyse. Filosofiassa voidaan ylipäätänsä ymmärtää olevan kyse nimenomaan tästä – käsitteellistyksien muodostamisesta. Väitän, että ymmärtääksemme meneillään olevaa yhteiskunnallista kehitystä, on relevanttia analysoida nimenomaan informaation käsitettä. Dreyfus on tässä suhteessa merkittävä filosofi, joka osoitti kehityksen ollessa vielä aluillaan, minkälaisiin asioihin meidän on kiinnitettävä huomiota, mikäli haluamme ymmärtää, mistä tässä teknologisessa kehityksessä on kyse.

2 TEKOÄLYKESKUSTELUN HISTORIAA

Jotta voimme tarkastella väitteitä informaation luonteesta, ihmisistä ja tietokoneista, on syytä tehdä lyhyt katsaus historiaan ja tarkastella sitä, mitä aiheesta on viimeisten vuosikymmenien aikana esitetty. Tämän vuoksi teen lyhyen katsauksen tekoälyn ja siihen liittyvän filosofisen keskustelun historiaan. Tekoälyn käsite on itsessään epämääräinen: sillä voidaan viitata joko ”älykkäästi” toimivaan tietokoneohjelmistoon tai tällaisten järjestelmien kehittämisen tieteenalaan (Raatikainen 2021, 7). Tässä yhteydessä käytän käsitettä etupäässä jälkimmäisessä merkityksessä: en puhu tekoälystä vain yksittäisenä ohjelmistona vaan laajempana tieteellisenä pyrkimyksenä kehittää keinotekoinen älykkyys.

Tekoäly-tieteenalan historia on monipuolinen ja linkittyy osaksi laajempaa informaatioteknologian historiaa. Se on mahdollista tulkita osana laajempia historioita ja kehityskulkuja, kuten Frankfurtin koulukunnan teoretisoiman välineellisen järjen voittokulkua (Weizenbaum 1976, 249–251) tai laajemmin länsimaisen filosofian pyrkimystä tiedon totaaliseen formalisointiin (Dreyfus 1993, 67–75). Seuraavaksi keskityn nimenomaan siihen, mitä väitteitä ihmisen ja tietokoneen samankaltaisuudesta on tämän tieteenalan historian saatossa esitetty.

Sinällään näkemyksiä siitä, että ihminen tai laajemmin elolliset oliot voidaan ymmärtää jonkinlaisina koneina, on filosofian historiassa esitetty jo kauan ennen tietokoneiden aikaa. Väitteillä on tyypillisesti pyritty perustelemaan näkemystä siitä, että ihmiset voidaan nähdä laitteina, joiden toiminta perustuu rationaalisiin periaatteisiin. Näiden periaatteiden avulla käytöstä voidaan säädellä ja skaalata sekä tavoitella spesifejä funktioita fysiikan lakien puitteissa. (Bongard & Levin 2021.) Ihmisten, tai elollisten olioiden, ja koneiden yhtäläisyydestä on esitetty erilaisia analogioita, jotka ovat pitkälti sidoksissa aikakauden teknologiaan: esimerkiksi 1600-luvulla elollisia olioita verrattiin kelloihin ja myöhemmin höyryllä toimiviin koneisiin (Esposito & Baravalle 2023). Näin ollen ihmisten vertaaminen tietokoneisiin voidaan nähdä osana laajempaa historiallista jatkumoa, jossa orgaanisten olioiden toiminta pyritään selittämään rationaalisten ja mekaanisten prinsiippien avulla.

Asia ei kuitenkaan ole näin yksinkertainen, ja uudempia analogioita on voitu käyttää nimenomaisesti vanhojen näkemysten osoittamiseksi vääräksi. Esimerkiksi Norbert Wienerin ajatus kyberneettisestä koneesta on eksplisiittinen vastinpari vanhoille koneille, joita hän nimittää ”jäykiksi koneiksi” (rigid machine). Jäykän koneen toiminta on determinististä, mekaanista, joustamatonta ja samalla alistavaa, mutta kyberneettisenä koneena ihminen voi säilyttää autonomiansa informaatiovirtojen välityspisteenä. (Hayles 1999, 105–108.) Dreyfus (1993, 67–69) kuitenkin näkee myös kybernetiikan osana Platonista lähtenyttä jatkumoa, joka tiivistyy tekoälyn kehittämiseen.

Varsinaisen tekoälyn lähtökohtana voidaan pitää 1950-lukua. Tietokoneita oli kehitetty jo aikaisemmin, ja esimerkiksi Alan Turing oli pohtinut kysymystä potentiaalisesta koneellisesta älykkyydestä 1940-luvulla (Raatikainen 2021, 8–9). Varsinaisena tutkimusalana tekoälyn voidaan katsoa syntyneen vuonna 1956 Dartmouth Collegessa järjestetyn kesäkonferenssin yhteydessä. Tämän uuden tutkimusalan julkilausuttuna tavoitteena oli tutkia ja selvittää oppimisen ja älykkyyden perustavanlaatuisimpia piirteitä siten, että on mahdollista rakentaa niitä simuloiva kone (McCarthy ym. 2006). Konferenssin osallistujat, eräänlaiset tieteenalan perustajajäsenet, ovat henkilöitä, joiden ajatuksia kohtaan Dreyfus suuntaa suuren osan kritiikistään. Heihin kuuluvat muun muassa artificial intelligence -termin keksinyt John McCarthy sekä Marvin Minsky, Allen Newell ja Herbert Simon. (Bringsjord & Govindarajulu 2018.) Heidän näkemyksensä voidaan lukea osaksi laajempaa, tekoälyoptimistista ajatusmaailmaa, jopa eräänlaista ideologiaa, jolle Joseph Weizenbaum (1976) on antanut nimen ”the artificial intelligentsia”.

Tekoälyllä ei edelleenkään ole tarkkaa määritelmää. Se on saanut lähtökohtansa pyrkimyksestä simuloida ihmisen älykkyyttä ja oppimista ja toisintaa tämä tietokoneella. Käsitteellisesti voidaan erottaa toisistaan niin sanottu heikko sekä vahva, tai yleinen tekoäly, joista jälkimmäinen on ainoastaan hypoteettinen. Siinä missä kaikki nykyiset järjestelmät edustavat heikkoa tekoälyä, sellainen järjestelmä, joka todella kykenisi laajasti ihmisen kaltaiseen toimintaan ja jolla olisi ihmisen tasoinen älykkyys, olisi vahva tekoäly. (Pesonen 2021.) Vaikka tämä käsitteellinen eronteko on peräisin vasta 1990-luvun lopulta, on tekoälyn kehittäminen alusta lähtien pyrkinyt nimenomaan kohti vahvaa tekoälyä eli ihmisen älykkyyden jäljittelyä. On kuitenkin selvää, että älykkyyden ja älykkyyttä vaativan tehtävän käsitteet ovat epämääräisiä. Sillä, millaiseksi älykkyys määritellään, on vahvoja implikaatioita paitsi sille, mitä pidetään tekoälynä, myös sille, mitä pidetään ihmisen älykkyytenä ja sen ilmiasuna. Nimenomaan tekoälyyn liittyi lähtökohtaisesti vahvoja ajatuksia siitä, kuinka ihmiset ja tietokoneet muistuttavat toisiaan. Jotta voidaan tuottaa älykäs järjestelmä, on käsitettävä äly tietyllä tavalla, mikä taas ihmiseen liittyen on laaja filosofinen kysymys.

Tekoälyllä voidaankin sanoa olevan juurensa filosofiassa (Bringsjord & Govindarajulu 2018), ja ajatukseen keinotekoisen älykkyyden mahdollisuudesta on sisäänrakennettu tiettyjä filosofisia sitoumuksia riippuen ennen kaikkea vahvasti siitä, kuinka määritellään älykkyys ja kuinka käsitteellistetään se toiminta, jota koneella pyritään tekemään. On tehtävissä olennainen ajallinen eronteko niin sanottuun klassiseen tekoälyyn sekä uudempaan konnektionistiseen paradigmaan. Näiden kahden pyrkimyksen välinen ero ilmentää merkittävää eroa siinä, kuinka älykkyyden käsite ja lähtökohdat ymmärretään. Klassisen tekoälyn, jota 1950-luvulta alkaen kehitettiin, idea lähtee liikkeelle Boolen kalkyylin sekä Turingin koneen

ideoista. Tämän mallin mukaan kaikki ajattelu on sekventiaalista komputointia, tietyn sanaston ja sääntöjen avulla tapahtuvaa laskemista. Älykkyys tarkoittaa tällöin sitä, että kykenee suorittamaan virheettömästi tällaisia toimituksia. Voidaan puhua myös ajattelun kielestä, joka tässä yhteydessä nähdään laskennallisesti reprodusoitavissa olevana (Haaparanta 1995, 41). Tämä käsitys ihmisen ja tietokoneen yhteisestä ”ajattelemisen tavasta” ja älykkyydestä symbolijonojen sääntöjen mukaisena manipulointina on ytimessä niissä väitteissä, joissa ihmisiä ja tietokoneita pidetään samankaltaisina.

Ajatus kognitiivisesta simulaatiosta oli keskeinen 1960-luvulla ja olennainen osa varhaista tekoälyn kehitystä. Kognitiivisessa simulaatiossa pyritään määritelmällisesti simuloimaan ihmisen kognitiivisia prosesseja, kuten ajattelua, ongelmanratkaisua ja hahmontunnistusta. Tämä edellyttää näiden prosessien käsittämistä tietynlaisina. Kognitiivisen simulaation ja ongelmanratkaisun parissa työskennelleet Newell ja Simon esittivät useita vahvoja ajatuksia ihmisten ja tietokoneiden yhteydestä. Heidän mukaansa sekä aivot että tietokoneet ovat yleisiä, symboleja manipuloivia laitteita ja tietokone on mahdollista ohjelmoida käsittelemään informaatiota samalla tavalla kuin ihmisen aivot. Heidän mukaansa subjektin käyttäytyminen perustuu ohjelmistoon, joka koostuu elementaarisista informaatioprosesseista. Ihmisen älykäs käyttäytyminen voidaan nähdä tuloksena monimutkaisesta mutta rajallisesta sääntöjen kokoelmasta. Kaikki tämä tekee mahdolliseksi sen, että ihmisen älykäs käyttäytyminen voidaan toisintaa tietokoneohjelman avulla. (Dreyfus 1993, 93–96; 163–174.)

Newellin ja Simonin, kuten muidenkin varhaisten tekoälytutkijoiden, lähestymistapa perustuu edellä mainittuun klassiseen tekoälyyn, jota on kutsuttu myös nimellä ”vanha kunnon tekoäly” (Good, Old-Fashioned AI, GOFAI). Dreyfus (1993, ix) nimeää GOFAI:n tieteellistä kehitystä haittaavaksi paradigmaksi. Sen idea voidaan tiivistää pyrkimykseksi tuottaa ihmisen kaltainen älykkyys toiminnallisten sääntöjen, symbolien manipuloinnin ja loogisen päättelyn avulla. Keskeistä on yksityiskohtainen ohjelmointi – tekoälyjärjestelmä toimii kuten se on ohjelmoitu toimimaan, ja sen käyttäytymistä säätelevät formaalit säännöt. (Niiniluoto 2021, 118; Raatikainen 2021, 11.) Lähestymistapaa on kutsuttu myös symboliseksi tekoälyksi, sillä siinä on kyse symbolien manipuloinnista: manipuloitavilla syötteillä on oltava representationaalinen sisältö sekä syntaksi (Stoutland 1995, 208–209).

Samaa lähestymistapaa edusti myös kognitiotieteilijä ja varhaisen tekoälytutkimuksen pioneeri Marvin Minsky. Minskyn esittämä määritelmä tekoälylle on tunnettu: tekoäly tarkoittaa tiedettä, joka pyrkii rakentamaan ihmiseltä älyä vaativia tehtäviä suorittavia koneita (Raatikainen 2021, 7). Määritelmän keskiössä on älyn käsite, ja on merkittävää, millainen käsitys älystä siihen ja samalla koko varhaiseen tekoälyn ajatukseen sisältyi. Minskyn keskeinen näkemys oli, kuinka

ihmisen älykäs käytös on heurististen sääntöjen määräämää toimintaa, jonka toisintaminen tietokoneella on mahdollista (Dreyfus 1993, 163–164). Hänen mukaansa mentaaliset prosessit vastaavat tietokoneohjelmien elementaarisia prosesseja: niissä on kyse arbitraarisesta symboliassosiaatiosta, skeemojen varastosta, ehdollisesta transferoituvuudesta ja vastaavasta. Minskyn mukaan ei ole syytä olettaa, miksi koneet eivät pystyisi siihen, mihin ihmisetkin. (Dreyfus 1993, 191–197.) Minsky esitti jopa suoran vertauksen siitä, että ihmiset ovat lihasta tehtyjä koneita muiden koneiden rinnalla (Weizenbaum 1976).

Tämän päivän tekoäly ja siihen liittyvä ajattelu on siinä mielessä ratkaisevasti erilaista, että kritisoidusta klassisen tekoälyn paradigmasta on siirrytty konnektionismiin. Konnektionismilla viitataan malliin siitä, kuinka älykkään toiminnan voidaan katsoa perustuvan hermoverkkoon ja kuinka tekoälyn luomisessa voidaan soveltaa keinotekoisia hermoverkkoja (Pesonen 2021). Dreyfus (1993, xiv–xv) itse tervehtii paradigman muutosta ilolla ja toteaa, että neuroverkkomallit voivat tarjota huomattavasti paremman vastauksen tekoälyn kehittämiseen kuin symbolinen malli. Myös Niiniluoto (2021, 123–129) toteaa syväoppivan neuroverkkomallin välttävän Dreyfusin klassiseen tekoälyyn kohdistaman kritiikin. Syväoppiva järjestelmä vaikuttaisi toimivan juuri sillä tavalla, miten Dreyfus kuvasi taitavan ekspertin toimintaa: ei käsitteellisesti eikä sääntöihin pohjautuen, vaan kokemukseen ja eräänlaiseen ”intuitioon” perustuen. Konnektionismi tuo kuitenkin mukanaan uusia kysymyksiä sekä uusia kritiikin kohteita, ja konnektionismi liittyy edelleen vahvasti niihin filosofisiin taustaoletuksiin, joita esimerkiksi Dreyfus kritisoi (Aizawa 1992).

Yleisesti voidaan sanoa, että tekoälyn kehittäminen nimenomaan teko*älynä* on osa laajempaa filosofista ajatusta, jossa älykäs toiminta, paikoin myös ihmisen toiminta, ymmärretään viime kädessä informaation käsittelyyn perustuvana, tietokoneen suoritettavissa olevana prosessina. Tekoälyn kehittämiseen on liittynyt erilaisia ajanjaksoja, ja voimme esittää erilaisia tulkintoja siitä, kuinka paljon vanhoja oletuksia on haastettu ja kuinka pitkälti kyse on samoista ajatuksista eri muodossa. Konnektionismi voidaan nähdä klassisen tekoälyn mallin suorana haastajana. Yhtä lailla tekoäly sinänsä voidaan nähdä osana samaa ja laajempaa käsitteellistämisen tapaa ihmisen älystä viime kädessä irrallisena konseptina, mitä esimerkiksi Aizawa (1992) ja Hayles (1999) ovat tulkinneet. Tämän tutkielman kannalta olennaista on ymmärtää tekoälykeskustelun laajemmat implikaatiot: ne ajatukset, jotka johtavat näkemykseen siitä, että voimme ylipäätänsä puhua ihmisen kaltaisesta älykkyydestä määriteltävänä konseptina sekä pitää mahdollisena toisintaa tämä tietyn teknisen järjestelmän avulla.

3 DREYFUSIN ESITTÄMÄ KRITIIKKI

Hubert Dreyfus suuntasi kritiikkinsä nimenomaan klassisen tekoälyn ajatusta kohtaan, joka oli vallitseva paradigma 1950-luvulta 1980-luvulle (kts. Pesonen 2021). Dreyfusin ansio on siinä, kuinka hän pyrki osoittamaan klassisen tekoälyn taustalla olevat filosofiset juuret (Kenaw 2008). Hän linkittää klassisen tekoälyn Platonista lähtöisin olevaan rationalistiseen filosofiaan, jossa suurena pyrkimyksenä on esittää kaikki tieto formaalina, erillisistä elementeistä ja säännöistä koostuvana rakenteena (Dreyfus 1993, 67–72; Dreyfus & Dreyfus 1986, 2–4). Tätä pyrkimystä Dreyfus kritisoi 1960-luvulta alkaen ja oli samalla ensimmäisiä ja tunnetuimpia tekoälyn kriitikkoja. Dreyfus itse toimi filosofian professorina Kaliforniassa, ja ammensi filosofiassaan eksistentialismista ja fenomenologiasta (Su & Luvaanjalba 2021). Hän hyödyntää kritiikissään fenomenologian näkökulmaa ihmisen kokemisen ja olemisen tavan erityislaatuisuudesta, mutta hänen kritiikkinsä on myös laajempaa.

Seuraavaksi käyn läpi Dreyfusin kritiikin keskeisen sisällön ja tiivistän hänen esittämänsä argumentit. Käyn myös läpi Dreyfusin esittämän olennaisen jäsennyksen siitä, millaiset olettamat klassisen tekoälyn idean taustalla voidaan tunnistaa. Luvussa neljä siirryn käsittelemään Dreyfusin esille nostamaa ajatusrakennelmaa tarkemmin informaation konseptin näkökulmasta.

3.1 Dreyfusin yleinen teesi

Dreyfusin väitteen ydin on, ettei ihmisen kaltaisen älykkyyden rakentaminen tietokoneella ole mahdollista. Tämä johtuu siitä, että ihmisen ajattelemisen ja olemisen tapa on perustavalla tavalla erilainen kuin koneen. Älykkyys edellyttää ymmärrystä, mikä puolestaan edellyttää taustalleen inhimillisen kokemuksen, jonka mahdollistavat kehollisuus, vuorovaikutus materiaalisen todellisuuden kanssa sekä kasvaminen kulttuuriin (Dreyfus 1993, 2–3). Ihmisen kyky havainnoida ja arvioida ympäristöä ja tilanteita perustuu ihmisenä olemisen pluraalisuuteen ja kulttuuriin, eikä sitä voida kuvata symbolisella rakenteella (Negrotti 2019).

Ihmisenä olemisessa ja ihmisen älykkäässä toiminnassa on siis Dreyfusin mukaan jotain sellaista, mitä ei mitenkään voida reprodusoida tietokoneella, sikäli kun tietokone ymmärretään yleiseksi, symboleja manipuloivaksi laitteeksi. Tietokoneen operaatiot perustuvat täysin eksaktiin ja määriteltyyn dataan, jota käsitellään tiukasti määrättyjen sääntöjen mukaan. Tietokone on passiivinen informaation prosessoija, joka ottaa vastaan syötteitä ja palauttaa tulosteita. (Dreyfus 1993, 240–243.) Vaikka esimerkiksi John McCarthy on esittänyt ihmisen olevan

vastaava informaation prosessoija (kts. Dreyfus 1993, 243), Dreyfus on jyrkästi eri mieltä. Hänen mukaansa ihmisen ajattelu ja toiminta nimenomaan eivät ole selitettävissä tietokoneen mallin mukaisesti.

Dreyfus kutsuu kritisoimaansa mallia teoksensa uudessa osassa representationalismiksi. Se, kuten ylipäätänsä klassinen tekoäly, perustuu Dreyfusin mukaan kartesiolaiseen ideaan siitä, että ymmärrys tarkoittaa sopivien symbolisten representaatioiden ymmärtämistä ja käyttämistä. Tekoälyn kehittäminen on Dreyfusin mukaan muuttanut Descartesin, sekä myöhempien rationalistien kuten Leibnizin ideat systemaattiseksi tutkimusohjelmaksi. Representationalismin ajatus on, että jokapäiväisen ymmärryksen taustalla on implisiittisten uskomusten systeemi. (Dreyfus 1993, xi–xvii.)

Representationalismi ja klassisen tekoälyn näkökulma ovat Dreyfusin mukaan väärässä. Tämän todistamiseen hänen kirjansa perustuvat. Dreyfus jäsentää toisaalta sitä, miksi klassiseen tekoälyyn liittyvät näkemykset ovat itsessään virheellisiä, sekä sitä, millä muulla tavalla ihmisen toiminta tulisi ymmärtää, kuin sääntöpohjaisena informaation prosessointina. Ihmisen erityislaatuisuuden keskiössä on kehollisuus. Kyse ei ole vain materiaalisuudesta sinänsä, vaan ihmisen kehollisuus mahdollistaa ihmiselle erityisen intuition, jota ei voida toisintaa tietokoneella. (Dreyfus 1993, 254–255; Dreyfus & Dreyfus 1986.) Dreyfus selittää asian hyvin seikkaperäisesti, ja hänen vuonna 1986 veljensä Stuartin kanssa julkaisema teos (Dreyfus & Dreyfus 1986) keskittyy perustelemaan nimenomaan tätä argumenttia. En selitä sitä tässä yksityiskohtaisemmin, mutta kertaan lyhyesti sen keskeisen sisällön.

Ihmisen kehollisille olemiselle on ainutlaatuista sisäinen horisontti, joka mahdollistaa erityisen ennakointikyvyn. Tämän ennakointikyvyn yleinen luonne on ainutlaatuinen, ja se on siirrettävissä aistielimestä toiseen, mikä ilmenee esimerkiksi silmän ja käden koordinaatiossa esinettä käsiteltäessä. Ihmisen kehollisen olemisen tapa mahdollistaa sen, että ihminen kykenee toimiessaan ohittamaan formaalin analyysin, joka koneilta vaaditaan. (Dreyfus 1993, 249–255.) Se, miten etenkin taitava ihminen käyttäytyy, perustuu tiedostamattomiin prosesseihin, joita ei voida jäljitellä ja toisintaa formaaleilla säännöillä. Keskeinen esimerkki tästä on aloittelijan taidon vertaaminen eksperttiin: aloittelijan on seurattava tietoisesti sääntöjä, jolloin hänen toimintansa ei ole vielä sujuvaa, mutta ekspertti voi toimia intuitionsa varassa, ja näin ollen ekspertin toimintaa voidaan kutsua arationaaliseksi. (Dreyfus & Dreyfus 1986, 35–36.)

Tästä huolimatta tekoälyn kehittäjät pyrkivät ymmärtämään ihmisen älykkään käytöksen nimenomaan sääntöpohjaisena informaation prosessointina, minkä johdosta klassisen tekoälyn paradigma on Dreyfusin (1993, ix) mukaan degeneratiivinen, se johtaa koko tieteenalaa väärään suuntaan. Kuten edellä todettua, Dreyfus ei pidä tätä ainoastaan aikakauden teknologian kehittäjien syynä, vaan

linkittää klassisen tekoälyn paradigman osaksi koko länsimaisen filosofian historiaa. Hänen teoksensa suurimpia ansioita on perusteellinen jäsennys siitä, mistä paradigmassa oikein on kyse. Mitä psykologisia, epistemologisia ja viime kädessä ontologisia oletuksia meidän on tehtävä, jotta voimme ajatella ihmisen älykkyyden olevan jäljiteltävissä tietokoneella?

3.2 Dreyfusin jäsentämät neljä olettamaa

Dreyfus luettelee neljä oletusta, jotka ovat tämänhetkisen tekoälyoptimismin taustalla. ”Tällä hetkellä” viitataan kirjan kirjoitusajankohtaan 70-lukuun, ja Dreyfus siteeraa runsaasti tekoälyn kehittämisen alkuaikojen keskeisiä henkilöitä, kuten Marvin Minskyä, Allen Newellia ja Herbert Simonia. Hän ei kuitenkaan osoita kritiikkiään ainoastaan yksittäisten henkilöiden ajatuksia kohtaan, vaan tarkastelee klassisen tekoälyn kehittämistä nimenomaan osana laajempaa filosofista ajatusmaailmaa.

Dreyfusin (1993, 156) jäsentämät olettamat ovat seuraavat

(1) biologinen olettama: aivot prosessoivat informaatiota diskreetteinä operaatioina, jotka ovat biologinen vastine on/off -kytkennöille
(2) psykologinen olettama: mieli voidaan käsittää välineenä, joka operoi formaaleihin sääntöihin perustuvan informaation perusteella; tietokone on malli mielen toiminnasta
(3) epistemologinen olettama: kaikki tieto on formalisoitavissa, ja siten esitettävissä loogisina relaatioina
(4) ontologinen olettama: kaikki relevantti informaatio maailmasta on pohjimmiltaan analysoitavissa tilanneriippumattomina määrättyinä elementteinä; kaikki, mitä on olemassa, on joukko faktoja, jotka ovat loogisesti erillisiä toisistaan

Nämä olettamat eivät ole toisistaan erilliset, vaan ne muodostavat perustelujen ketjun. Ensimmäinen argumentti on tekoälyn kehittämisen tavoite: On mahdollista rakentaa täysin ihmisen kaltainen tekoäly. Tämä pätee sen vuoksi, että (1) aivot prosessoivat informaatiota diskreetteinä, eli toisistaan irrallisina, erillisinä operaatioina. Tämä pätee sen vuoksi, että (2) mieli on väline, joka operoi formaaleihin sääntöihin perustuvan informaation perusteella. Tämä pätee sen vuoksi, että (3) kaikki tieto, kaikki mitä älykäs toiminta vaatii, on formalisoitavissa ja esitettävissä loogisina relaatioina. Ja tämä pätee siksi, että (4) kaikki relevantti informaatio maailmasta ylipäätänsä on analysoitavissa tilanneriippumattomina elementteinä, siksi että kaikki mitä ylipäätänsä on olemassa, on joukko toisistaan loogisesti erillisiä faktoja, toisin sanoen diskreettiä dataa.

Jotta voimme todistaa tekoälyn kehittämisen tavoitteen kunnolla vääräksi, on esitettävä jokaisen olettaman virheellisyys. Sinällään väite kumoutuu jo ensimmäiseen olettamaan: Dreyfus (1993, 159–162) esittelee lyhyesti, ettei empiirinen tutkimus tue oletusta aivojen toiminnasta tietokoneen kaltaisena laitteena on-off-kytkentöineen. Ja vaikka tukisikin, väite mielen toiminnasta sääntöihin perustuvana informaation prosessointina on väärä. Tämän tutkielman kannalta kiinnostavimpia ovat kuitenkin epistemologinen ja ontologinen taso. Dreyfus (1993, 190) tarkentaa, mitä epistemologisella olettamalla tarkoittaa: se pitää sisällään ensinnäkin oletuksen siitä, että kaikki ei-sattumanvarainen käytös voidaan formalisoida, ja toiseksi siitä, että tämän formalisoinnin avulla käytös voidaan toisintaa. Mikä tahansa ihmisen toiminta olisi siis abstrahoitavissa formaaliksi rakenteeksi, ja tämän jälkeen siirrettävissä sellaisenaan toiseen kohteeseen. Kuten luvussa kaksi osoitettiin, klassinen tekoäly perustui lähtökohtaisesti nimenomaan tähän ajatukseen. Dreyfus (1993, 225–227) huomauttaa kriittisesti, ettei edellä esitettyjä olettamia lainkaan kyseenalaistettu tekoälytutkijoiden piirissä, vaan niitä pidettiin välttämättöminä lähtökohtina. Vasta neuroverkkomenetelmien kehittäminen muutti tilannetta – mikä ei kuitenkaan tarkoittanut, ettei perimmäisiä ontologisia olettamia voisi nähdä myös niiden taustalla (Dreyfus 1993, xxx–xxxv).

Pidän Dreyfusin esittämää näkemystä toimivana ja varteenotettavana. Sitä on kuitenkin mahdollista tarkentaa, ja avainasemassa tässä on informaation käsite. Käsitykset informaatiosta ovat keskeinen osa Dreyfusin argumenttia, mutta hän ei kuitenkaan paneudu niihin erityisen systemaattisesti, vaikka mainitseekin niiden olennaisuuden tekoälykeskustelulle (Dreyfus 1993, 166). Seuraavaksi tarkoituksenani on täydentää Dreyfusin argumenttia tarkastelemalla kysymystä ihmisen ja koneen rinnasteisuudesta nimenomaan informaation käsitteen kautta.

4 ARGUMENTTI IHMISEN JA TIETOKONEEN ANALOGIASTA

Dreyfus argumentoi, kuinka ajatus ihmisen ja tietokoneen samankaltaisuudesta perustui viime kädessä ontologiseen olettamaan siitä, että kaikki relevantti informaatio maailmasta on analysoitavissa erillisinä, riippumattomina elementteinä. Näiden sääntöpohjainen käsittely loogisena rakenteena on älykkyyttä, ja tämä rakenne olisi mahdollista toisintaa tietokoneella. Tämä olettama on kuitenkin Dreyfusin mukaan väärä – ei ole niin että kaikki relevantti informaatio olisi esitettävissä erillisistä elementeistä koostuvana datana, eikä myöskään niin, että ihmisen mieli prosessoisi sitä sääntöpohjaisesti. (Dreyfus 1993, 206–207; 225–227.)

Otan Dreyfusin ajatuksen lähtökohdakseni. Sitä on kuitenkin mahdollista täydentää siihen implisiittisesti olennaisesti kuuluvalla informaation käsitteellä. On huomionarvoista, ettei Dreyfus itse käytä informaation käsitettä johdonmukaisesti. Tarkalleen hän kirjoittaa ontologisen oletuksen olevan, että ”kaikki olennainen älykkäälle käytökselle voidaan ymmärtää määriteltyinä yksittäisinä elementteinä[1]” (Dreyfus 1993, 206). Toisin sanoen, informaation käsite ei suoraan ole hänen esittämänsä ontologisen argumentin keskiössä. Toisissa yhteyksissä hän käyttää sekä informaation että datan käsitteitä, mutta osittain ristiin, eikä myöskään rakenna argumentaatiotaan näiden käsitteiden varaan.

Väitän kuitenkin, että nimenomaan informaation käsitteen avulla voimme saada täsmällisemmän käsityksen siitä, mistä ihmistä ja konetta rinnastavassa ajattelussa ja toisaalta tekoälyn ajatuksessa on kyse. Tämä johtuu käsitteen keskeisestä roolista alkuperäisissä tekoälyargumenteissa, joissa ihmisiä ja tietokoneita pidettiin yleisinä, informaatiota prosessoivina systeemeinä (Dreyfus 1993, 164–165). Toisin sanoen, ei ole pohdittava ainoastaan sitä, millaisia ihmisten oletetaan olevan ja miksi nämä oletukset eivät päde, vaan on keskeistä kysyä, mistä informaatiossa itsessään on kyse. Samalla avautuu uusia mahdollisuuksia asettaa tekoälyn idea kritiikin alle, ja tämä kritiikki täydentää Dreyfusin omaa, ennen kaikkea fenomenologiaan pohjautuvaa kritiikkiä.

Esitän kysymykseni muodossa *millä tavalla käsitys informaation luonteesta vaikuttaa käsitykseen ihmisen ja tietokoneen välisestä samankaltaisuudesta?* Seuraavaksi tarkastelen tarkemmin, mistä väitteessä nimenomaan tässä muodossa on varsinaisesti kyse, ja jäsennän auki sen argumenttirakenteen. Ensiksi on tarkasteltava, mitä ylipäätänsä tarkoitan samankaltaisuudella. Kuten aiemmin todettua, ei ole mielekästä lähestyä samankaltaisuutta, yhtäläisyyttä tai analogisuutta yleisen tason luonnehdintana, vaan kysyä, mitä sellaisia yhteisiä aspekteja ihmisellä ja tietokoneella on, jotka oikeuttavat

[1] “everything essential to intelligent behavior must in principle be understandable in terms of a set of determinate independent elements”

näiden rinnastamisen. Kuten kysymyksen ensimmäisestä osasta käy ilmi, tarkoitus on tarkastella nimenomaan informaatioon liittyviä aspekteja.

Tutkielmani kannalta olennaista on todeta, että ihmisen ja tietokoneen toiminnassa voidaan ylipäätänsä tulkita olevan jonkin asteen samankaltaisuus. Tämä samankaltaisuus voidaan nimetä ”informaation prosessoinniksi”, ”älyksi”, ”älykkääksi toiminnaksi” tai vastaavilla käsitteillä. Tällaista näkemystä edustaa esimerkiksi Minskyn väite siitä, että mentaaliset prosessit vastaavat tietokoneiden elementaarisia prosesseja. Voimme toistaiseksi jättää älyn ja älykkyyden hämäriksi käsitteiksi, mutta todeta samalla, kuinka ne tulevat spesifillä tavalla ymmärretyksi tiettyä teknologiaa kehittäessä ja ennen kaikkea käsitteellistettäessä ja kehystettäessä tämä teknologia tietyllä tavalla.

Toisin sanoen ihmisen ja tietokoneen voidaan katsoa jakavan jonkinlaisen yhteisen ajattelemisen, päättelyn tai mentaalisen prosessoinnin tavan. Se tulee useissa käsityksissä nimetyksi informaation prosessoinniksi, mutta Dreyfusin oma argumentti on, että ihmisen tapauksessa kyseessä ei nimenomaan ole informaation prosessointi. Siksi hän puhuukin ”informaation prosessoinnista” lainausmerkeissä, aina kun puhuu ihmisen mentaalisesta toiminnasta.

Väite, jota olemme tarkastelemassa, on alustavalta muodoltaan seuraavanlainen päättelyketju, jossa tietyistä premisseistä (P) seuraa tiettyjä johtopäätöksiä (J). On huomionarvoista, ettei yhteys kahden johtopäätöksen, analogisuuden ja tekoälyn rakentamisen mahdollisuuden kanssa ole itsestään selvä.

P(1): Informaatio on [määritelmä]
↓
P(2): …
↓
P(3): …
↓
…
↓
J(1): Ihmisten ja tietokoneiden välillä vallitsee analogia
↓
(J(2): On mahdollista rakentaa ihmisen kaltainen tekoäly)

Koska seuraamme Espositon ja Baravallen (2023) käsitystä mielekkäästä analogiasta, voimme tarkentaa johtopäätökseen johtavia perusteluja määritelmillä siitä, mitä analogialla tässä yhteydessä tarkoitamme. Kuten yllä todettiin, kyse on ajatuksesta, että ihminen ja tietokone jakaisivat jonkinlaisen yhteisen ”älyn”, ”mentaalisen prosessoinnin” tai ”informaation prosessoinnin” tyylin. Kutsun sitä alustavasti Dreyfusin tavoin ”informaation prosessoinniksi ” lainausmerkeissä, tiedostaen, että

sen sisältämä informaation käsite on kiistanalainen. Päättelyketjumme näyttää nyt tältä.

P(1): Informaatio on [määritelmä]
↓
P(2): …
↓
P(3): …
↓
…
↓
P(n-1): Ihmisten älykäs toiminta on pohjimmiltaan samanlaista ”informaation prosessointia” kuin se, mitä tietokoneet suorittavat
↓
P(n): Ihmiset ja tietokoneet jakavat sellaisen yhteisen ominaisuuden, että ihmisen toiminta on mielekkäällä tavalla rinnastettavissa tietokoneen toimintaan ja toisin päin
↓
J(1): Ihmisten ja tietokoneiden välillä vallitsee analogia
↓
(J(2): On mahdollista rakentaa ihmisen kaltainen tekoäly)

Seuraavaksi tehtävänämme on jäsentää väitteen muut osat. Mitä informaatiosta on tarkalleen ottaen väitettävä, jotta voimme muodostaa esitetyt johtopäätökset? On syytä huomioida, että kysymys on erittäin laaja. Sen sijaan, että lähtisin kartoittamaan kaikkia mahdollisia suuntia, joihin kysymyksestä voi edetä, pyrin ennen kaikkea osoittamaan kuinka informaation käsitteen avulla on ylipäätänsä mielekästä täydentää Dreyfusin esittämää ajatusta. Tämä edellyttää, että jäsennän yleisellä tasolla sen, kuinka informaatioon liittyvät väitteet tai oletukset liittyvät toisiinsa, ja kuinka ne vastaavasti liittyvät laajempaan argumenttiin, jossa ihmisen ja tietokoneen välillä vallitsee analogia.

On syytä huomioida, että oletukset informaatiosta johtavat itse asiassa kahteen johtopäätökseen, joiden välinen yhteys ei ole niin selvä, kuin se on oheisessa kaaviossa esitetty. Ensinnäkin, ne johtavat ihmisen ja tietokoneen analogisuuteen, joka on selitetty samanlaisten ominaisuuksien jakamisella. Toiseksi, ne johtavat mahdollisuuteen rakentaa ihmisen kaltainen tekoäly. Tämä mahdollisuus on otettu itsestään selvänä analogian myötä sinänsä, mikä ilmenee esimerkiksi Newellin ja Simonin väitteessä siitä, että koska tietokoneet ja ihmisaivot ovat samanlaisia symboleja manipuloivia laitteita, voidaan ohjelmoida ihmisen tavoin käyttäytyvä tietokone (Dreyfus 1993, 170–176). Tämä on kuitenkin syytä problematisoida. Ne oletukset informaatiosta, jotka oikeuttavat väitteen ihmisen ja tietokoneen

samankaltaisuudesta, eivät suoraan oikeuta väitettä tekoälyn mahdollisuudesta. Se vaatii lisää oletuksia, jotka liittyvät paitsi informaatioon, myös ihmiseen ja älykkyyteen sinänsä.

Tämä ilmenee havainnollisesti Dreyfusin (1993, 206–207) huomiossa siitä, että datan, jolla tietokone operoi, on oltava diskreettiä, eksplisiittistä ja määriteltyä. Tällainen data on sitä informaatiota, jota tietokone prosessoi. Jotta ihmisen ja tietokoneen välillä olisi analogia, myös ihmisen prosessoiman informaation on oltava sellaista. Asiaan liittyy kuitenkin kaksi kyseenalaistusta, jotka Dreyfus esittää toistensa yhteydessä, mutta jotka johtavat itse asiassa kahteen erikseen tarkasteltavaan oletukseen. Dreyfus (1993, 206–207) toteaa, ettei ole syytä olettaa ensinnäkään, että tällaista dataa on ihmisten maailmassa *saatavilla*, eikä toiseksikaan, että tällaista dataa on *olemassa*. Kysymys oikeanlaisen datan olemassaolosta on metafyysinen ja se vaaditaan, jotta argumenttirakenteen ensimmäinen osa pätee, eli jotta ihmisen ja tietokoneen välisen analogian voidaan katsoa pätevän. Vaaditaan kuitenkin myös tämän datan saatavuus sekä ihmisen mahdollisuus hyödyntää sitä, toisin sanoen konkreettisia keinoja, jotta varsinainen tekoäly voitaisiin rakentaa.

Lähden seuraavaksi jäsentämään tarkemmin näitä oletuksia, toisaalta ihmisen ja tietokoneen analogisuuden ja toisaalta ihmisen kaltaisen tekoälyn mahdollisuuden taustalla. Dreyfusin itsensä jäsentämässä olettamassa oli neljä tasoa: biologinen, psykologinen, epistemologinen ja ontologinen olettama. Tarkastelemani käsitykset informaatiosta poikkileikkaavat tätä jaottelua, mutta eivät ole suoraan rinnasteisia sille. Tarkoitukseni on ikään kuin yhdistää Dreyfusin epistemologinen ja ontologinen olettama tietyiksi väitteiksi informaatiosta, ja tämän lisäksi täydentää niitä muilla väitteillä.

4.1 Immateriaalinen informaatio

Lähden liikkeelle yleisen tason kysymyksestä: mitä informaation on pohjimmiltaan ajateltava olevan, jotta rinnastus voidaan tehdä. Millainen olio, entiteetti tai käsite on se, jonka prosessoinnin sanotaan yhdistävän ihmisiä ja tietokoneita? Kuten edellä on todettu, Dreyfusin mukaan tämä prosessoitava informaatio on ymmärrettävä diskreeteistä elementeistä koostuvana ja määriteltynä. Tätä on kuitenkin mahdollista täydentää. Otan avukseni kybernetiikan ja posthumanismin historiaa tutkineen N. Katherine Haylesin (1999) väitteen siitä, että tämän informaation on lisäksi oltava olennaisesti immateriaalista – abstraktia ja materiasta riippumatonta. Tämä väite täydentää Dreyfusin esittämää ontologista olettamaa, eli oletusta irrallisista elementeistä koostuvasta prosessoivasta todellisuudesta. Huomio siitä, että tämä irrallisuus on olennaisella tavalla irrallisuutta materiaalisesta todellisuudesta, on

olennainen. On mahdollista käsitteellistää materiaalisia olioita yhdistävä informaation prosessoinnin taso, joka on tästä materiasta irrallinen. Sinällään tämänkaltainen näkemys voidaan nähdä perustyyppisenä länsimaiselle filosofialle sinänsä, joka keskittyy nimenomaan ajatuksiin, käsitteisiin ja muuhun olennaisesi immateriaaliseen (Hayles 1999, 195). Alustava väitteeni on siis, että informaation on oltava olemukseltaan immateriaalista, jotta johtopäätös ihmisen ja tietokoneen analogisuudesta voidaan muodostaa.

Kuinka tämä käsitys on suhteessa siihen, mitä informaation käsitteellä ylipäätänsä on tarkoitettu? Informaatiota on määritelty eri tavoin, niin semanttisesti kuin laskennallisestikin. Siitä, mitä informaatio varsinaisesti on ja mikä on esimerkiksi sen materiaalinen status, ei ole yhtäläistä jaettua käsitystä. (Adriaans 2020.) Informaatioteorian perustajana ja modernin informaation käsitteen luojana pidetään Claude Shannonia, joka julkaisi vuonna 1948 tutkimuksen ”The Mathematical Theory of Communication”. Shannon määritteli informaation puhtaan matemaattisesti: informaatio on tietyn tyyppisten merkkien relationaalinen ominaisuus, joka määrittyy ainoastaan merkkien tilastollisesta todennäköisyydestä. Shannonin teoriassa informaatio määritellään merkkien esiintymisen todennäköisyytenä, tarkemmin todennäköisyyden negatiivisena logaritmina. (Adriaans 2020; Niiniluoto 1988, 11–12; 30-35.) Tämä vastaa oletusta informaatiosta materiasta erillisenä abstraktiona.

Vaikka yhtäläistä informaation määritelmää ei ole kyetty muodostamaan, voidaan varteenotettavana teoriana tänä päivänä pitää Luciano Floridin muodostamaa informaation yleistä määritelmää (General Definition of Information, GDI). GDI:n mukaan informaatio on *hyvin muodostettua, merkityksellistä dataa* (Floridi 2010, 20–21). Määritelmä pitää sisällään kolme aspektia: (1) informaatio on olemukseltaan dataa, (2) datan on oltava hyvin muodostettua ja (3) datan on oltava merkityksellistä. Hyvin muodostetulla viitataan siihen, että datan on noudatettava jotakin syntaksia tai laajemmin sääntöjä. Merkityksellisyydellä viitataan siihen, että datalla on oltava jokin merkityssisältö kyseessä olevassa systeemissä tai kielessä. (Floridi 2010, 20–22.) Mutta sillä, mitä informaation konstituoiva tekijä eli data itsessään on, on ratkaiseva merkitys sille, mitä informaatio itse asiassa on. Niiniluoto (1988, 43) huomauttaa informaatiosta puhuttavan usein, niin kuin se olisi ainetta tai energiaa, mutta asia ei ole näin yksinkertainen.

Merkittäviä kysymyksiä liittyy edelleen siihen, mitä informaatio varsinaisesti on luonteeltaan. Onko se fysikaalinen suure, matemaattinen todennäköisyysfunktio vai jotakin muuta? Mikä on sen varsinainen ontologinen status, onko sillä väistämättä materiaalinen perusta, vai onko se sekä materiasta että energiasta irrallinen, oma ontologinen kategoriansa, kuten Norbert Wiener on ehdottanut? (Floridi 2011, 30–31; 42–43.) Muitakin avoimia kysymyksiä on, kuten se, mikä on informaation dynamiikka, onko informaatiolla totuusarvoa ja kuinka tämä totuusarvo on mahdollinen, mihin

merkityksellisen datan merkitys perustuu, sekä voidaanko luonto käsittää pohjimmiltaan informaationaalisena (Floridi 2011, 32–44). Tämän tutkielman kannalta olennaisia ovat ennen kaikkea kysymykset informaation ontologisesta statuksesta. Vaikka Floridin esittämä yleisnäkemys siitä, että informaatio koostuu datasta, ei itsessään kerro informaation perimmäisestä ontologisesta statuksesta, tarjoaa se kuitenkin selkeän kategorisoinnin sille, miten informaation ja datan käsitteet suhteutuvat toisiinsa. Allekirjoitan Floridin näkemyksen nimenomaan käsitteiden välisestä suhteesta ja hyödynnän sitä analyysissani.

Minkä vuoksi informaatio nimenomaan materiasta irrallisena kategoriana mahdollistaa ihmisten ja tietokoneiden samaistamisen? Tämä on Haylesin (1999) keskeisin väite: informaatio on ”menettänyt ruumiinsa” (lost its body), ja nimenomaan tämä ruumiin menettäminen on mahdollistanut ihmisen ja teknologian rajojen hälventymisen. Haylesin (1999) esittämässä kehityskulussa informaatio käsitetään ensin teoreettisena entiteettinä, sellaisena, joka on nimenomaisen aineetonta. Tämän jälkeen ihmisen hermojärjestelmä ja aivojen rakenne käsitetään informaatiota prosessoivina järjestelminä. Sitten luodaan artefakteja, jotka kääntävät informaatiovirrat havaittavissa oleviksi operaatioiksi. (Hayles 1999, 50–51.) Hayles (1999, 50) kutsuu tätä informaation *reifikaatioksi,* esineellistymiseksi tai vieraantumiseksi. Kyseessä on vastakkainasettelu informaation ja materiaalisuuden välillä, ja informaatio edustaa nimenomaan aineettomuutta.

Olennaista on *ymmärtää informaatio joksikin sellaiseksi, jonka prosessointi yhdistää sekä ihmisiä että tietokoneita*. Argumentin kannalta keskeinen on tämä *prosessoinnin toimenpide*. Informaatio esiintyy konseptina, jonka avulla analogia on mahdollista tehdä, eikä se itse analogiassa tule välttämättä määritellyksi tarkemmin. Informaation käsite mahdollistaa abstraktion tason, jonka avulla erilaiset entiteetit voidaan nähdä samankaltaisina. Dreyfus (1993, 166) toteaa suoraan, että kognitiivisen simulaation argumentit perustuvat pitkälti nimenomaan siihen, että informaation käsitettä käytetään epätarkasti. Se on konsepti, jonka avulla niin biologisten kuin keinotekoistenkin toimijoiden toiminta voidaan määritellä. Hayles (1999) osoittaa, kuinka Shannonin laskennallinen informaatioteoria syntyi tiettyyn ajalliseen kontekstiin, osana kyberneettistä tutkimusta, jossa pyrkimyksenä oli jo sinänsä ihmisen ja koneen rajojen hälventäminen. Ajatus informaatiosta puhtaan laskennallisena, merkityksestä ja materiasta irrallisena ominaisuutena, asettui palvelemaan näkemyksiä ja edistämään tulkintaa, jossa ihmiset ja koneet kuuluvat samaan kategoriaan.

Olennaista siinä, että informaatio on materiaalista erillistä, on mahdollisuus käsittää se olemukseltaan muodoksi tai kuvioksi (form, pattern). Teknologiaoptimisti Hans Moravec esittää tämän kuvaavasti: hänen mukaansa ihmisen identiteetti voi olla keho- tai muotoidentiteetti. Muotoidentiteetin omaava ymmärtää, että minuuden

olemus on materiasta irrallinen muoto ja prosessi, keho on lopulta ”pelkkää hyytelöä” (Moravec 1988). Muodon ja materian erottaminen toisistaan, sekä oletus materiasta riippumattomasta, abstraktista muodosta osana todellisuuden rakennetta, voidaan nähdä platonistisena ajatuksena. Informaation konseptista on tullut perimmäinen Platonin idea, kuten Hayles (1999, 13) toteaa.

Olemme todenneet, kuinka informaation käsittäminen materiasta irrallisena mahdollistaa sen käsittämisen abstraktina, kaikkea yhdistävänä muotona, jonka käsittelijöinä kaikki oliot voidaan määritellä. On lisäksi tarkennettava, että tässä käsityksessä data, josta informaatio koostuu, voidaan ymmärtää relationaalisena käsitteenä. Data ei perustu materiaaliseen pohjaan, vaan on viime kädessä kyse eroavaisuudesta muuttujien välillä (Floridi 2010, 23). Todellisuus on siis viime kädessä pelkistettävissä yksittäisiin erottuvuuksiin, esimerkiksi sähkösignaaliin ja sen puuttumiseen tai erottuvuuteen kahden värin välillä. Kyse on diskreeteistä elementeistä koostuvasta todellisuudesta – aivan kuten Dreyfus (1993) asian ontologisen olettaman käsitteellä ilmaisee. Todellisuus on dataa, josta voidaan prosessoida informaatiota, ja jonka prosessointi on kaikkia olioita yhdistävä tekijä.

Tämä on laajennus siihen, mitä Dreyfus (1993) esitti tekoälyn taustalla oleviksi sekä ontologisiksi että epistemologisiksi sitoumuksiksi. Informaatio abstraktina muotona, jonka käsittely yhdistää ihmisiä ja tietokoneita, tekee mahdolliseksi niiden samaistamisen. On kuitenkin tehtävä merkittävä huomio – oheinen ajatusrakennelma ei itsessään johda johtopäätökseen tekoälyn mahdollisuudesta. Tarkasteluni kohteena on nimenomaisesti ihmisen ja tietokoneen välinen samankaltaisuus. Mikäli informaatio oletetaan koko todellisuuden perustavanlaatuiseksi tekijäksi, Platonin idean kaltaiseksi, seuraisi siitä kaikkien olioiden samankaltaisuus. Tämä pitäisi sisällään paitsi ihmiset ja tietokoneet, myös muut koneet, kasvit, eläimet, elottoman luonnon ja niin edelleen. Johtopäätös voisi näin olla yhtä lailla se, että ihminen voidaan rinnastaa karhuun tai kolibakteeriin, koska molemmat prosessoivat informaatiota. Siihen, että ihmisen informaationprosessointikyky yhdistyy juuri tietokoneeseen ja on mahdollista erottaa ihmisen kehosta, liittyy muitakin oletuksia.

4.2 Formalisoinnista, hallittavuudesta ja älystä

Informaatio on oletettava materiasta irralliseksi kategoriaksi, jotta tietynlaisten analogioiden tekeminen erilaisten olioiden välillä on ylipäätänsä mahdollista, mutta nimenomaiseen analogiaan ihmisen ja tietokoneen välillä tarvitaan muutakin. Lisäksi tarvitaan vielä lisää oletuksia väitteeseen siitä, että ihmisen kaltaisen tekoälyn rakentaminen on mahdollista.

Olen tähän mennessä päätynyt siihen, kuinka informaatio on oletettava materiasta irralliseksi. Tähän voidaan suoraan lisätä Dreyfusin itsensä esittämä ontologinen olettama: informaatio on oletettava diskreeteistä elementeistä koostuvaksi ja määritellyksi. Nimeämme nämä diskreetit elementit dataksi, näin myös Dreyfus (1993, 206) itse tekee. Hän tosin ei käytä käsitettä johdonmukaisesti, vaan puhuu sen sijaan informaation ja datan käsitteistä ristiin. Pidän kuitenkin tässä lähtökohtanani Floridin (2010; 2011) näkemystä siitä, että informaatio koostuu datasta. Antamalla tälle datalle tietynlainen ontologinen status, voimme esittää tulkinnan informaation luonteesta. Kun data on pelkkiä erottuvuuksia, ei informaatiolla ole materiaalista perustaa.

Pyrin seuraavaksi osoittamaan, kuinka tarvitsemme lisäksi oletuksia formalisoinnin mahdollisuudesta, hallittavuudesta sekä älyn käsitteen sisällöstä. Jokainen näistä on oma, suuri filosofinen kysymyksensä, ja käyn ne läpi vain yleisellä tasolla. Teen tämän kuitenkin informaation käsitteen kautta. Osoitan, kuinka lisäämällä hieman lisää elementtejä käsitykseen informaation immateriaalisuudesta sekä käsitykseen informaation palautuvuudesta irrallisista elementeistä koostuvaan ja määriteltävissä olevaan dataan, voidaan muodostaa johdonmukainen käsitys paitsi ihmisen ja koneen analogisuudesta, myös konkreettisesta mahdollisuudesta rakentaa ihmistä vastaava tekoäly. Koska teen tämän informaation käsitettä hyödyntäen, myös tämän käsityksen haastamiseen tarjoutuu uusia elementtejä informaation käsitteen kautta.

Formalisoinnin mahdollisuus. Kysymys siitä, mitä ylipäätänsä voimme esittää formaalin järjestelmän puitteissa, ja mikä on formaalin järjestelmän suhde ontologiaan sinänsä, on laaja filosofinen kysymys (kts. Hofweber 2004). Voimme kuitenkin todeta, että nimenomaan oletus tiettyjen ilmiöiden formalisoitavuudesta on olennainen tekijä tekoälyoletuksessa. Tämä johtuu siitä, että ajatus yleisestä formalisoitavuudesta sinänsä on määritelmällisesti sama kuin ajatus tietokoneen laskentamahdollisuudesta. Tämä havainnollistuu Turingin koneen ideassa.

Formaali järjestelmä tarkoittaa järjestelmää, johon sisältyy kieli ja sen lauseenmuodostussäännöt, peruslauseet eli aksioomat, sekä päättelysäännöt (Raatikainen 1995, 189). Päättelysäännöt ovat laskettavia operaatioita lauseiden joukossa; on siis oltava mahdollista ratkaista mekaanisesti, onko oletettu päätelmä päättelysääntöjen mukainen (Raatikainen 2005). Mekaaninen ratkaiseminen on tässä tapauksessa olennaista. Ongelma on mekaanisesti ratkaistavissa, täsmälleen silloin, kuin se on ratkaistavissa Turing-koneella, joka on eräänlainen teoreettinen abstrahointi tietokoneen toimintaperiaatteesta. (Raatikainen 2005.) Toisin sanoen, formalisoitavissa oleva ilmiö on määritelmällisesti käsiteltävissä tietokoneella.

Edellä on todettu, kuinka informaation, jota ihmiset ja tietokoneet prosessoivat yhdenmukaisesti, on oltava materiasta erillistä. Tämän lisäksi itse prosessoinnin on

oltava esitettävissä formaalina järjestelmänä. Sen on oltava ilmaistavissa mekaanisesti laskettavien sääntöjen kokoelmana. Ei kuitenkaan riitä, että mikä tahansa käytös sinänsä on formalisoitavissa, kuten Dreyfus (1993, 195–196) huomauttaa – tällöin päätyisimme jälleen geneeriseen näkemykseen kaikkien olioiden samanlaisuudesta. Olennaista on yhdistää käsitys tietystä informaation luonteesta tähän käsitykseen tietystä prosessoinnin luonteesta. Tietokoneet käsittelevät, ja ihmiset samoin tässä analogiassa, informaatiota, joka on muodoltaan *representatiivista dataa* (kts. Dreyfus 1993, 196). Kyse on siis tosiasioita suoraan vastaavasta datasta, jota saadaan maailmasta sinänsä. Data ei ole tuotettua tai muodostettua, vaan käsitteen alkuperäisen latinankielisen merkityksen[2] mukaisesti annettua. Informaatio on paitsi materiasta erillistä, myös olemukseltaan irrallista dataa, joka on itsessään todellisuuden representaatio. Kyse ei siis ole ainoastaan esimerkiksi kehojen välittämistä fysikaalisista signaaleista, vaan olennaisella tavalla jostakin sellaisesta, joka on kuva todellisuudesta. Kyse on representationaalisesta datakäsityksestä (Leonelli 2019). Tämä idea on tunnistettavissa myös Dreyfusin (1993, xvii) muotoilemassa representationalismin käsitteessä, jolla hän kuvaa tekoälyidean taustalla olevaa käsitystä.

Hallittavuus. Ihmiset ja tietokoneet prosessoivat formaalien sääntöjen mukaisesti informaatiota, joka on materiasta erillistä, diskreettiä dataa, joka sellaisenaan representoi todellisuutta. Tämä oletus on pätevä peruste analogian muodostamiselle – ihmisiä ja koneita yhdistää jokin tekijä. Itsessään se ei kuitenkaan mahdollista sitä, että tekoäly olisi konkreettisesti rakennettavissa. Tarvitaan vielä oletus siitä, että formalisoitavissa oleva järjestelmä on merkittävällä tavalla myös teknisesti toteutettavissa. Tätä oletusta kuvaa Weizenbaumin (1976, 12) tekemä olennainen havainto. Länsimaisen filosofian kehitystä on hänen mukaansa ohjannut kysymys siitä, mitkä elämän aspektiti ovat formalisoitavissa. Pitkään kysymys tarkoitti sitä, mitä ja miten voimme tietää ihmisen mahdollisuuksista, vastuista ja rajoitteista. Tiedon saavuttaminen mahdollisti formaalin jäsennyksen tekemisen. Uuden teknologian myötä kysymys elämän formalisoitavuudesta on kuitenkin muuttunut kysymykseksi siitä, mihin teknologiseen sukuun ihmisen voidaan katsoa kuuluvan. (Weizenbaum 1976, 12.) Voidaan todeta, kuinka formalisointi ja teknologia katsottiin yhdeksi – formalisoitavissa oleva asia on määritelmällisesti teknologiaa.

Näin huomaamme, ettei tekoälyoletukseen sisälly oletuksia ainoastaan siitä, mitä ihminen tai hänen ajattelunsa on, tai mitä informaatio on. Olennainen on myös näkemys siitä, mikä kone on. Teknologian olemus on jälleen suuri ja kompleksinen filosofinen kysymys, johon paneutuminen olisi oma tehtävänsä. Tässä yhteydessä on keskeistä tunnistaa, kuinka teknologia tulee ymmärretyksi kaiken olevaisen oletettua abstraktia ylätasoa hallitsevaksi kategoriaksi. On mahdollista rakentaa kone, joka

[2] *datum* (lat): ”se mikä on annettu”

prosessoi informaatiota sinänsä sen abstraktissa muodossaan, sen sijaan että nähtäisiin sen prosessoivan esimerkiksi sähkösignaaleja. Oletus suorasta pääsystä tähän informaation tasoon, oletus siitä, että formaali järjestelmä voidaan implementoida sellaisenaan, on taustalla siinä ajatuksessa, että ihmiselle analogisen teknisen laitteen rakentaminen on todella mahdollista.

Älyn käsite. Viimeinen ratkaiseva oletus koskee älyn käsitteen sisältöä. Kuten luvussa kaksi todettiin, ajatus tekoälystä pitää sisällään ajatuksen siitä, että äly sinänsä on määriteltävissä, formalisoitavissa ja irrotettava sen materiaalisesta toteutumasta eli ihmisruumiista. Kun tuotetaan keinotekoinen älykkyys, mitä itse asiassa tuotetaan? On merkittävää havaita, kuinka älykkyyden käsite ei ole neutraali, vaan sillä on pitkä, muun muassa koloniaalisiin valtasuhteisiin kytkeytyvä historiansa (Adams 2021; Cave 2020). Tämän problematiikan selvittäminen on laaja tehtävä, joka on tämän tutkimuksen fokuksen ulkopuolella. Voimme kuitenkin tehdä aiheesta muutaman tärkeän noston.

Älykkyys on tullut modernissa tieteessä abstrahoiduksi yksittäiseksi, selkeäksi entiteetiksi. Se voidaan nähdä jopa tietyllä asteikolla mitattavana suureena, mistä älykkyysosamäärän konsepti kertoo. (Cave 2020; Weizenbaum 1976, 203–205.) Kun älykkyys oletetaan irrotettavaksi ja toiseen kohteeseen siirrettäväksi, oletetaan se samalla selkeärajaiseksi ja kontekstiriippumattomaksi. Toisin sanoen se oletetaan informaation prosessoinniksi. Toisin päin käännettynä, älykäs ihminen tai muu toimija on sellainen, joka hallitsee mahdollisimman tehokkaan loogisen päättelyn, sääntöpohjaisen informaation prosessoinnin, eikä sellaisille ominaisuuksille, jotka eivät ole formalisoitavissa, ole sijaa älykkyydessä. Abstrakti päättelykyky on korkein kognitiivinen kyky, ja se simuloimalla saataisiin aikaan älykkyys (Adams 2021).

Toinen huomio on kyseessä olevan älykkyyden konseptin irrallisuus paitsi materiaalisuudesta, myös kehollisuudesta. Irrotettava, kontekstista toiseen siirrettävä äly, jonka olemassaolo on irrallista materiaalisesta ja kehollisesta todellisuudesta, omalla tavallaan toisintaa länsimaisessa filosofiassa vallinnutta ajatusta ajattelun sfäärin irrallisuudesta (Hayles 1999, 195–196). On tosin huomioitava, ettei tekoälyn ajatuksen sinänsä tarvitse hylätä ajatusta kehosta, itse asiassa se voidaan jopa nähdä välttämättömänä edellytyksenä sille (Telakivi & Arstila 2021). On tosin eri asia todeta älykkyyden edellyttävän kehollista olemista sinänsä, kuin väittää sen voivan olemavan implementoitavissa materiaalisesta toteutumasta riippumatta, ja toteutuvan ikään kuin missä tahansa kehossa. Nämä ovat kuitenkin tämän tutkimuksen ulkopuolelle jääviä kysymyksiä.

4.3 Argumentin yhteenveto

Mitä meidän on oletettava informaatioon liittyen, jotta päädymme lopputulokseen ihmisen ja tietokoneen analogisuudesta ja siitä edelleen lopputulokseen ihmisen kaltaisen tekoälyn mahdollisuudesta? Luvun neljä alussa esittelemäni argumenttikaavio näyttää nyt tältä

P(1): Informaatio on luonteeltaan materiaalista irrallista
↓
P(2): Informaatio koostuu datasta, joka on diskreettejä, yksittäisiä elementtejä
↓
P(3): Informaatio muodostaa todellisuuden tason, jonka prosessoijina kaikki oliot voidaan ymmärtää
↓
P(4): Sekä ihmiset ja tietokoneet prosessoivat erityisenlaatuista informaatiota: representationaalista dataa joka se on sellaisenaan kuva todellisuudesta
↓
P(5): Sekä ihmisten että tietokoneiden harjoittama informaation prosessointi voidaan esittää formaalin järjestelmän puitteissa
↓
P(6): Ihmisten älykäs toiminta on pohjimmiltaan samanlaista ”informaation prosessointia” kuin se, mitä tietokoneet suorittavat
↓
P(7): Ihmiset ja tietokoneet jakavat sellaisen yhteisen ominaisuuden, että ihmisen toiminta on mielekkäällä tavalla rinnastettavissa tietokoneen toimintaan ja toisin päin
↓
J(1): Ihmisten ja tietokoneiden välillä vallitsee analogia
↓
P(8) Formaali järjestelmä on mahdollista toteuttaa konkreettisen teknologian avulla
↓
P(9) Ihmisen älykkyys on viime kädessä yksittäinen, mitattava suure joka on riippumaton kontekstistaan
↓
P(10) Ihmisen älykkyydestä voidaan tuottaa formaali ideaalimalli ja implementoida se tietokonejärjestelmään
↓
J(2): On mahdollista rakentaa ihmisen kaltainen tekoäly

Palatkaamme Dreyfusin alkuperäiseen väitteeseen. Hän esitti tekoälyoletuksen taustalla olevat neljä olettamaa: biologisen, psykologisen, epistemologisen sekä ontologisen. Tässä esittämäni jäsennys on pitkälti sama, mutta lähden liikkeelle informaation käsitteestä. Hyödynnän Floridin käsitteellistystä siitä, kuinka informaatio koostuu datasta, data on siis informaation elementtejä. Käsitys sitä, että ihmiset ja tietokoneet prosessoivat representationaalista dataa, on yhteensopiva Dreyfusin esittämän ontologisen olettaman kanssa: kaikki olemassaoleva voidaan analysoida joukkona loogisesti erillisiä faktoja.

Kuten luvussa 4.2 totesin, keskeistä tulkitsemassani kokonaisuudessa on sen kriittinen potentiaali. Dreyfus jäsentää perusteellisesti tekoälyoletuksen taustalla olevia, viime kädessä filosofisia ajatusrakennelmia nimenomaan siksi, että pyrkii väittämään niitä vastaan. Kyse on ihmisen ja tietokoneen analogisuudesta, missä Dreyfus keskittyy ensisijaisesti siihen, kuinka oletukset ihmisestä ovat virheellisiä. Ihmisen toiminta perustuu ymmärrykseen, joka syntyy vain ihmisenä olemisesta, ja johon väistämättä kuuluvat kehollisuus, vuorovaikutus materiaalisen todellisuuden kanssa sekä inhimillinen kulttuuri (Dreyfus 1993, 2–3). Lisäksi se on väistämättä situationaalista, sidoksissa tiettyyn kontekstiin, josta sitä on mahdotonta irrottaa (Dreyfus 1993, 273–282). Nämä ovat sinällään hyviä näkemyksiä, joiden avulla ihmisen ja tietokoneen analogian kumoaminen onnistuu. Niitä voidaan kuitenkin myös kritisoida. Tässä jäsentämäni informaation käsitteeseen perustuva näkökulma tarjoaa kuitenkin myös uusia välineitä Dreyfusin näkemyksen tueksi.

Voimme kyseenalaistaa oletuksen informaatiosta materiaalista irrallisena, ja todeta kuinka data vaatii väistämättä materiaalisen toteutuman ollakseen ylipäätänsä olemassa. Ei ole mitään aineetonta kerrosta, joka sellaisenaan siirtyisi materiaalisesta ruumiista toiseen (kts. Hayles 1999, 48–49). Voimme myös kyseenalaista datan representatiivisen luonteen eli sen, että se edustaisi todellisuutta sellaisenaan jossakin erillisessä kerroksessa, ikään kuin mallina todellisuudesta. Tämän sijasta data tulisi nähdä tuotettuna ja kontekstisidonnaisena, eri toimijoiden välisestä vuorovaikutuksesta syntyneenä (Leonelli 2019). Sille, ettei ihmisen mielen toiminta ole formalisoitavissa, on Dreyfus itse esittänyt hyvät perusteet. Viime kädessä Gödelin epätäydellisyysteoreema todistaa, ettei minkään järjestelmän totaalinen formalisointi ole mahdollista (Raatikainen 2005). Representaatioihin ja muotoihin perustuva järjestelmä jää väistämättä vajavaiseksi, ja sen avulla on määritelmällisesti mahdotonta toisintaa ihmisen älykkyys tai ajattelukyky – vaikka tämä kyky olisikin jollakin hypoteettisella tavalla irrotettavissa erilleen ihmisen ruumiista ja kontekstista.

Mielenkiintoinen lisähuomio on, kuinka edellä kumotut käsitykset kontekstiriippumattomasta, irralliseen muotoon perustuvasta sekä materiasta toiseen siirrettävästä älykkyydestä johtavat paitsi ideaan tekoälystä, myös vielä kyseenalaisempiin ideoihin. Hans Moravec (1988) menee niin pitkälle, että rinnastaa

koko ihmisen mielen dekontekstualisoituun informaatioon – mistä seuraa se looginen johtopäätös, ettei mieli ole sidoksissa tiettyyn materiaaliseen tai keholliseen toteutumaan. Koko ihmisen minuus on vain muoto tai kuvio. Näin ollen se voidaan myös ladata tietokoneelle, mikä voisi mahdollistaa biologisesta kehosta riippumattoman, lähes ikuisen elämän. (Moravec 1988.) Tämä oletus tosin menee vielä pidemmälle, kuin tekoälyoletus, jossa oletettiin voitavan simuloida ihmisen yhtä, näennäisesti irrallista ominaisuutta. Joka tapauksessa, Moravecin sinänsä sisäisesti johdonmukaisista argumenteista voidaan huomata, millaisia ajatuskulkuja oletus informaatiosta perustavana, abstraktina muotona mahdollistaa.

Voidaan todeta, kuinka ihminen ja tietokone eivät ainakaan tällä perusteella ole toisilleen analogisia, ja vaikka olisivatkin, keinotekoisen älykkyyden rakentaminen formalisoimalla ja toisintamalla symbolisen informaation prosessoinnin kyky ei onnistu. Näkemys ei kuitenkaan välttämättä ole riittävä. Dreyfus on saanut jo aikanaan osakseen kritiikkiä siitä, ettei hänen argumenttinsa ole välttämättä riittävä (kts. Weizenbaum 1976, 12–13). Tämän lisäksi tekoälyteknologian kehittyminen ja konnektionistinen paradigma ovat pitkälti syrjäyttäneet pyrkimyksen kehittää tekoäly edellä mainittua reittiä. Tämän myötä on avautunut uusia mahdollisuuksia ihmisen ja tietokoneen välisten analogioiden tekemiseen. Näin ollen esittämääni argumenttirakennelmaa, Dreyfusin omaa argumenttia ja sen informaation käsitteellä laajentamaani tulkintaa, voidaan kokonaisuudessaan kritisoida. Luvussa 5 luodaan lyhyt katsaus tähän kritiikkiin.

5 RAJOITTEET JA VAIHTOEHDOT

Olen todennut, mitä oletuksia ihmisen ja koneen välinen analogia edellyttää, ja kuinka nämä oletukset kumoamalla voidaan myös itse analogia kumota. Tämä ei kuitenkaan tarkoita, että analogia sinänsä tulisi kaikissa tapauksissa kumotuksi, eikä edes sitä, ettemmekö voisi olettaa ”ihmisen kaltaisen” tekoälyn olevan mahdollinen. Dreyfusin mallia vastaan on esitetty kritiikkiä sekä hänen aikanaan että nykyään – huolellinen ja perusteellinen jäsennys ei välttämättä sittenkään onnistu tavoitteessaan puolustaa ihmisen ainutlaatuisuutta koneisiin verrattuna.

Seuraavaksi kiinnitän huomiota kahteen erityyppiseen rajoitteeseen, jotka koskevat toisaalta Dreyfusin mallia sinänsä, toisaalta siitä esittämääni laajennettua tulkintaa. Ensimmäinen liittyy jo johdannossa mainitsemaani teknologian kehitykseen: tekoälyn kehittäminen ei ole enää pitkään aikaan tarkoittanut pyrkimystä formalisoida ihmisen älykkyys eksaktille symbolikielelle, vaan lähestymistapa perustuu hermoverkkoihin, minkä voidaan katsoa muuttaneen tilanteen merkittävällä tavalla. Toisaalta, muutokset eivät välttämättä ole niin merkittäviä kuin saatetaan esittää, ja juuri tämä voi altistaa vääristyneille argumenteille. Toinen rajoite liittyy muotoilemaani johtopäätökseen siitä, kuinka ihmisen ja tietokoneen analogian kannalta olennaista on käsittää informaatio nimenomaan aineettomana, abstraktina muotona. Osoitan, kuinka lähtökohta ei välttämättä ole esitetyllä tavalla erityinen.

Dreyfusin tekoälykritiikin lähtökohtana oli, kuinka älykkyys ei voi perustua symbolien manipulointiin tarkkojen sääntöjen mukaisesti. Klassinen ”vanha kunnon tekoäly” pyrki tutkimusprojektina toisintamaan irrallisen älykkyyden, joka ymmärrettiin nimenomaan representatiivisen datan käsittelynä formalisoitavissa olevien sääntöjen mukaisesti. Konnektionistinen lähestymistapa on kutienkin ratkaisevasti erilainen, sillä älykkyys ymmärretään siinä nimenomaisesti ei-symboliseksi. Älykkyys, tai ylipäätänsä monimutkaisten tehtävien suorittamisen kyky, on sen sijaan emergentti ilmiö, joka perustuu monikerroksisessa neuronien verkostossa tapahtuviin kytkeytymisiin (Bringsjord & Govindarajulu 2018).

Keinotekoinen neuroverkko ei prosessoi informaatiota samassa merkityksessä kuin eksplisiittisesti formalisoitava järjestelmä. Sen harjoittama ”prosessointi” muistuttaa itse asiassa nimenomaan sitä, millaisena Dreyfus itse kuvaili ihmisen ainutlaatuisen, kokemukseen perustuvan toiminnan: se perustuu kokemukseen ja on intuitiivista, ei-käsitteellistä ja ei sääntöihin pohjautuvaa (Niiniluoto 2021, 127). Kokemus viittaa tässä yhteydessä syväoppimiseen, jossa järjestelmä mukautuu käsittelemällä valtavan määrän dataa, joka muuttaa verkon tilaa ja kytkentöjä. Kyse ei kuitenkaan ole mistään käsitteellisestä; vaikka kyse on laskennallisista prosesseista, niillä ei ole representatiivista sisältöä. (Niiniluoto 2021, 119–120.)

Voidaan siis todeta, kuinka nykyisten, neuroverkkoihin perustuvien tekoälyjärjestelmien taustalla ei ole ainakaan kokonaisuudessaan samanlaista ajatusrakennelmaa kuin luvussa 4 esitettiin. Oletukset formalisoitavuudesta ja representationaalisuudesta putoavat neuroverkon myötä pois. On kuitenkin monimutkaisempi, ja tämän tutkimuksen ulkopuolelle jäävä kysymys, miten informaatio sinänsä tulee käsitetyksi neuroverkkojärjestelmässä. Myös neuroverkkoon perustuva tekoäly prosessoi jotakin, mutta tämä jokin on merkittävästi eri tyyppistä kuin klassisen tekoälyn tapauksessa. Tämän käsitteellistäminen on oma projektinsa. Voidaan kuitenkin huomauttaa, että käsitykset ovat kaikesta huolimatta todennäköisesti pitkälti yhtäläiset (kts. Aizawa 1992). Myöskään esimerkiksi kritiikkiä kontekstista irrotetun, atomisoidun älykkyyden olettamisesta neuroverkkoparadigma ei väistä.

Toinen olennainen rajoite liittyy johtopäätökseeni siitä, että informaatio on käsitettävä materiaalisesta irrallisena. Immateriaalinen informaatio voi toimia ”platonisena muotona” (Hayles 1999, 13), se voidaan nähdä kaikkea yhdistävänä todellisuuden tasona, jonka prosessoinnin kautta jokainen olio kuuluu itse asiassa samaan sukuun. Johtopäätöksenä tämä ei kuitenkaan ole välttämättä uniikki. Mikäli päädymme käsittämään kaikki oliot informaation käsitteen kautta, tämä ei edellytä sitä, että informaatio ymmärretään aineettomana muotona, vaan myös päinvastaisen perustavanlaatuisen materialismin kautta voidaan päätyä samaan lopputulokseen. Voidaan sanoa informaation olevan fundamentaalisesti fysikaalista, ja todeta sen nimenomaan siksi olevan konstitutiivinen tekijä materiaalisessa universumissa – aina alkeishiukkastasolle saakka (Bawden 2013). Ei tarvitse olettaa materiasta erillistä todellisuuden tasoa, jonkinlaista muotojen ja ideoiden maailmaa, jotta informaatio voidaan määritellä perustavanlaatuiseksi tekijäksi. On myös mahdollista olettaa fundamentaalisesti materiasta koostuva, fysikaalinen todellisuus, joka elementaarisella tasollaan perustuu informaatioon (Bawden 2013).

Tämä ei kuitenkaan tee koko käsittelemääni argumenttia tyhjäksi, sillä kaikkea yhdistävä informaatio ei ylipäätänsä ole riittävä käsitys tarkastelemalleni ihmisen ja tietokoneen analogialle. Se, millaisten käsitteiden kautta ymmärrämme kaikki oliot, ja millaisena näemme todellisuuden perustavanlaatuiset rakenteet, voi itse asiassa olla toissijaista sen kannalta, miten ymmärrämme nimenomaan ihmiset ja tietokoneet. Tietyt näkökulmat voivat joka tapauksessa johtaa vahvempiin johtopäätöksiin kuin toiset, ja filosofisella keskustelulla tekoälystä on viime kädessä metafyysinen ulottuvuutensa.

Voidaan kuitenkin todeta niin Dreyfusin argumentin kuin siitä esittämäni laajennetun tulkinnan olevan sinänsä pätevä, mutta kuitenkin riittämätön. Siinä esitettyjen oletusten kumoaminen ei takaa sitä, ettei johtopäätökseen ihmisen ja tietokoneen analogisuudesta voitaisi silti päätyä. Sen käsitteellistäminen, mitä

neuroverkon ei-käsitteellisellä tasolla tapahtuu, jää myös omaksi filosofiseksi ongelmakseen. Mutta kuten Niiniluoto (2021, 127) huomauttaa, neuroverkkojärjestelmälle ei silti tarvitse olettaa ymmärrystä tai älykkyyttä. Sen toiminta on esikäsitteellistä, ei-propositionaalista ja tulee sellaisena lähelle Dreyfusin luonnehtimaa ihmisen ainutlaatuista toimintaa, mutta hyppäys siihen, että voitaisiin puhua mielestä tai älystä, on edelleen suuri. Myöskään kokonaisen mielen siirtämistä tietokoneelle se ei sellaisenaan mahdollista.

6 JOHTOPÄÄTÖKSET

Lähdin tutkielmassani tarkastelemaan ihmisen ja tietokoneen eroa ja yhtäläisyyttä jäsentävää argumentaatiota. Pyrin Dreyfusia seuraten jäsentämään, mitä oletuksia vaaditaan, jotta voidaan väittää ihmisen ja tietokoneen olevan toisilleen analogisia. Tarkastelin tätä informaation käsitteeseen keskittyen: millaiset käsitykset informaatiosta johtavat siihen, että ihmiset ja tietokoneet voidaan nähdä olennaisesti samanlaisina. Kytkin keskustelun osaksi tekoälyn ajatuksen historiaa nimenomaan laajassa merkityksessään, eli sen ajatuksen historiaa, että voimme ylipäätänsä rakentaa ihmisen älykkyyttä vastaavan keinotekoisen järjestelmän. Nämä käsitykset voidaan tiivistää siihen, kuinka *ihmiset ja tietokoneet prosessoivat formaalien sääntöjen mukaisesti informaatiota, joka on materiasta erillistä, diskreettiä dataa, joka sellaisenaan representoi todellisuutta.*

Muodostin jäsennyksen, jossa on ensiksi oletettava informaation luonne materiasta irrallisena todellisuuden tasona, eräänlaisena kaikkia olioita yhdistävänä muotona. Tämä mahdollistaa sen, että voimme tehdä analogian kaikkien olioiden välille, niin ihmisten, eläinten kuin koneidenkin. Ihmisen ja tietokoneen välille muodostava analogia tarvitsee kuitenkin tarkempia määritteitä. Olennainen on väite siitä, että ihmiset ja tietokoneet prosessoivat informaatiota samalla tavalla. Keskeinen on prosessoinnin toimenpide, ja kysymme, millaista tämän informaation on silloin oltava. Totesin, kuinka sen on oltava diskreettiä, eli irrallisista yksittäisistä elementeistä muodostunutta dataa, mikä sinällään on laajasti hyväksytty näkemys informaation luonteesta. Tämän datan on kuitenkin oltava myös representationaalista, eli kuvattava todellisuutta sellaisenaan. Kyse on maailmasta sinänsä saatavista tosiasioista, jotka viime kädessä edustavat todellisuutta sinänsä sen sijaan että olisivat esimerkiksi vain sähkösignaalien eroja.

Lisäksi totesin, kuinka sekä ihmisten että tietokoneiden harjoittama informaation prosessointi on käsitettävä formaalin järjestelmän puitteissa, joka on ilmaistavissa mekaanisesti laskettavien sääntöjen kokoelmana. Jotta varsinainen keinotekoinen älykkyys voitaisiin rakentaa, tämän formaalin järjestelmän on oltava teknisesti toteutettavissa. Lopulta, jotta voimme nimetä sitä tekoälyksi, nimenomaan keinotekoiseksi älykkyydeksi, on älykkyys itsessään oletettava kontekstista irralliseksi, tarkkarajaiseksi kokonaisuudeksi, joka on irrallaan materiaalisuudesta ja kehollisuudesta. On toisin sanoen ajateltava universaali älyn malli ja mahdollinen universaali-ihminen, jolle tämä kuuluu ja jonka älyä koneeseen ollaan implementoimassa

Näin ollen johtopäätökseni on, että käsitys materiasta irrallisesta, erillisistä elementeistä koostuvasta ja representatiivisesta informaatiosta, sekä tämän informaation prosessoinnista formaalien sääntöjen mukaisesti, mahdollistaa ihmisen

ja tietokoneen välille tehtävän analogian. Ihmisen mieli on käsitettävä formaalien sääntöjen mukaisesti toimivaksi systeemiksi ja ihmisen älykkyys abstraktiksi symbolien manipuloinniksi, joka on ensinnäkin tarkasti määriteltävissä ja toisekseen sellaisenaan implementoitavissa toiseen kohteeseen. On sitouduttava viime kädessä idealistiseen todellisuuskäsitykseen, jossa ajattelu ja älykkyys eivät ole sidoksissa ruumiiseen ja materiaan, vaan voivat sen sijaan siirtyä vapaasti sellaisenaan materiaalisesta toteutumasta toiseen.

Kokonaisuudessaan tässä käsityksessä on monia kohtia, joita vastaan voimme argumentoida, ja osoittaa siten, ettei keinotekoinen älykkyys ole mahdollinen. Dreyfusin oma tarkastelu, samoin kuin yllä esitetty, keskittyi ennen kaikkea siihen, kuinka se ei ole mahdollinen klassisella, symbolisella sääntöpohjaisella järjestelmällä. Konnektionismi, toisin sanoen keinotekoisten hermoverkkojen rakentaminen, on mutkistanut tilanteen. Esittämäni oletukseen sisältyvät väitteet siitä, millaista informaation prosessoinnin olisi oltava, on helppo kumota ja todeta, kuinka ne eivät vastaa ihmisen toimintaa. Kuitenkin myös konnektionistiset järjestelmät kumoavat samat oletukset, ja osoittavat, kuinka keinotekoisen ”älyn” rakentaminen on silti mahdollista.

Tulkintani muodostaa pohjan, jota voidaan edelleen tarkentaa. En varsinaisesti keskittynyt siihen, kuinka Dreyfusin näkemystä voidaan täydentää nimenomaan konnektionismin näkökulmasta, vaan siihen, kuinka sitä ylipäätänsä voidaan täydentää. Hyödynsin informaation käsitettä, mutta vielä tarkempi informaation luonteen analysointi voi tuoda uusia näkökulmia. Väitän, että näihin suuntiin lähtemällä meille tarjoutuu lopulta relevantteja mahdollisuuksia suhtautua kriittisesti myös keinotekoisen hermoverkon avulla tuotettuun ”älykkyyteen”. Viime kädessä voimme todeta, kuinka älykkyyden haltuun ottaminen kontekstistaan irrallisena ominaisuutena on mahdotonta, ellemme tarkoita älykkyydellä jotakin hyvin rajattua toimintoa, jolloin kyseessä ei ole enää ihmisen älyn jäljittely.

Kontekstisidonnaisuuden argumentti ei kuitenkaan kumoa sinällään ideaa ihmisen kaltaisesta tekoälystä, sillä keho ja konteksti on myös mahdollista rakentaa keinotekoisesti, eräänlaisena robottiruumiina (kts. Telakivi & Arstila 2021). Ihmisen ja koneen välisen kompleksisen suhteen käsitteellistämiseen tarvitaan lopulta nimenomaan informaation konseptia. On tarkennettava sen ajatuksen kritiikkiä, että informaatio olisi mahdollista dekontekstualisoida ja siirtää materiaalisesta toteutumasta toiseen, siis että se olisi materiasta irrallinen todellisuuden taso. Mikäli tätä ei hyväksytä, ei keinotekoinen järjestelmä vastaa koskaan ihmistä, vaikka sillä olisi ruumis ja vaikka se ilmaisi älykkyyden tai jopa tunteiden kaltaisia asioita. Se yksinkertaisesti perustuu fundamentaalisesti erilaiseen materiaaliseen alustaan.

Tämän tarkempi pohdiskelu on muiden tutkimusten aihe. Toisaalta, on pohdittava myös sitä, millä perusteella haluamme toisaalta väittää koneen voivan

muistuttaa ihmistä, sekä toisaalta, millä perusteella haluamme kieltää sen. Viime kädessä kysymys siitä, onko ihmisen kaltainen keinotekoinen älykkyys mahdollinen, ei välttämättä ole relevantti, vaikka onkin filosofisesti kiinnostava. Emme myöskään pääse eroon kysymyksenasetteluun väistämättä sisältyvästä ongelmallisuudesta, nimittäin oletuksesta tietynlaisesta universaalista ihmisestä.

Huomionarvoista on, että kaikista näistä pohdinnoista voidaan esittää vielä erilaisia johtopäätöksiä. Totesin kysymyksen ihmisen ja tietokoneen välisestä analogiasta olevan yhteiskunnallisesti keskeinen siksi, koska teknologia kehittyy ja tietokonejärjestelmät vaikuttavat pystyvän yhä ihmismäisempiin toimintoihin. Weizenbaum (1976) otti kuitenkin jo 1970-luvulla erilaisen kannan. Hänen mukaansa se, vaikka tietokoneet olisivatkin muistuttaneet ihmisiä kaikilla mahdollisilla osa-alueilla, ei oikeuttanut yhteiskunnallisiin johtopäätöksiin. Keskeinen kysymys ei hänen mukaansa ollut, mihin tietokoneet pystyvät, vaan mitä tietokoneiden tulisi tehdä ja mitä ei (Weizenbaum 1976, 12–13). Tämä oli ja on edelleen moraalinen kysymys, joka voidaan asettaa irralleen siitä kysymyksestä, mitä tietokoneet tosiasiallisesti pystyvät tekemään. Voidaan toki kuitenkin kysyä miksi, jos keinotekoiset järjestelmät todella muistuttaisivat ihmisiä, tai ainakin sitä minkä oletamme ihmiseksi, riittävän suurissa määrin. Myös tämän spekulointi on erillisen tutkimuksen aihe.

LÄHTEET

Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews, 46*(1–2), 176–197. https://doi.org/10.1080/03080188.2020.1840225

Adriaans, P. (2020). "Information", *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (toim.), https://plato.stanford.edu/archives/fall2020/entries/information/

Aizawa, K. (1992). Connectionism and artificial intelligence: History and philosophical interpretation. *Journal of Experimental & Theoretical Artificial Intelligence 4*(4), 295-313. https://doi.org/10.1080/09528139208953753

Bawden, D. (2013). "Deep Down Things": In What Ways is Information Physical, and Why Does it Matter for Information Science? *Information Research*, *18*(3), Special.

Bongard, J., & Levin, M. (2021). Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior. *Frontiers in Ecology and Evolution, 9,* 650726. https://doi.org/10.3389/fevo.2021.650726

Bringsjord, S. & Govindarajulu, N. S. (2018). "Artificial Intelligence", *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (toim.), https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/

Cave, S. (2020). The Problem with Intelligence: Its Value-Laden History and the Future of AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 29–35. https://doi.org/10.1145/3375627.3375813

Dreyfus, H. L. (1993). *What computers still can't do: A critique of artificial reason* (3th pr). MIT Press.

Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* Free Press.

Esposito, M., & Baravalle, L. (2023). The machine-organism relation revisited. *History and Philosophy of the Life Sciences*, *45*(3), 34. https://doi.org/10.1007/s40656-023-00587-2

Floridi, L. (2010). *Information. A Very Short Introduction.* Oxford University Press.

Floridi, L. (2011). *Philosophy of Information.* Oxford University Press.

Haaparanta, L. (1995). Älykäsitykset ja moderni logiikka. Teoksessa Haaparanta, L. (toim.), *Älyn ulottuvuudet ja oppihistoria: Matka logiikan, psykologian ja tekoälyn juurille*. Suomen tekoälyseura.

Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Hofweber, T. (2004). "Logic and Ontology". *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (toim.), https://plato.stanford.edu/archives/sum2023/entries/logic-ontology/

Kenaw, S. (2008). Hubert L. Dreyfus's Critique of Classical AI and its Rationalist Assumptions. *Minds and Machines*, *18*(2), 227–238. https://doi.org/10.1007/s11023-008-9093-7

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science*, *9*(2), 22. https://doi.org/10.1007/s13194-018-0246-0

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine, 27*(4), 12.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Negrotti, M. (2019). Hubert Dreyfus, the artificial and the perspective of a doubled philosophy. *AI & Society*, *34*(2), 195–201. https://doi.org/10.1007/s00146-018-0800-5

Niiniluoto, I. (1988). *Informaatio, tieto ja yhteiskunta. Filosofinen käsiteanalyysi*. Valtiohallinnon kehittämiskeskus.

Niiniluoto, I. (2021). Syväoppimisen filosofiaa. Teoksessa Raatikainen, P. (toim.), (2021). *Tekoäly, ihminen ja yhteiskunta: Filosofisia näkökulmia*. Gaudeamus.

Pesonen, R. (2021). "Tekoäly". Logos-ensyklopedia. Eurooppalaisen filosofian seura ry. Verkossa https://filosofia.fi/fi/ensyklopedia/tekoaly

Raatikainen, P. (1995). Laskettavuuden teorian varhaishistoriaa. Teoksessa Haaparanta, L. (toim.), *Älyn ulottuvuudet ja oppihistoria: Matka logiikan, psykologian ja tekoälyn juurille*. Suomen tekoälyseura.

Raatikainen, P. (2005). Formalismin rajat. *Niin & näin, 44*(2), 31–35.

Raatikainen, P. (2021). Tekoäly, ihminen ja yhteiskunta – johdatusta teemaan. Teoksessa Raatikainen, P. (toim.), *Tekoäly, ihminen ja yhteiskunta: Filosofisia näkökulmia*. Gaudeamus.

Stoutland, F. (1995). Connectionism and philosophy of action. Teoksessa Haaparanta, L. & Heinämaa, S. (toim.), *Mind and Cognition. Philosophical Perspectives on Cognitive Science and Artificial Intelligence.* Acta Philosophica Fennica Vol. 58.

Su, Bc., Luvaanjalba, B. (2021). The Effect of Hubert Dreyfus's Epistemological Assumption on the Philosophy of Artificial Intelligence. Teoksessa Nah, F.FH., Siau, K. (toim.), *HCI in Business, Government and Organizations. HCII 2021. Lecture Notes in Computer Science, vol 12783*. Springer, Cham. https://doi.org/10.1007/978-3-030-77750-0_42

Telakivi, P. & Arstila, V. (2021). Onko mahdollista rakentaa keinotekoinen mieli? Teoksessa Raatikainen, P. (toim.), *Tekoäly, ihminen ja yhteiskunta: Filosofisia näkökulmia*. Gaudeamus.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman.