TIETOISUUS TEKOÄLYSYSTEEMEISSÄ

TURUN YLIOPISTO
Informaatioteknologian laitos
Tietojenkäsittelytiede
Pro gradu -tutkielma
31.05.2008
Jussi Antila

TURUN YLIOPISTO
Informaatioteknologian laitos

ANTILA, JUSSI Tietoisuus tekoälysysteemeissä

Pro gradu -tutkielma, 99 s.
Tietojenkäsittelytiede
Toukokuu 2008

Tietokoneiden teho on kasvanut eksponentiaalisesti siitä lähtien, kun ensimmäiset elektroniset tietokoneet rakennettiin. Eikä tähän ole odotettavissa muutosta. Koneista on tullut niin kykeneviä, että ne korvaavat ihmisen jo monissa asioissa. Tietokoneet itseasiassa päihittävät ihmisen monissa tehtävissä, jotka ovat aikaisemmin oletettu sellaisiksi, että vain ihminen kykenee ne ratkaisemaan. Sellaisessakin älykkyyttä vaativassa lajissa kuin šakkipelissä, on tietokone kyennyt voittamaan jopa lajin maailmanmestarin.

Mutta mitä tuo tietokoneiden älykkyys on? Voidaanko sitä verrata ihmisälyyn? Erityisesti, voivatko koneet olla tietoisia? Tieteistarinat ovat jo pitkään kertoneet meille tietoisista koneista, jotka saattavat jopa syrjäyttää ihmislajin maapallolta. Mutta voiko tekoälytutkimus todella tuottaa koneita, jotka aidosti kykenevät samankaltaiseen tietoisuuteen, mikä ihmisillä on?

Käsittelen tutkielmassa juuri näitä kysymyksiä. Aihe on hyvin filosofinen. Näiden kysymysten ratkaisemiseksi täytyy perehtyä hyvin perustavanlaatuisiin kysymyksiin maailman, tietoisuuden, ja laskennallisuuden luonteesta. Syvennynkin tutkielmassa nimenomaan tekoälyn filosofiseen puoleen, jättäen tekniset kysymykset vähemmälle huomiolle.

Komputationalismi on filosofian näkemys, jonka mukaan tietoisuus on perimmäiseltä luonteeltaan laskentaa. Tämän näkemyksen mukaan siis ihmisaivotkin suorittavat pohjimmiltaan vain laskentaa. Aivot ovat vain monimutkainen tietokone. Tutkielman tutkimuskysymys voidaankin asettaa näin: onko komputationalismi tosi? Näkemyksen puolesta ja sitä vastaan on 1900-luvun puolesta välistä asti esitetty kirjallisuudessa ja tieteellisissä artikkeleissa hyvin paljon erilaisia argumentteja. Käsittelen tutkielmassa tärkeimpiä näistä argumenteista. Komputationalismilta on kysytty erityisesti sitä, miten tietokone voi saada tietoa merkityksistä. Tietokoneidenhan ajatellaan yleensä vain manipuloivan täysin merkityksettömiä bittejä: ykkösiä ja nollia. Miten näistä merkityksettömistä symboleista voi syntyä aitoja merkityksiä?

Tutkimuskysymyksen ratkaisu riippuu hyvin paljon siitä, millaisia metafyysisiä oletuksia kysyjä on valmis hyväksymään. Tulen kuitenkin osoittamaan, että on hyviä perusteita sen puolesta, että komputationalismi todella on tosi, riippumatta oletuksista.

Avainsanat: tekoäly, tekoälyn filosofia, komputationalismi, funktionalismi, Turingin testi, kiinalainen huone

UNIVERSITY OF TURKU
Department of Information Technology

ANTILA, JUSSI Consciousness in Artificial Intelligence Systems

Master's Thesis, 99 p.
Computer Science
May 2008

Power of computers has grown exponentially ever since the first electronic computers were built. And there is no reason to suppose that this trend will end. Machines have become so capable that they can substitute human in many tasks. In fact, computers actually beat human in many tasks that were previously thought to be solvable only by human. Even in such intelligence-requiring game as chess, a computer has managed to beat even the world champion.

But what is computer intelligence? Is it comparable to human intelligence? In particular, can machines be conscious? Science fiction has long depicted conscious machines, which could even depose the human race from planet earth. But can artificial intelligence research really produce machines which could have the same kind of consciousness as human beings have?

In this thesis I will study precisely these questions. The topic is very philosophical. To answer these questions, we have to explore very fundamental questions about the nature of the world, consciousness and computability. I will focus especially on the philosophical aspects of artificial intelligence and therefore I will not pay much attention to technical questions.

Computationalism is a philosophical view, which asserts that consciousness is essentially just computation. So, in this view, even human brain essentially just computes. Human brain is just a complex computer. The research question of this thesis can be presented as follows: Is computationalism true? Since mid-1900s many arguments have been published for and against computationalism in literature and in scientific articles. I will discuss the most important ones of these arguments in this thesis. For example, computationalism has been required to answer the question how a computer can become conscious about meanings. Computers are usually thought only to manipulate entirely meaningless bits: ones and zeros. How can real meanings emerge from these meaningless symbols?

The answer to these questions depends much on what kinds of metaphysical assumptions the inquirer is ready to accept. Nonetheless, I will argue that there is plenty of evidence to support computationalism, regardless of one's assumptions.

Keywords: artificial intelligence, philosophy of artificial intelligence,
computationalism, functionalism, Turing test, Chinese room

SISÄLLYS

1 JOHDANTO

1.1 Motivaatio

Jos täytyisi nimetä yksi tekijä, joka on muuttanut ihmiskuntaa eniten 1900-luvulla ja sen jälkeen, olisi yksi varteenotettava vaihtoehto ehdottomasti tietokoneiden ja informaatioyhteiskunnan synty ja kehitys. Tietokoneiden huiman kehityksen myötä niistä on tullut välttämätön työväline jokapäiväisten asioiden hoitamiseen, eikä yhteiskunta tulisi enää toimeen ilman niitä. Mutta miksi tietokoneista on tullut välttämättömiä? Mikä on se ominaisuus, joka tekee niistä niin hyödyllisiä? Se on niiden ylivertainen kyky käsitellä, hallita, välittää ja tallentaa informaatiota. Kuten paperi ja kirjoitustaito aikanaan mullistivat ihmisten tiedonvälityskyvyn, saman on tehnyt tietokone vielä paljon suuremmassa mittakaavassa. Tietokoneiden myötä ihminen pystyy tuottamaan ja hallitsemaan huomattavasti enemmän informaatiota, kuin aikana ennen tietokoneita. Koneista on tullut ikäänkuin jatke tai laajennus ihmisen älykkyydelle.

Hyvin pian tietokoneiden ilmestymisen jälkeen huomattiin tietty yhteys, joka vallitsee ihmisaivojen ja tietokoneen välillä: molemmat ovat jollakin tapaa kykeneviä käsittelemään informaatiota. Nousi esiin ajatus, että tämä yhteys ei olekaan ainoastaan näennäistä tai sattumaa. Ehkä taustalla onkin se, että ihmisten ja koneiden älykkyys syntyy jostakin hyvin samankaltaisesta ilmiöstä. Syntyi hypoteesi siitä, että ihmisaivot ja tietokone toimivat itseasiassa täysin samalla periaatteella. Molempien perimmäiseen luonteeseen kuuluu informaation käsitteleminen ja juuri tämä ominaisuus mahdollistaa sen, että tietokoneet voivat toimia jatkeena ihmisen älylle. Ihmisaivot ovat siis hypoteesin perusteella tietokone, vain sillä erotuksella perinteisiin tietokoneisiin, että aivot ovat orgaanista materiaa, kun taas tietokoneiden prosessorit ovat piistä tai muusta ei-orgaanisesta materiasta valmistettuja. Ajatuksesta tuli hyvin menestyksekäs laajoissa tieteellisissä piireissä, erityisesti tietojenkäsittelijöiden ja muiden luonnontieteilijöiden keskuudessa.

Pian kantautui kuitenkin myös useiden epäilijöiden ääni. He halusivat kumota aivot-tietokone -analogian pätevyyden (esim. Dreyfus, 1965). Heidän mielestään kone ei

koskaan voisi olla tietoinen samalla tavoin kuin ihminen. Vastustajat eivät välttämättä halunneet kumota sitä ajatusta, että aivojen ja tietokoneen välillä vallitsee jonkinlainen yhteneväisyys. Jotkut vastustajat saattoivat jopa hyväksyä, että tietokone saattaa joskus tulevaisuudessa olla yhtä älykäs kuin ihminen siinä mielessä, että se pystyy käyttäytymään täsmälleen samalla tavoin kuin ihminenkin. Toisin sanoen, että kone voisi jonakin päivänä kyetä imitoimaan ihmistä. Tärkeäksi käsitteeksi nousi kuitenkin tietoisuus. Tietoisuus haluttiin pitää erillään älykkyydestä. Se koettiin joksikin, joka on riippumatonta ulkoisesta käyttäytymisestä. Tietoisuus on jollain tapaa enemmän kuin pelkkää mekaanista sääntöjen noudattamista. Tietokoneethan ajatellaan yleensä ainoastaan automaateiksi, jotka noudattavat formaaleja sääntöjä, ilman että ne kokisivat tai ymmärtäisivät mitään siitä mitä ne tekevät. Ihminen sen sijaan tietoisena olentona kykenee tuntemaan, aistimaan, ymmärtämään ja ajattelemaan. Kun esimerkiksi katselen sinistä taivasta, tai maistan kuumaa kaakaota, liittyy näihin kokemuksiin aina erityinen tuntu siitä, millainen tämä kokemus on. Kun valonsäde saapuu silmiemme kautta aistinelimiimme ja neurologisten prosessiemme työstettäväksi, se ei ole ainoastaan mekaaninen prosessi, vaan tämä tapahtuma aiheuttaa minussa myös tietynlaisen elämyksellisen ilmiön. Tekoälyn vastustajien puolelta tuotiin esiin useita argumentteja, jotka yrittivät kumota näkemyksen, että myös tietokoneella voisi olla vastaavia elämyksellisiä ilmiöitä.

Väittely eri osapuolten kesken on jatkunut yli puoli vuosisataa, eikä loppua näy. Saattaa olla, että jopa lähitulevaisuudessa kehitetään robotteja, jotka kykenevät kaikkeen samaan kuin ihminenkin (Kurzweil, 2005). Ne voivat kyetä keskustelemaan ihmisten kanssa siten, että ihmiset eivät huomaa keskustelevansa robotin kanssa. Täytyykö meidän olettaa, että robotti on tällöin tietoinen samalla tavoin kuin ihminen? Lakkaako väittely koneiden tietoisuudesta silloin? Tai voidaanko kehittää jokin testi, jolla voimme selvittää onko jokin asia tietoinen? Tieteistarinat ovat jo jonkin aikaa valmistelleet meitä kohtaamaan sen hetken, jolloin joudumme oikeasti tähän tilanteeseen. Joudumme tulevaisuudessa todennäköisesti pohtimaan esimerkiksi moraalikysymyksiä robottien suhteen. Mutta onko meillä oikeasti keinoja ratkaista niitä?

Kysymys koneiden tietoisuudesta ei ole helppo, se on hyvin kaukana siitä. Sen ratkaisemiseen ei tunnu olevan empiirisiä keinoja. Vastauksen löytämiseksi täytyy

lähteä liikkeellä hyvin matalalta tasolta, perinteisten filosofisten kysymysten parista. Kysymys heijastelee ikivanhoja mielenfilosofian ongelmia, kuten mieli-ruumis -ongelmaa, sekä muitten mielten ongelmaa. Täytyy pystyä määrittelemään tarkasti olennaisia käsitteitä, kuten tietoisuus, tietokone, komputaatio jne. Kysymys koneiden tietoisuudesta on noussut ehkä tärkeimmäksi kysymykseksi, mikä nykyfilosofialla on ratkaistavanaan.

Tässä tutkielmassa käsitellään juuri tätä kysymystä ja yritetään löytää vastauksia siihen, voivatko koneet joskus todella ajatella ja olla tietoisia, kuten me ihmiset. Entä voidaanko määritellä joitain vaatimuksia, jotka koneen täytyy toteuttaa, ennen kuin tietoisuus voisi syntyä? Tekoälyn filosofiasta on kirjoitettu hyvin paljon viime vuosikymmeninä. Argumentteja sekä puolesta että vastaan on esitetty hyvin paljon.

Voidaanko koneiden tietoisuudesta antaa mitään lopullista vastausta? Se riippuu hyvin pitkälti siitä millaisia metafyysisiä oletuksia kysyjä on valmis hyväksymään. Näiden perusteella voidaan antaa hyvinkin suoraviivaisia vastauksia. Kaikkia tyydyttävää vastausta sen sijaan voi olla vaikeampi antaa. Sen lisäksi, että tutkielmassa määritellään tärkeät tekoälyyn liittyvät käsitteet, täytyy meidän myös käydä läpi millaisia filosofisia oletuksia ja perusteita on näiden käsitteiden taustalla. Tutkimalla tekoälyn kannalta tärkeimpiä filosofisia teorioita saamme käsityksen myös näistä asioista. Samalla tulee selkeämmäksi mitkä näistä teorioista ovat yhteensopivia konetietoisuuden kanssa. Yritän kehittää uskottavimpien teorioiden pohjalta myös omaa näkemystäni aiheeseen.

Yksi olennaisimmista kysymyksistä tekoälyn ratkaisemiseen on kysymys merkityksistä. Toisin sanoen kysymys siitä, miten semantiikka syntyy. Merkitysten ymmärtämisen ajatellaan olevan välttämätöntä ihmisen tietoisuudelle. Ihminen ei voi ymmärtää esimerkiksi puhettaan, jollei hän kykene ymmärtämään mitä hänen käyttämänsä sanat merkitsevät, tai mihin ne viittaavat. Tietokoneitten sen sijaan ajatellaan usein vain käsittelevän merkityksettömiä symboleita, ykkösiä ja nollia. Miten näiden yksinkertaisten bittien manipulaatiosta voi syntyä jotain merkityksellistä. Miten syntaksista voi syntyä semantiikka? Tämä on yksi avainkysymyksistä, jos halutaan ratkaista kysymys koneitten tietoisuudesta. Tätä tullaankin käsittelemään useaan otteeseen.

1.2 Mitä on tekoäly?

Käsitteellä tekoäly on monia merkityksiä. Kun yliopistolla käydään kursseja tekoälystä, niiden päällimmäisenä tavoitteena ei ole välttämättä opettaa miten luodaan tietoinen kone, tietoinen siinä mielessä, miten ihminen koetaan tietoisena. Tavoitteena on ennemminkin opettaa, miten koneeseen saadaan luotua tietty rationaalisuus. Siis jonkinlainen järkevyys ja loogisuus. Pyritään löytämään keinoja, jolla kone saadaan ratkaisemaan tietty ongelma tällä tavoin. Tämä on ehkä yleisin ja tieteellinen käsitys tekoälystä. Tekoälyprojektissa yritetään siis kehittää rationaalisesti käyttäytyvä kone (Russell & Norvig, 1995, 5).

1.2.1 Perinteinen tekoäly vs. kognitivismi

Ne tavat, joilla koneille saadaan luotua tämä rationaalisuus, voivat olla kuitenkin hyvin erilaisia. Yksi tapa saada kone ratkaisemaan tietty ongelma on yrittää mallintaa sitä, miten ihminen tuon ongelman yleensä ratkaisee, ja ohjelmoida se tietokoneelle. Ratkaisutapa voi olla myös täysin erilainen. Jos esimerkiksi tietokoneeessa riittää laskentatehoa, voimme käskeä sitä yrittämään jokaista mahdollista ratkaisuvaihtoehtoa ongelmaan, jolloin se välttämättä joskus löytää jonkun ratkaisukeinon, jos ongelmaan ylipäänsä on ratkaisu olemassa.

Tämä jako siitä, miten rationaalisuuteen päästään on ollut merkittävässä roolissa tekoälyn kehityksessä. Siis se, pyritäänkö rationaalisuuteen mallintamalla ihmisen ajattelua vai pidetäänkö tärkeänä vain sitä, että kone löytää ongelmaan jonkun ratkaisun. Voidaan hyvin karkeasti tehdä jako perinteiseen tekoälytutkimukseen ja kognitivismiin (Anderson, 1989, 73). Perinteisessä ei pidetty niinkään tärkeänä sitä simuloiko systeemi ihmismieltä, riittää että systeemi ratkaisee ongelman keinoista riippumatta. Esimerkkinä perinteisesti tekoälystä voidaan pitää šakkitietokoneita. Vuonna 1997 IBM:n Deep Blue -šakkitietokone voitti šakin maailmanmestarin Garri Kasparovin. Tämä saavutus perustui kuitenkin pelkkään numeronmurskaukseen. Šakkikone pyrki vain tiettyjä heuristisia apuvälineitä käyttäen tutkimaan mahdollisimman suuren määrän erilaisia siirtomahdollisuuksia ja etsimään niistä parhaan. Sen tapa etsiä ratkaisu ongelmiin oli hyvin erilainen kuin ihmisellä. Ihmisenhän tapa pelata šakkia perustuu hyvin vähän

siirtojen eteenpäin laskemiseen. Ihmisen muisti ei kykene hallitsemaan kuin aivan muutaman siirron eteenpäin, joten tässä ominaisuudessa tietokone lyö ihmisen täysin selvästi. Ihminen sen sijaan on mestari näkemään kokonaistilanteen šakkilaudalla ja huomaamaan oleelliset ja tärkeät siirrot pelissä. Ihminen pystyy esimerkiksi huomaamaan tiettyjä tilanteita, jotka aikaisemmissa peleissä ovat olleet samankaltaisia, ja tekemään näistä peleistä saadun kokemuksen perusteella oikeita siirtoja. Ihmisen hahmontunnistus ja yleistyskyky ovat siis omaa luokkaansa, eikä šakkikoneisiin ole kyetty luomaan vastaavaa kykyä.

Perinteisen tekoälyn yhteydessä puhutaan usein myös bottom-up -mallintamisesta tai forward engineeringistä. Näissä idea on se, että ratkaistavaa ongelmaa lähdetään ratkomaan ikäänkuin puhtaalta pöydältä, etsien ratkaisuja jotka tuovat meitä lähemmäs lopullista ratkaisua, kunnes se lopulta tulee vastaan. Otetaan esimerkiksi tehtävä, jossa minun tulisi matkustaa Turusta Tallinnaan. Pyrin tekemään ratkaisun, jolla pääsen lähimmäksi päätepysäkkiä. Jos vaihtoehtona olisivat juna Tampereelle, juna Helsinkiin, laiva Tukholmaan tai lentokone Ouluun, vaihtoehto, joka toisi minut lähimmäs päätepysäkkiä, olisi ilmiselvästi juna Helsinkiin. Tehtävä Helsingistä Tallinnaan ratkeaisi samalla periaatteella, ja näin löytäisimme ratkaisun koko tehtävän ratkaisemiseksi. Toki myös ihminen ratkaisee monet ongelmat samankaltaisesti, muttei välttämättä. Edellisessä tehtävässä ihminen tuskin olisi edes harkinnut vaihtoehtoa matkustaa ensiksi Turusta lentokoneella Ouluun.

Toisenlainen tapa kuvata perinteistä tekoälytutkimusta voisi olla tällainen: Meillä on ongelma, kutsutaan sitä syötteeksi (input). Toisekseen meillä on jokin toivottu ratkaisu, jota voidaan kutsua tulokseksi (output). Nyt pyritään löytämään jokin funktio, jolla pääsemme syötteestä haluttuun tulokseen. Siitä millainen tämä funktio on, ei olla niinkään kiinnostuneita.

Kognitivismissa sen sijaan pyritään tarkemmin tutkimaan sitä, miten juuri ihmismieli toimii. Etsitään ongelmiin ratkaisua siitä, miten ihmismieli tuon ongelman ratkaisee. Neurologian ja psykologian tutkimustuloksia apuna käyttäen pyritään mallintamaan tietokoneella ihmismieltä. Voidaan puhua myös top-down -mallintamisesta tai reverse engineeringistä. Meillä on valmis toimiva älykäs systeemi, ihmisaivot. Miksi siis

turhaan yrittäisimme keksiä tyhjästä ratkaisukeinoja, kun me löydämme ratkaisut ihmisaivoja tutkimalla. Kognitivismissa halutaan löytää syöte-tulos -parin välille juuri se funktio, joka vastaa sitä funktiota, minkä ihmismieli toteuttaa. Analysoimalla valmista mallia, aivoja, löydämme ratkaisun. On kuitenkin täysin toinen asia, kuinka helppo tehtävä tällainen ihmismielen analysointi on, tähän asti se on ollut erittäin haastava.

Nämä erottelut eivät kerro kuitenkaan juuri mitään tämän tutkielman tutkimuskysymyksen ratkaisusta. Vaikka kykenemme rakentamaan koneen, joka käyttäytyy täysin rationaalisesti, tai ainakin ihmisen tavoin, ei se kuitenkaan välttämättä kerro sitä, onko tämä kone tietoinen. Nämä edellä mainitut erottelut ovat pikemminkin vain metodologisia eroja. Onko ihmismieltä tarkkaan mallintava tietokone yhtään sen kykenevämpi tietoisuuteen, kuin tietokone, joka käyttäytyy kuin ihminenkin, mutta jonka prosessi käyttäytymisen luomiseen on erilainen? Tekoäly onkin harhaanjohtava termi tässä tutkielmassa siinä mielessä, että toki kone voi olla älykäs olematta tietoinen. Nykyään usein puhutaankin konetietoisuudesta, kun halutaan tuoda esille myös tietoisuuden rooli tekoälytutkimuksessa.

1.2.2 Symbolinen tekoäly vs. konnektionismi

Perinteisen tekoälyn ja kognitivismin lisäksi on myös yksi toinen merkittävä jako tekoälytutkimuksen sisällä. Tämä jako pyrkii jo suoremmin vastaamaan kysymykseen, milloin kone voi olla aidosti tietoinen, samalla tavoin kuin ihminen. Se heijastelee eroa perinteisen tekoälytutkimuksen ja kognitivismin välillä, mutta ei vastaa kuitenkaan aivan sitä. Pikemminkin se on jako kognitivismin sisällä. Kyse on siitä pyritäänkö koneella mallintamaan ihmismieltä vai ihmisaivoja. Tämä on luonut kaksi kilpailevaa tutkimusohjelmaa, jotka varsinkin tekoälytutkimuksen alkutaipaleella olivat hyvin eri mieltä toistensa onnistumismahdollisuudesta. Tosin nykyään kilpailu ei ole enää aivan yhtä kiihkeää, ja on pyritty yhdistämään molempien hyvät puolet. Kyse on symbolisesta tekoälystä ja konnektionismista.

Symbolisen tekoälyn mukaan ihmismielen toiminta on perimmäiseltä luonteeltaan symbolien manipulaatiota. Tietokoneiden toiminta taas on myös luontaisesti symbolien

manipulointia, joten tästä voidaan suoraviivaisesti johtaa teoria tietoiselle koneelle. Kone, joka simuloi ihmismielen käyttämien symbolien manipulaatiota samalla funktionaalisella tasolla kuin ihmismieli manipuloi käyttämiään symboleita, on symbolisen tekoälyn mukaan kykenevä samanlaiseen tietoisuuteen kuin ihminen.

Konnektionismi sen sijaan kieltää, että pelkkä symbolien manipulointi olisi riittävä keino luoda aito tietoisuus. Sen mukaan tietoisuuteen tarvitaan funktionaalinen organisaatio, joka vastaa sekä funktionaalisesti että rakenteellisesti ihmisaivoja. Konnektionismi puoltaa tekoälyn luomiseen niin sanottuja konnektionistisia verkkoja, jotka pyrkivät mallintamaan ihmisaivojen neuroneita ja synapsien välistä sähköimpulssien liikettä.

Näillä molemmilla tekoälyohjelmilla on omat vahvuutensa ja heikkoutensa. On olemassa tiettyjä syitä, joiden takia voidaan nähdä ero näiden eri tekoälyohjelmien tavoissa ratkaista tietoisen koneen ongelma. Tarvitaanko tietoisuuteen esimerkiksi juuri ihmisaivojen funktionaalista organisaatiota vastaava systeemi vai eikö jokin täysin erityyppinen malli olisi kykenevä tietoisuuteen? Tämä kysymys on hyvin kiinnostava ja tulemme myöhemmin tässä tutkielmassa paneutumaan kysymykseen siitä, millaisia filosofisia seurauksia näillä eri tekoälyohjelmilla on.

1.3 Tekoäly filosofisena ongelmana

Edellä mainitut tekoälyohjelmat tekevät kuitenkin jo varsin pitkälle meneviä oletuksia tietoisuuden luonteesta. Ne olettavat, että siihen, että kone voisi olla aidosti tietoinen, riittää se, että kone on oikealla tavalla ohjelmoitu. Niille ongelma on se, miten löytää ja luoda se oikea ohjelma, joka mahdollistaa tietoisuuden. Kenties suurempi ongelma on kuitenkin se filosofinen ongelma. Voivatko nämä tutkimusohjelmat ylipäätään onnistua? Onko mielen toiminta todellakin pelkästään tietokoneohjelman suoritusta?

Kuten huomaamme, tekoäly on ongelma sekä filosofeille että insinööreille. Voidaan melkein sanoa, että tekoäly on filosofiaa (Glymour, 1988, 195). Itseasiassa suuri osa jopa aivan yleisistä tekoälyssä käytetyistä menetelmistä ja ideoista onkin peräisin filosofisen logiikan ja tieteenfilosofian kirjallisuudesta. Usein insinööreillä ja

filosofeilla on kuitenkin varsin vastakkainen näkemys siitä kuinka tekoälyn ongelma ratkaistaan. Insinöörit pitävät filosofiaa aivan liian korkealentoisena ja epämääräisenä. He empiristeinä luottavat siihen, että jossain vaiheessa saadaan tarpeeksi tieteellistä todistusaineistoa jonkin ilmiön suhteen, jonka perusteella voidaan tehdä tarvittavat johtopäätelmät. Filosofit sen sijaan moittivat insinöörejä usein pinnallisuudesta ja siitä, että vaikka insinöörien oppikirjat ovat täynnä hienoja tieteellisiä kaavoja, nämä kaavat eivät kuitenkaan kerro mitään asioiden todellisesta luonteesta.

Kysymyksessä tekoälystä voin tunnustaa, että sympatiani kallistuu hienoisesti filosofien puolelle. Siitä huolimatta, että luonnontieteellisen tutkimuksen menestys on ollut yhtä riemuvoittoa jo vuosisatojen ajan. Vaikka monet fysiikan ja biologian huomattavat saavutukset ovat saaneet meidät uskomaan, että näiden avulla voimme ratkaista kaikki maailmaa koskevat kysymykset, on yksi asia, jonka ratkaisu ei ole tullut juuri yhtään lähemmäs tieteellisen kehityksen edetessä. Se on kysymys juuri tietoisuuden synnystä. Miten fysikaalisesta materiasta voi syntyä tietoinen mieli? Vastaukset tietoisuuden syntyyn jopa ihmisissä ovat erittäin kiistanalaisia, joten voimmeko olettaa, että kysymys koneitten tietoisuudesta olisi ratkaistavissa?

Tässä tutkielmassa tullaankin lähestymään tekoälyä nimenomaan filosofisena, erityisesti mielenfilosofisena ongelmana. Tietoisuudesta on kirjoitettu paljon teorioita. Suurin osa niistä keskittyy kuitenkin sellaisten funktionaalisten asioiden kuten aistien tai muistin toiminnan selittämiseen, jättämättä vastaamatta tärkeimpään kysymykseen filosofian kannalta, eli siihen miten nämä funktionaaliset prosessit tuottavat sen elämyksellisen ja kokemuksellisen ilmiön, jonka aistimme meille välittävät. Tässä tutkielmassa pyrin ensisijaisesti vastaamaan tähän laiminlyötyyn kysymykseen. Lähtökohtana on se, miten komputationaalinen systeemi, kuten tietokone, voisi tämän elämyksellisen ilmiön tuottaa.

Tarkastellaan johdantokappaleen lopuksi vielä hieman termiä tekoäly. Sen alkupuoli, teko-, viittaa keinotekoiseen. Tekoäly viittaa siis keinotekoiseen älyyn, tai tässä tutkielmassa voisi puhua kai keinotekoisesta tietoisuudesta. Käsitteeseen keinotekoinen liittyy eräs mielenkiintoinen piirre. Onko keinotekoinen nimittäin jotenkin huonompaa kuin aito? Joissain tapauksissa näin näyttää olevan: esimerkiksi tekokukat eivät ole

sama asia kuin aidot kukat. Sen sijaan on olemassa keinotekoista valoa, joka ei ole luonnonvaloa. Esimerkiksi loisteputket tuottavat keinovaloa. Eikö tämä ole kuitenkin aivan yhtä aitoa valoa kuin se mitä aurinkokin tuottaa? Kumpi on siis totuus tekoälyn tapauksessa? Mikä tekee asiasta aidon ja mikä on pelkkää simulaatiota? Tämä on yksi tärkeistä kysymyksistä tekoälyn kannalta, ja sitä tullaan pohtimaan kun tutkimme simulaation ja duplikaation eroja.

Seuraavaksi luvussa esittelen tarkemmin sen erityislaatuisen ominaisuuden, josta puhumme kun tutkimme voivatko tietokoneet olla aidosti tietoisia. Mitä on tietoisuus?

2 TIETOISUUS

Tietoisuus oli pitkään vaiettu käsite tekoälytutkimuksessa. Se johtuu suureksi osaksi siitä, että tietoisuus on hyvin epämääräinen käsite. Siitä ei ole olemassa yleisesti hyväksyttyä yksiselitteistä määritelmää. Se on pikemminkin klusterikonsepti (Sloman & Chrisley, 2003) koostuen useista läheisistä käsitteistä kuten aistiminen, tunteminen, kokeminen, uskominen, haluaminen, muistaminen, oppiminen jne. Osa näistä käsitteistä voidaan määritellä suhteellisen yksiselitteisesti, mutta kaikki tutkijat eivät ole esimerkiksi samaa mieltä siitä, mitkä näistä käsitteistä ovat välttämättömiä tietoisuuden synnylle. Näitä tietoisuuden tai kognition eri piirteitä voisi ja ehkä pitäisikin käsitellä huomattavasti. Ne kaikki ovat myös tekoälytutkimuksessa tärkeitä käsitteitä, ja on paljon tutkimusta siitä, miten ne voitaisiin implementoida tietokoneelle. Tämän tutkielman ongelmaa ratkaistaessa niiden rooli ei ole kuitenkaan ensisijainen ja siirrynkin suoraan tämän tutkielman kannalta tärkeimpään ilmiöön, joka tietoisuuteen liittyy, eli kvalioihin.

2.1 Kvaliat

Tärkein tietoisuuteen liittyvä ilmiö on se tietty elämyksellinen tai kokemuksellinen piirre, joka siihen liittyy. Mielenfilosofiassa näitä ilmiötä kutsutaan fenomenaalisiksi kvaliteeteiksi, lyhyesti kvalioiksi. Niillä tarkoitetaan juuri sitä tiettyä tunnetta, joka mentaaliseen tilaan liittyy: on jotain, miltä tuntuu olla mentaalisessa tilassa.

Yksi asia, mikä tekee kvalioista erityislaatuisia ja ongelmallisia, on niiden poikkeuksellinen subjektiivisuus. Ongelma, jonka Thomas Nagel toi nykyfilosofiaan artikkelissaan What Is It Like To Be A Bat? (1974). Kellään toisella ihmisellä ei ole mitään keinoa päästä käsiksi minun tuntemuksiini, eikä myöskään minulla toisten tuntemuksiin. Kukaan ei voi koskaan tarkasti tietää miltä minusta tuntuu. Minulla sen sijaan on suora ja välitön pääsy omiin tuntemuksiini. En tarvitse ulkopuolista todistusta tietääkseni miltä minusta tuntuu. Kun katsomme toisen ihmisen kanssa sinistä taivasta, voimme kyllä puhua siitä kuinka kirkkaan sininen se tänään on ja käyttää yhteisiä käsitteitä kuten "sininen", jolla tarkoitamme sitä elämyksellistä ilmiötä, jonka olemme tottuneet kokemaan katsoessamme pilvetöntä taivasta. Se miltä sininen tuntuu minulle,

ei kuitenkaan välttämättä ole sama kuin se miltä se tuntuu sinulle. On loogisesti täysin mahdollista, että se miltä sinisen esineen katsominen tuntuu minusta, tuntuu sinulle samalta kuin se miltä minusta tuntuu katsoessani punaista esinettä.

Käsite kvalia tuntuu siis olevan vaikeasti määriteltävissä tarkasti, ja käsitys täytyykin muodostaa jokseenkin intuitiivisesti. Tärkeää on kuitenkin se, mitä kvaliat eivät ole. Ne eivät tarkoita niitä funktionaalisia tai rakenteellisia ominaisuuksia, jotka ovat mukana mielemme toiminnassa. Tällaisia ovat esimerkiksi prosessit, jotka hakevat muististamme vanhoja kokemuksia tietoisuutemme kohteeksi tai prosessi, jolla mielemme tunnistaa näköhavaintomme kohteen tietyksi esineeksi. Kvaliat toki saattavat liittyä hyvinkin kiinteästi myös näihin funktionaalisiin tapahtumiin. Saattaa jopa olla, että kaikkiin mielen funktionaalisiin tapahtumiin liittyy myös fenomenaalinen ilmiö, eli kvalia. Tärkeää on kuitenkin se, että kvalialla ei tarkoiteta funktionaalista prosessia vaan nimenomaan sitä tuntua, joka mielen tapahtumiin liittyy.

Voidaan toki kyseenalaistaa onko tällaisia elämyksellisiä ilmiöitä oikeasti olemassa, vai ovatko ne vain jonkinlaista harhaa, kuten esimerkiksi Daniel Dennett (1991) on väittänyt. Hänen mukaansa ne ovat vain eräänlaisia mielleyhtymien ryppäitä, redusoitavissa siis muihin mielen ominaisuuksiin. Tässä tutkielmassa oletetaan kuitenkin, ainakin aluksi, että kvaliat ovat todellinen ilmiö. Näin tekee myös suurin osa muista mielenfilosofeista. Sillä kun tarkastelemme tietoisuuttamme, lähimpänä meitä tulee vastaan juuri tämä subjektiivinen ja fenomenaalinen tuntu, joka hallitsee tietoisuuttamme. Kuten Descartes (1641) jo aikanaan huomasi, voimme epäillä lähes kaikkea mahdollista. Voin epäillä esimerkiksi sitä, että edessäni on tietokoneen näyttö, saatanhan olla kenties unessa. Kuitenkin siitä voin olla varma, että minusta *tuntuu* siltä, että edessäni olisi tietokoneen näyttö. Ilman erittäin hyviä perusteita kvalioiden olemassaolon kieltäminen ei siis ole mahdollista.

Kvalian olemassaoloon ei tarvita myöskään erityistä tietoisuutta ympäristöstä. Voin olla esimerkiksi unessa tai huumattuna, ja täysin tiedoton ympäristöstäni. Silti voin näissä tiloissa kokea ja tuntea monia asioita. Vaikka olen tiedoton ympäristöstäni, voin kuitenkin olla tietoinen omista sisäisistä tiloistani.

2.2 Miksi tietoisuus on ongelma?

Monet tietoisuuden piirteistä voidaan selittää ongelmitta perinteisen fysiikan ja biologian keinoin. Tai ainakin meillä on jonkinlainen käsitys siitä miten niiden tutkimusta voidaan jatkaa. Esimerkiksi se miten ihminen kykenee oppimaan tai miten ihmisen muisti toimii, voidaan selittää ainakin jollakin tasolla neurobiologian keinoin. Nämä kyvyt on myös helppo implementoida tietokoneelle. Tietokoneilla on ollut muisti, eli ne ovat kyenneet tallettamaan tietoa, jo lähes niiden ensi-ilmestymisestä asti, ja tietokoneet pystyvät nykyään myös oppimaan monia asioita, vaikka se hieman hankalampi ongelma onkin. Mutta kysymys siitä, miksi ihmisillä on kvalioita ja miten ne syntyvät, tai miten ne voitaisiin luoda myös tietokoneelle, on edelleen täysin hämärän peitossa.

Neurobiologia on tutkinut valtavasti miten ihmisaivot toimivat. Se on kyennytkin paikallistamaan useita funktionaalisia toimintoja tiettyyn kohtaan aivoissa. Näköaistimukset ovat läheisesti korrelaatiossa neokorteksin aktiivisuuden kanssa (Crick & Koch, 1995), puheen tuotto ja havaitseminen taas Brocan ja Wernicken alueilla. Myös tietoisuuden neurokorrelaattia on yritetty paikallistaa suurella tarmolla. Ja aivoista ei varmaan yhtään pistettä löydykään enää, jota ei olisi ehdotettu tietoisuuden tyyssijaksi. Tunnetuin ehdotus on varmaan Descartesin esittämä käpylisäke. Kuitenkaan tietoisuudelle ei vakuuttavasti ole löytynyt mitään yksittäistä paikkaa. Luultavasti tietoisuus syntyykin eripuolilla aivoja tapahtuvasta aktiivisuudesta hajautetusti. Toisaalta vaikka tietoisuudelle jokin neurokorrelaatti löytyisikin vielä, tämä ei kuitenkaan riittäisi. Miksi jonkin tietyn aivoalueen, tai hajautetusti koko aivojen, aktiivisuus aiheuttaisi kvaliat?

Neurobiologian ongelma on siis se, että se pystyy löytämään vain tiettyä aktiivisuutta aivoista jonkin mielen tapahtuman yhteydessä. Korrelaatio ei ole kuitenkaan riittävä selitys. Neurobiologian löydös saattaisi olla esimerkiksi, että punaisen värin kokeminen on yhtä kuin se, että synapsi X lähettää sähköimpulssin synapsiin Y. Vaikka todella olisikin niin, että aina ihmisen aistiessa punaista, sähköimpulssi lähtisi synapsista X synapsiin Y, ei tämä kerro kuitenkaan mitään siitä, miksi sähköimpulssi aiheuttaa juuri punaisen väriaistimuksen, eikä esimerkiksi sinistä. Eihän esimerkiksi fakta "kaikilla

olioilla joilla on sydän, on myös maksa" tarkoita sitä, että sydän jollain tapaa aiheuttaisi maksan olemassaolon. Niiden korrelaatio johtuu täysin niistä riippumattomista syistä. Niillä on kausaalinen suhde johonkin yhteiseen tekijään, mutta keskenään ne eivät ole suhteessa. Eikä tämä ongelma rajoitu vain neurobiologiaan, vaan on kaiken fysikalismin (ks. 3.1) ongelma. Jos lisättäisiin tutkimuksen tarkkuutta vaikka atomi- tai kvanttitasolle, löytäisimme silti vain lisää mentaalisen kanssa korreloivia ilmiöitä, emmekä itse syytä. Jos korrelaatio ei riitä selitykseksi, niin mikä sitten olisi riittävä selitys kvalian syntyyn? Miksi tietty fysikaalinen prosessi tuottaisi kvaliat?

David Chalmers (1996, 24) tekeekin kuuluisan jaottelunsa tietoisuuden helppoihin ja vaikeisiin ongelmiin. Helpot ovat niitä, jotka näyttäisivät olevan ratkaistavissa perinteisin fysiikan, biologian yms. keinoin. Vaikea ongelma taas on se, miksi meillä on kvalioita. On kyse todellakin ikivanhasta mieli-materia -ongelmasta. Mielen ja materian välillä vallitsee selityksellinen kuilu. Onko mitään keinoa kutistaa tuota kuilua. Jo Leibniz (1714, §17) aikanaan huomasi tuon kuilun olemassaolon kuuluisassa tuulimyllyvertauksessaan, joka myös ennakoi tekoälyn syntyä: Kuvitellaan tietoinen mekaaninen kone, joka on niin iso, että ihminen mahtuu kulkemaan sen sisuksissa. Vaikka ihminen kuinka etsisi tietoisuutta sieltä, ei hän löydä sitä sieltä. Sama pätee ihmisaivoihin. Voidaan kuvitella, että kutistettaisiin ihminen niin pieneksi, että hän mahtuisi kulkemaan toisen ihmisen aivoissa. Silti, vaikka hän kuinka etsisi sieltä tietoisuutta, niin hän löytäisi vain aivojen synapseja ja niiden välisiä sähköimpulsseja. Kuitenkin jollain mystisellä tavalla näiden harmaiden aivosolujen konjunktiosta syntyy tietoinen mieli. Fysikalismin keinot eivät näytä riittävän ongelman ratkaisemiseksi.

Chalmersin mukaan tämän kuilun ylittämiseksi täytyy luoda täysin uusi teoria maailman rakenteesta. Tässä tutkielmassa en lähde uutta teoriaa maailman rakenteesta luomaan, mutta on tärkeää kuitenkin tietää millaisia teorioita mielen ja materian suhteesta on olemassa. Chalmersin jako helppoihin ja vaikeisiin kysymyksiin heijastelee nimittäin ajatusta, että maailmalla on jollakin tapaa dualistinen luonne. Dualismia, fysikalismia jne. käsitelläänkin hieman tuonnempana. Palataan hetkeksi ensin tietokoneiden pariin.

2.3 Turingin testi - testi tietoisuudelle?

Alan Turingin oli yksi tärkeimmistä tietokoneen kehittäjistä. Sen lisäksi, että hän kehitti teoreettista pohjaa tietokoneen synnylle, hän oli myös tietoinen tietokoneen mullistavasta vaikutuksesta maailmalle. Turing näki, että jonain päivänä tietokoneeseen voidaan luoda ihmisen kaltainen älykkyys ja hän oli myös kiinnostunut tämän filosofisista seurauksista. Olisiko tietokone tuolloin myös aidosti tietoinen, jos se kykenisi käyttäytymään kuten ihminen? Turing tiesi hyvin myös selityksellisen kuilun olemassaolon. Hänen mielestään sen ylittäminen oli mahdotonta. Hän julkaisi vuonna 1950 artikkelin Computing Machinery and Intelligence (Turing, 1950), josta tuli lähtölaukaus koko tekoälyn filosofialle. Siinä hän esitteli imitaatiopelin, jota on myöhemmin kutsuttu Turingin testiksi. Koska Turingin mielestä tietoisuuden todentaminen tieteen apuvälineitä käyttäen oli mahdotonta, hän päätyi kehittämään toisenlaisen testin, jolla olisi mahdollista selvittää pystyikö kone aidosti olemaan tietoinen.

Turingin testi on hyvin yksinkertainen. Siinä ns. kyselijä pyrkii selvittämään, kumpi kahdesta vastaajasta on ihminen ja kumpi tietokone, pelkästään kyselemällä näiltä erilaisia kysymyksiä. Kone pyrkii parhaansa mukaan harhauttamaan kyselijän uskomaan, että se on ihminen. Kysymykset ja vastaukset annetaan siten, etteivät osapuolet voi nähdä toisiaan, eli käytännössä kirjallisesti. Jos kyselijä ei tietyn kysymysmäärän jälkeen pysty erottamaan, kumpi vastaajista on ihminen ja kumpi tietokone, voidaan Turingin mukaan olettaa tietokoneen oikeasti kykenevän ajattelemaan ja ymmärtämään kieltä.

Hän piti siis riittävänä syynä olettaa, että kone on tietoinen, samaa syytä minkä takia pidämme muita ihmisiä tietoisina. Ainoa syy olettaa, että muut ihmiset ovat tietoisia, on heidän käyttäytymisensä. Muut ihmiset käyttäytyvät yleensä samoissa tilanteissa samalla tavoin kuin minä käyttäytyisin niissä. Käyttäytymistäni taas ohjaavat omat haluni, ajatukseni ja tietoisuuteni. Siispä, muutkin ihmiset ajattelevat ja ovat tietoisia. Perinteinen analogia-argumentti. (Russell, 1948, 482) Tämä ei tietenkään ole pätevä argumentti, mutta houkutteleva ja paljon käytetty. Itseasiassa niin houkutteleva, että käytämme sitä jatkuvasti sosiaalisessa kanssakäymisessä. Emme voi muiden ihmisten

suhteen olla yhtään sen varmempia heidän tietoisuutensa olemassaolosta, joten Turingin mielestä ei ollut mitään syytä kieltää koneiltakaan älykkyyttä, jos ne kerran pystyvät antamaan niin ihmismäisiä vastauksia, ettei niitä pystynyt erottamaan ihmisen antamista (Turing, 1950, 433).

Turingin testin uskottavuus on kuitenkin kyseenalaistettu useaan otteeseen. Ja harva tekoälyn filosofian tutkija nykyään pitääkään sitä täysin luotettavana. Se on kuitenkin tietyllä tapaa hyvin perusteltu. Se pitää kuitenkin sisällään paljon oletuksia tietoisuudesta ja maailman luonteesta. Seuraavassa luvussa lähden perehtymään näihin asioihin. Millaisia tapoja siis on selittää mielen ja maailman suhde? Samalla huomaamme mitä Turing oletti testiä luodessaan. Turingin testiin palaamme vielä useasti tämän tutkielman aikana.

3 TIETOISUUSTEORIOISTA

Tutkielman johdannossa esitettiin väite, että ihmisaivot ovat vain monimutkainen tietokone. Tietoisuus syntyy siis kun realisoidaan tarpeeksi monimutkainen tietokone. Tämä teoria mielen ja materian suhteesta kantaa nimeä komputationalismi. Yksi tapa ilmaista tämän tutkielman aihe on siis kysymys: onko komputationalismi tosi? Komputationalismi on nykyään yksi suosituimmista ja varteenotettavimmista vaihtoehdoista mielen ja materian suhteesta, mutta se ei ole suinkaan ainoa. Se on varsin nuori tulokas muiden joukossa. Esittelenkin nyt tiiviisti muut tärkeimmät teoriat mielen ja materian suhteesta, jotta saamme kuvan siitä millainen historia komputationalismia on edeltänyt, ja mitä ongelmia muissa teorioissa on ollut.

Huomautuksena se, että seuraavat aliotsikot eivät suinkaan ole täysin erillisiä ja toisensa poissulkevia teorioita. Ne ovat sen sijaan vain nimiä tietyille suosituille teorioille. Kukin niistä hyväksyy joukon teesejä, ja toisaalta ne hylkäävät toisen joukon teesejä. Eri teoriat voivat jakaa samoja teesejä. Toisaalta sama teoria voi olla ottamatta kantaa johonkin kysymykseen, jolloin teoria voi jakautua kahteen alateoriaan saman teorian sisällä. Teoriat eivät myöskään ole kronologisessa järjestyksessä.

3.1 Fysikalismi

Aloitan fysikalismista, koska se on teoria, jonka pohjalle käytännössä kaikki nykypäivän mielenfilosofia rakentuu. Sitä pidetäänkin lähes oletusarvoisesti totena, vaikkakin kuten jo edellisessä kappaleessa (2.2) huomasimme, että myös se sisältää ongelmia, päällimmäisenä ongelmana kvalioiden selittäminen.

Kun esitämme seuraavia väitteitä: "ihmisellä on mieli", "ihminen on tietoinen" tai "ihminen kokee jonkun fenomenaalisen ilmiön", tulemme huomaamattamme antaneeksi näille mentaalisille käsitteille tietyn ontologisen statuksen. Lause "ihmisellä on mieli" nimittäin voi saada meidät ajattelemaan, että pitäisi olla olemassa jokin entiteetti, johon sana "mieli" viittaa. Lauseen "ihminen on tietoinen" voi sen sijaan olettaa tarkoittavan, että tietoisuus on tietty tila, jossa ihminen voi olla: ihminen on tietoisessa tilassa. "Ihminen kokee fenomenaalisen ilmiön" taas pitää mentaalisuutta jonkinlaisena

tapahtumana tai prosessina: tietoisuus on tapahtuma tai tapahtumien joukko. Ja muistutan vielä, että tässä tutkielmassa tarkoitetaan mielellä, tietoisuudella, fenomenaalisilla ilmiöillä, mentaalisuudella jne. käytännössä samaa asiaa: ne ovat sellaisia ilmiöitä joihin liittyy kiinteästi kvaliat.

Nämä edelliset lauseet ovat kaikki hieman erilaisia muotoiluja mentaalisuudesta. Kirjaimellisesti otettuna ne antavatkin täysin erilaisen ontologisen statuksen mentaalisuudelle. Ensimmäinen lause saattaa pitää mentaalisuutta täysin erillisenä substanssina, joka on irti fysikaalisesta materiasta. Jälkimmäiset muotoilut taas pitävät mentaalisuutta pikemminkin jonkinlaisena ominaisuutena: ihmisaivot ovat tietyssä tilassa, jolla on sellainen erityislaatuinen ominaisuus, että se on mentaalinen tila, tai ihmisaivot käyvät läpi tietynlaista prosessia, jolla on sellainen erityislaatuinen ominaisuus, että se on mentaalinen prosessi. Nämä eivät siis pidä mentaalisuutta erillisenä substanssina, vaan pitävät sitä vain ominaisuutena, joka fysikaalisella materialla on erityistapauksissa.

Fysikalismi pitää mentaalisuutta nimenomaan vain ominaisuutena, joka on sidoksissa fysikaaliseen materiaan. Toisin sanoen se kieltää, että perimmiltään olisi olemassa mitään muuta kuin fysikaalista materiaa. Mentaalisuus on vain fysikaalisen materian yhdistelmästä syntyvä ominaisuus ja näin fysikalismi siis antaa fysikaaliselle ensisijaisen statuksen verrattuna mentaalisuuteen. Materian ensisijaisuudesta tulee myös käsite materialismi, joka on käytännössä synonyymi fysikalismille. Hieman vanhahtavampi nimitys vain. Fysikalismi on yleensä pohjana myös komputationalismille. Kun fysikaalinen tietokone ohjelmoidaan tietyllä tavalla, tälle materialle syntyy myös mentaalinen ominaisuus.

Fysikalismiin, kuten yleisesti kaikkeen mielenfilosofiaan liittyy läheisesti mieli-materia supervenienssi -teesi. Yleinen supervenienssin käsite on seuraava: B-tyypin ominaisuudet ovat supervenientteja A-tyypin ominaisuuksien suhteen, jos ei ole kahta sellaista tilannetta, joissa B-tyypin ominaisuudet eroaisivat toisistaan A-tyypin ominaisuuksien pysyessä samoina. Mieli-materia supervenienssi -teesi, toisin sanoen väite, että mieli on materian supervenientti ominaisuus, voidaan muotoilla siis seuraavasti: Jos kaksi fysikaalista objektia ovat fysikaalisesti identtisiä, ne ovat

välttämättä myös mentaalisilta ominaisuuksiltaan identtisiä. Esimerkiksi, jos kaksi ihmistä olisivat fysikaalisesti täysin identtisessä tilassa, he olisivat välttämättä tuolloin myös samanlaisessa mentaalisessa tilassa, toisin sanoen heillä olisi samanlaiset ajatukset ja aistimukset. Mieli-materia -teesi on intuitiivisesti erittäin uskottava ja myös yleisesti hyväksytty lähes kaikessa nykymielenfilosofiassa, kuten myös fysikalismissa. Teesistä käytännössä seuraa se, mitä fysikalismi väittääkin, eli että mentaalisuus on fysikaalisesta materiasta riippuvainen, tai että fysikaalinen materia determinoi mentaaliset ominaisuudet.

Fysikalismi jakautuu kuitenkin kahteen tärkeään haaraan. Erottava kysymys on se millaisina fysikalistit pitävät mentaalisia ominaisuuksia. Mitä ne ovat? Ovatko ne erotettavissa fysikaalisista ominaisuuksista, vai ovatko ne nimenomaan vain fysikaalisia ominaisuuksia? Ovatko mentaaliset ominaisuudet redusoitavissa tai identifioitavissa fysikaalisiin ominaisuuksiin?

Ei-reduktiivinen fysikalismi ja ominaisuusdualismi pitävät mentaalisia ominaisuuksia erillisinä ja redusoimattomina ominaisuuksina. (Ominaisuusdualismi erottuu ei-reduktiivisesta fysikalismista siten, että se ei välttämättä hyväksy mieli-materia supervenienssiä, joten se ei välttämättä ole myöskään fysikalismia.) Reduktiivinen fysikalismi, eli ominaisuusmonismi sen sijaan pitää mentaalisia ominaisuuksia redusoitavissa fysikaalisiin ominaisuuksiin. Reduktiivisen fysikalismin mukaan on siis aidosti olemassa vain fysikaalisia ominaisuuksia. On olemassa myös vahva versio reduktiivisesta fysikalismista nimeltään eliminativismi. Sen mukaan tosia lauseita, joissa esiintyy mentaalisia ominaisuuksia, ei ole olemassa. Sen sijaan perinteisessä reduktiivisessa fysikalismissa voimme silti puhua mentaalisista ominaisuuksista ja käyttää niitä tieteellisissä teorioissa, vaikka pohjimmiltaan onkin olemassa vain fysikaalisia ominaisuuksia.

Suuri osa nykymielenfilosofiasta keskittyy pohtimaan näiden vaihtoehtojen mielekkyyttä. Palaan ennen näihin paneutumista vielä kuitenkin siihen vaihtoehtoon jonka fysikalismi torjuu, eli substanssidualismiin, jolla on historiallisesti erittäin suuri merkitys.

3.2 Dualismi

Dualismi on näkemys joka on hallinnut ajattelun historiaa hyvin suuresti. Suurin osa kulttuureista, joissa uskonnoilla on ollut jonkinlainen asema, ovat olleet syvästi dualistisia. Sielun ja ruumiin erottelu on ollut keskeinen teesi lähes kaikissa uskonnoissa. Myös filosofiassa dualismi on ollut merkittävässä roolissa. Antiikin Kreikassa Platon erotteli maailman ja ideoitten maailman. Descartes taas muotoili yhden tunnetuimmista substanssidualismin versioista: karteesiolaisen dualismin. Nykyään dualismi on tieteen ja filosofian piirissä varsin kielletty oppi, johon ovat syynä monet ongelmat, jotka se tuo mukanaan. Pieni vähemmistö on kuitenkin yhä nykyään, joka sitä kannattaa, mutta heidänkin dualistinen näkemyksensä on kuitenkin hyvin erilainen teoria verrattuna perinteisiin näkemyksiin sielusta ja ruumiista.

Substanssidualismin mukaan on olemassa kaksi fundamentaalista substanssia: mentaalinen ja fysikaalinen. Niiden olemassaolo on täysin riippumattomia toisistaan. Voidaan esimerkiksi ajatella, että ihmisellä on fysikaalisen ruumiin lisäksi mentaalinen mieli, ja molemmat näistä substansseista voivat olla olemassa ilman, että toinen niistä olisi. Tässä ajatuksessa sinänsä ei ole mitään vikaa eikä ongelmaa. Sen sijaan ongelma seuraa yhdestä teesistä, minkä dualismi useimmiten hyväksyy. Se on teesi, jonka mukaan fysikaalinen ja mentaalinen voivat vaikuttaa toisiinsa kausaalisesti. Eli joillakin fysikaalisilla tapahtumilla voi olla mentaalinen syy, ja päinvastoin.

Mitä syitä voisi sitten olla, että tarvittaisiin myös mentaalinen substanssi? Miksei pelkkä fysikaalinen materia riittäisi? Tähän on olemassa yleisesti ottaen kahdenlaisia perusteluita. Ensimmäiseksi, on olemassa epistemologisia syitä. Nämä vetoavat siihen, että tietomme ruumiistamme ja mielestämme on perustavalla tavalla erilaista ja asymmetristä. Puhuin tästä jo kvalioitten yhteydessä (2.1). Descartes huomasi aikanaan "cogito"-argumentissaan, että tietomme mielestämme on suoraa ja välitöntä, ja samalla subjektiivista. "Ajattelen, olen". En voi epäillä mieleni olemassaoloa. Sen sijaan tieto ruumiistani ei ole samanlaista. Saan tiedon ruumiistani aistieni kautta ja se on siis myös ei-subjektiivista. Aistini voivat kuitenkin myös pettää, joten tietoani ruumiistani voi myös epäillä. Ja näin koska saan tietoa mielestäni ja ruumiistani hyvin eri tavoin, eivät ruumis ja mieli voi olla myöskään identtisiä.

Toisen tyyppisiä perusteluita dualismille ovat ontologiset syyt. Samaa "cogito" -argumenttia on käytetty myös näiden syiden lähtökohtana. Olennainen asia näissä on se, että mieli ajatellaan tietyllä tapaa essentiaalisesti tai ainakin mahdollisesti ei-spatiaaliseksi. Ajatukseni eivät ole spatiaalisia, ts. ne eivät sijaitse missään tietyssä paikassa. Ei voida löytää mitään avaruudellista pistettä, jossa ajatukseni sijaitsevat. Sen sijaan ruumiillani on välttämättä tarkka fysikaalinen sijainti. Näin ollen koska mieli on mahdollisesti ei-spatiaalinen ja siten immateriaalinen, se ei voi olla identtinen ruumiini kanssa.

Nämä molemmat argumentit ovat jossain määrin uskottavia, vaikka niitä voidaan toki monella tapaa myös kritisoida. Jätän nyt kuitenkin nämä argumentit ja palaan siihen syyhyn, jonka vuoksi dualismi edellä mainituista hyvistä argumenteista huolimatta yleensä hylätään. Eli siihen, että mieli ja ruumiis voisivat olla keskenään vuorovaikutuksessa. Tässä on ylitsepääsemätön ongelma. Koska mieli dualismin mukaan yleensä ajatellaan ei-spatiaaliseksi ja ulotteettomaksi (periaatteessa voidaan ajatella myös spatiaaliseksi, mutta tästä seuraa usein vielä suurempia ongelmia), näemme helposti tästä seuraavan ongelman: miten jokin ei-spatiaalinen voisi vaikuttaa spatiaaliseen materiaan. Fysiikan kielellä sanoen, kaikkiin fysikaalisiin tapahtumiin tarvitaan syyksi jokin fysikaalinen voima. Immateriaalisessa ulottuvuudessa ei voi olla fysikaalisia voimia, eikä sieltä voi siis olla keinoa, jolla voisi vaikuttaa fysikaaliseen materiaan. Näin ollen vuorovaikutussuhdetta ei voi olla.

Dualismi ainakin perinteisissä versioissaan ei ole houkutteleva, mutta on silti pieni joukko puolustajia, jotka sitä kannattavat ja kehittelevät. Dualismin hyvä puoli on se, että sille kvaliat eivät ole ongelma, toisin kuin fysikalismille. Dualismissa kvalioille annetaan itsenäisen entiteetin status, joten ne eivät tarvitse fysikaalista selitystä olemassaololleen.

Fysikalistiset teoriat ovat joka tapauksessa nykymielenfilosofiassa kaikista merkittävimpiä ja siirryn seuraavaksi niihin. Fysikalismin mukaan on vain yksi fundamentaalinen substanssi, materia. Tämän vuoksi dualismia vaivaavaa vuorovaikutusongelmaa ei ole. Fysikalismi on siis monistinen teoria. On olemassa myös muita monistisia teorioita fysikalismin lisäksi, tärkeimpänä idealismi. Sen mukaan on

aidosti olemassa vain mentaalisia entiteettejä. Perinteinen idealismi on kuitenkin nykyään vielä harvinaisempi näkemys kuin dualismi, joten en puutu siihen enempää. Sen sijaan nykyään on olemassa useita monistisia teorioita, joiden mukaan fundamentaalinen entiteetti ei ole materia, eikä myöskään mentaalinen vaan jokin vielä perustavampi entiteetti. Näiden yhteydessä puhutaan usein myös kaksoisaspektiteorioista. En perehdy nyt myöskään näihin, vaan siirryn fysikalismin eri versioihin.

3.3 Behaviorismi

Fysikalismi on teoriana dualismin tapaan hyvin vanha. Uuden ajan filosofeista Thomas Hobbes oli ehkä ensimmäinen joka kannatti fysikalismia, jättäen näin mentaalisen vaille merkittävää asemaa. Vaikka fysikalismilla on ollut jonkinlainen asema pitkään, länsimaista ajattelua hallitsi aina 1900-luvun alkuun asti kuitenkin hyvin pitkälti karteesiolainen ajattelu, joka jakoi siis fysikaalisen ja mentaalisen tiukasti eri lokeroihin. 1900-luvun alussa kuitenkin syntyi näkemys, joka halusi kumota täysin tuon ajattelun. Tämä liike kantaa nimeä behaviorismi.

Karteesiolaisen ajattelun mukaan ihmisellä on suora ja välitön pääsy vain omiin mentaalisiin tiloihinsa. Kaikki muu tieto on saatava aistien kautta. Aistit voivat kuitenkin pettää. Tästä seuraa se, että tietomme ulkopuolisesta maailmasta ja myös muista ihmisistä on aina epäilyksellistä. Emme siis voi koskaan saada tietoa muiden ihmisten mielen tiloista varmuudella. Voimme nähdä vain muiden ihmisten käytöksen, ja tämän perusteella tehdä johtopäätelmiä siitä, mitä näiden ihmisten mielessä liikkuu. Jos ihminen esimerkiksi käyttäytyy kuin hän olisi kovassa kivussa, en voi kuitenkaan varmuudella tietää aiheuttaako tämän käyttäytymisen samanlainen tuntemus, kuin se tuntemus, joka saa minut käyttäytymään kuin olisin kovassa kivussa. Myös tästä aiheesta puhuin jo kvalioitten yhteydessä (luku 2.1).

Behaviorismi kuitenkin halusi kiistää tämän tuloksen. Se ei voinut hyväksyä, että olisi olemassa tällaista subjektiivista tietoa, joka ei olisi muiden ihmisten saatavilla. Behaviorismin mukaan kaikki tieteellinen, mukaanlukien psykologinen, tieto täytyy olla objektiivisesti verifioitavissa tieteellisiä keinoja käyttäen. Näin ollen behaviorismi

halusi kieltää kaikkien mentaalisten ominaisuuksien olemassaolon, tai ainakaan niitä ei tulisi käyttää osana psykologian teoriaa. Ainoa verifioitavissa oleva asia on ihmisen käyttäytyminen, jonka perusteella voidaan rakentaa psykologian teoriaa. Sisäisten mentaalisten tilojen olemassaoloa ei voida olettaa.

Behaviorismi psykologisena teoriana oli hyvin vahva aina 1900-luvun jälkimmäiselle puoliskolle asti. Psykologiassa behaviorismi kielsi, ettei mentaalisia ominaisuuksia tulisi käyttää tieteen tekemisessä. Se ei kuitenkaan välttämättä sanonut mitään niiden todellisesta olemassaolosta. Filosofiassa behavioristiset näkemykset kieltävät usein täysin myös mentaalisten ominaisuuksien olemassaolon. Nämä näkemykset ovat siis reduktiivista fysikalismia, usein vielä sen vahvaa versiota, eliminativismia. Niissä ns. mentaaliset ominaisuudet voidaan korvata käyttäytymiskaavoilla. Mielen tilat ovat aidosti vain dispositioita, eli taipumuksia käyttäytyä tietyllä tapaa tietyissä tilanteessa, eikä mitään muuta. Loogisessa behaviorismissa esimerkiksi käsite pelko voisi olla vain joukko seuraavankaltaisia käyttäytymisdispositioita: "jos näen käärmeen, lähden juoksemaan karkuun, jne." Sellaista mentaalista tilaa kuin pelko ei aidosti ole olemassa, on vain ulkoista käyttäytymistä.

Behaviorismi ajautuu kuitenkin suuriin vaikeuksiin. Ensinnäkin voidaanko löytää selvä käyttäytymisvastine kaikille mentaalisille tiloille. Onko esimerkiksi niin, että kaikki ihmiset käyttäytyvät ollessaan pelkotilassa samalla tavalla? Korkeamman tason mentaalisille tiloille on vielä vaikeampaa löytää vastinetta. Onko esimerkiksi olemassa yleistä kaikkien ihmisten noudattamaa käyttäytymisvastinetta uskomukselle, että suorakulmaisen kolmion hypotenuusan pituus on tangenttien neliöiden summan neliöjuuri?

Vaikka tällaiset ongelmat pystyttäisiinkin ratkaisemaan, suurin ongelma behaviorismissa on yksinkertaisesti se, että se ei hyväksy mentaalisten tilojen olemassaoloa. Jos vaikka lyön vasaralla sormeeni, on hyvin vaikea olla kieltämättä etteikö se tuntuisi miltään. On vaikea väittää, että vasaralla lyönti aiheuttaa vain huudon ja mustan sormen, eikä mitään sisäistä kivun tunnetta. Tästä syystä behaviorismi on nykyisin täysin hylätty teoria. Sillä on kuitenkin ollut suuri vaikutus nykyisiin teorioihin. Ja tekoälyn filosofiassa nimenomaan luvussa 2.3 esitellyssä Turingin testissä

näkyy suuresti behaviorismin vaikutus. Turingin testin mukaanhan peruste sille, että kone voisi olla tietoinen samalla tavalla kuin ihminen, on se, että se kykenee käyttäytymään kuten ihminen. Turingin testi ei ole siis behaviorismin tavoin lainkaan kiinnostunut koneen sisäisestä toteutuksesta. Toisin kuin behaviorismi, Turing ei kuitenkaan kieltänyt mentaalisten tilojen olemassaoloa, vaan hyväksyi ne täysin. Hänen perusteensa behavioristisen testin laatimiselle oli tieteellisen testin mahdottomuus.

3.4 Identiteettiteoria

Behaviorismin jälkeen seuraava merkittävä teoria mielen ja materian suhteesta oli identiteettiteoria. Se on yhä nykyään jokseenkin yleinen teoria. Osaltaan siitä syystä, että se on varsin yksinkertainen ja elegantti teoria mielen ja materian suhteesta. Ihmismieli ja tietoisuushan hyvin vahvoin tieteellisin todistein sijaitsevat aivoissa, toisin sanoen ne korreloivat aivojen fysikaalisten tilojen kanssa. Jos aivomme nimittäin vaurioituvat, sillä on myös tietoisuuteemme hyvin suuri vaikutus. Tästä ei tarvitse kiistellä. Vain hieman vahvempi teesi on se, että itseasiassa ihmismieli on yhtä kuin ihmisaivot. Aivot eivät siis aiheuta mentaalisuutta, vaan itseasiassa tietty mentaalinen tapahtuma on identtinen tietyn aivotapahtuman kanssa. Tämä on identiteettiteorian keskeinen teesi.

Jotta pääsemme sisälle tähän väitteeseen, muodostetaan ensin tarkemmin se hieman heikompi teesi, eli se, että mieli sijaitsee aivoissa, ts. mieli-aivot korrelaatio teesi: Jokaiselle mentaaliselle tapahtumatyypille M, joka ilmenee organismille o, voidaan löytää fysikaalinen tapahtumatyyppi F organismilla o, siten että M ilmenee organismille o hetkellä t, jos ja vain jos F ilmenee organismille o hetkellä t. (Kim, 2005, 82)

Jokaisen mentaalisen tapahtuman yhteydessä siis tapahtuu myös fysikaalinen tapahtuma. Ne ovat korrelaatiossa keskenään. Korrelaatio ei ole kuitenkaan aivan yksinkertainen ilmiö. Kaksi tapahtumaa voivat nimittäin korreloida keskenään monesta eri syystä. Ne voivat olla keskenään kausaalisessa vuorovaikutuksessa, jolloin jompi kumpi tai toinen aiheuttaa tapahtuman myös toisessa. Esimerkiksi syy miksi lämpötilan laskiessa nollaan myös vesi jäätyy, on se, että lämpötilan lasku yksinkertaisesti aiheuttaa veden jäätymisen. Lämpötilan lasku kausaalisesti vaikuttaa veden

mikrorakenteeseen, josta seuraa jäätyminen. Kausaalinen vuorovaikutus oli ajatus kartesiolaisessa dualismissa mieli-aivot korrelaation syynä. Tämä todettiin dualismin yhteydessä hyvin ongelmalliseksi. Mentaalisen vaikutus fysikaaliseen ei tuntunut mahdolliselta.

Jos kausaalinen linkki toimii vain toiseen suuntaan, puhutaan epifenomenalismista. Epifenomenalismi on jokseenkin yleinen teoria nykyään myös aivot-mieli korrelaation syynä, mutta sitä pidetään kuitenkin myös ongelmallisena. Sen mukaan siis aivot aiheuttavat mentaalisen, mutta mentaalisella ei voi olla mitään vaikutusta fysikaaliseen. Mentaalisuus on epifenomenalismin mukaan siis vain fysikaalisen tapahtuman sivutuote. Epifenomenalismin suurin ongelma on se, että se tekee ihmisen vapaan tahdon käsitteestä hyvin ongelmallisen.

Toinen syy kahden ilmiön korrelaatioon voi olla se, että niillä on jokin yhteinen kausaalinen tekijä menneisyydessä, joka saa ilmiöt korreloimaan keskenään. Esimerkiksi, miksi kellokaupassa kellot lyövät yhtä aikaa? Koska joku kausaalinen tekijä on ajastanut ne kaikki lyömään yhtä aikaa. Aivot-mieli korrelaation syynä tällainen vaihtoehto on täysin unohdettu vaihtoehto. Se on nykynäkemyksen mukaan täysin epäuskottava väite. Vain klassiset Leibnizin ja Malebranchen teoriat vuosisatojen takaa ehdottivat tämän tyyppisiä ratkaisuja. Näissä jumala toimi fysikaalisen ja mentaalisen "ajastajana".

Kolmas vaihtoehto on intuitiivisesti varsin epäuskottava, jos mietitään yleisellä tasolla miksi kaksi ilmiötä voisivat toistuvasti korreloida keskenään. Tämän vaihtoehdon mukaan voisi olla mahdollista, että korrelaatioon ei ole mitään syytä. Tämä ei kuitenkaan tunnu uskottavalta. Voin esimerkiksi ehkä uskoa, että kun jokin aamu herään ja aurinko paistaa silmiini, niin näillä tapahtumilla ei välttämättä ole mitään tekemistä keskenään. Se voi olla vain sattumaa. Mutta, jos joka ikinen aamu tämä sama toistuisi, niin voisi olla vaikea uskoa, että auringon valo ei vaikuttaisi mitenkään heräämiseeni. Mielenfilosofiassa tätä vaihtoehtoa kannattava teoria kantaa nimeä emergentismi. Sen mukaan mentaalisten ja aivotapahtumien korrelaatiolle ei ole olemassa mitään syytä tai selitystä. Se on vain raaka fakta, että näin tapahtuu. Myös tämä väite on jokseenkin

epäuskottava, mutta silti emergentismiä on kannatettu ja edelleen kannatetaan varsin paljon, ja tulemme palaamaan emergentismiin myöhemmin.

Neljäs ja viimeinen vaihtoehto kahden ilmiön korrelaation selittämiseksi on se, että mitään korrelaatiota ei itseasiassa olekaan, vaan kyse on vain yhdestä ainoasta ilmiöstä, ei kahdesta. Miksi salaman iskiessä tapahtuu myös sähköpurkaus maan ja pilvien välillä? Siksi, että on kyse yhdestä ja samasta ilmiöstä. Identiteettiteoria kannattaa juuri tätä vaihtoehtoa selittämään tai itseasiassa redusoimaan mielen ja aivojen välisen korrelaation. Kyse on samasta ilmiöstä. Kyse on reduktiivisesta fysikalismista juuri sen puhtaimmassa muodossa. (Kim, 2005, 85-88)

Kaikki aiemmat vaihtoehdot mielen ja materian suhteesta ovat olleet jokseenkin ongelmallisia. Identiteettiteoria sen sijaan on hyvin yksinkertainen ja elegantti teoria. Juuri yksinkertaisuus on ollut suurin syy identiteettiteorian suosioon. Miksi turhaan esittelisimme mentaalisia entiteettejä, jos pystymme täydellisesti selittämään maailman toiminnan pelkillä fysikaalisilla entiteeteillä. Mutta pystymmekö tähän todella?

Identiteettiteorian yhteydessä täytyy tarkentaa vielä identiteetin käsitettä. Tässä yhteydessä identiteetillä tarkoitetaan tiukkaa identiteettiä, joka määritellään identtisten erottamattomuus -periaatteessa (myös Leibnizin laiksi kutsuttu): Jos kaksi objektia ovat identtisiä, niin mikä tahansa ominaisuus, joka on toisella objektilla, on myös toisella. Ja jos kahdella objektilla on kaikki samat ominaisuudet, tällöin on siis kyse tietenkin vain yhdestä objektista. Nyt identiteettiteorian vastustajan tulee vain löytää yksi mikä tahansa ominaisuus, joka on mentaalisuudella, mutta ei fysikaalisella, tai päinvastoin. Tämä osoittaisi, ettei mentaalinen voi olla identtinen fysikaalisen kanssa. (Kim, 2005, 101)

Tällaiseksi ominaisuudeksi on ehdoteltu pitkälti samoja ominaisuuksia, joita on käytetty argumenteissa substanssidualismin puolesta (ks. 3.2). Erityisesti epistemologisia ominaisuuksia. Tietomme mentaalisista ominaisuuksista on hyvin erilaista, kuin fysikaalisista ominaisuuksista. Tietomme mentaalisista tiloistamme on suoraa ja välitöntä, kun taas tietomme fysikaalisesta on epäsuoraa, aistiemme kautta välittynyttä, joten mentaalinen ei voi olla identtinen fysikaalisen kanssa. On kuitenkin epäilty

voidaanko tällaisia epistemologisia ominaisuuksia pitää sellaisina ominaisuuksina, että ne voisivat estää identiteetin olemassaoloa.

Tärkein argumentti identiteettiteoriaa vastaan on kuitenkin moninainen realisoitavuus -argumentti, jonka puolesta muun muassa Hilary Putnam (1967) kirjoitti. Identiteettiteorian mukaan tietty mentaalinen tila vastaa yhtä ainoaa tiettyä fysikaalista tilaa. Ei tunnu kuitenkaan luonnolliselta, että esimerkiksi kaikilla ihmisillä kivun tunteen aiheuttaisi juuri samanlainen fysikaalinen tila, tai että uskomus "Tarja Halonen on tasavallan presidentti" aiheuttaa kaikilla täsmälleen samanlaisen fysikaalisen tilan ihmisaivoissa. Tämän lisäksi identiteettiteoria sulkee pois mentaalisen kaikilta olioilta, joilla aivojen rakenne ei vastaa ihmisaivoja. Esimerkiksi monilla eläimillä aivojen organisaatio on hyvin erilainen kuin ihmisillä, silti voisi kuvitella, että myös eläimet ovat kykeneviä tuntemaan kipua. Identiteettiteoria kieltää mentaalisen myös sellaisilta olioilta, kuten "avaruusolioilta", joilla aivot kenties olisivat jotain täysin muuta materiaalia kuin hiiliyhdisteistä orgaanista materiaa, mistä ihmisaivot ovat rakentuneet. Tästä seuraa se, että teoria kieltää mentaalisen myös koneilta. Kone, jolle kyettäisiin luomaan ihmisälyn kaltainen tekoäly, ei identiteettiteorian mukaan ole kykenevä tietoisuuteen. Identiteettiteoria ei voi siis toimia pohjana näkemykselle, että tekoälytutkimus voisi jonakin päivänä luoda ihmisen kaltaisen tietoisuuden myös koneille.

Tämän moninainen realisoitavuus -argumentin myötä identiteettiteorian valta-asema kuihtui 1960-luvulla, koska koettiin ongelmalliseksi se, että identiteettiteoria rajoitti tietoisuuden vain hyvin ihmisenkaltaisille olioille. Tämän argumentin pohjalta lähtikin näkemys, joka yhä nykyään on suosituin teoria mielen ja materian suhteesta. Se hyväksyy mentaalisen myös ns. "avaruusolioille", sekä roboteille toisin kuin identiteettiteoria. Tämä näkemys kantaa nimeä funktionalismi.

3.5 Funktionalismi

Funktionalismin esitti ensimmäisen kerran Putnam samaisessa yllämainitussa artikkelissaan, missä hän torjui identiteettiteorian. Toki funktionalismin piirteitä omaavia ajatuksia on ollut aikaisemminkin olemassa (esim. Turing), mutta Putnamin

artikkeli oli se, joka mullisti mielenfilosofian antamalla funktionalismille formaalin esityksen. Putnamin versio funktionalismista on juuri se, jota käytännössä koko tämä tutkielma tulee tutkimaan. Sitä voidaan kutsua konefunktionalismiksi tai komputationalismiksi. Seuraavassa luvussa perehdymme tarkemmin juuri tähän teoriaan. Sitä ennen esittelen kuitenkin yleisiä piirteitä funktionalismista, jonka alalaji komputationalismi siis on.

Funktionalismi on Putnamin artikkelista lähtien ollut suosituin teoria mielenfilosofiassa. Sen perusidea on seuraavanlainen: Ajatellaan esimerkiksi, että luokittelemme erilaisia fysikaalisia objekteja. Millainen objekti esimerkiksi lasketaan kuuluvaksi luokkaan kello? Emme ole niinkään kiinnostuneita siitä minkä näköinen tuo objekti on, emmekä siitä, mistä materiaalista se on valmistettu. Olemme kiinnostuneita vain siitä, että objekti toimii halutussa tehtävässään. Jos se toteuttaa halutun funktion, olemme valmis luokittelemaan sen kyseiseen luokkaan.

Funktionalismissa "luokitellaan" mentaalisia ominaisuuksia samalla tavoin. Ajatellaan esimerkiksi kipua. Emme ole kiinnostuneita siitä, millainen prosessi toteuttaa funktion, joka kivulla on. Se voi realisoitua ihmisaivoissa tiettyjen synapsien aktiivisuutena, tai se voi realisoitua tietokoneen prosessorissa. Tärkeää on vain, että prosessi toteuttaa kivun funktionaalisen määritelmän, joka voisi korkealla tasolla olla seuraavankaltainen: Jos aistimme huomaa elimistössämme sairauden, kipu on se funktio tai kausaalinen tekijä, joka saa elimistömme toimimaan siten, että se saisi häädettyä sairauden ruumiistamme.

Funktionalismilla on siis myös tietty yhteys behaviorismiin. Se on tietyllä tapaa kehittyneempi versio siitä. Behaviorismissa mentaalisuus identifioidaan käyttäytymisen kanssa. Funktionalismissa käyttäytymisellä on myös tärkeä rooli, mutta se on vain jälkimmäinen puolisko mentaalisesta funktiosta. Stimulaatio on se ensimmäinen puolisko. Funktionalismissa mentaalinen ominaisuus identifioidaan siis funktionaaliseen ominaisuuteen (tai tarkemmin sanottuna sille annetaan funktionaalinen määritelmä), sitä ei identifioida käyttäytymiseen behaviorismin tavoin. Ja muistutuksena, identiteettiteoriassahan mentaalisuus identifioitiin fysikaalisen tapahtuman kanssa.

Funktionalismissa ja behaviorismissa on kuitenkin hyvin merkittäviä eroja. Ensinnäkin toisin kuin behaviorismi, funktionalismi tunnustaa mentaaliset ominaisuudet todellisina sisäisinä tiloina, siten että niillä voi olla myös kausaalisia voimia. Behaviorismihan ei tunnustanut sisäisiä mentaalisia tiloja, vaan hyväksyi vain objektiivisesti havainnoitavat entiteetit. Toinen merkittävä ero, joka seuraa edellisestä, on se, että behaviorismi ei hyväksy, että jokin "mentaalinen" tila voisi aiheuttaa toisen "mentaalisen" tilan. Funktionalismissa sen sijaan, kuten myös arkijärjen mukaan, pidetään täysin normaalina, että yleensä ajatukset saavat meissä aikaiseksi muita ajatuksia. (Kim, 2005, 122-123)

Behaviorismi ja identiteettiteoria olivat reduktionistisia teorioita. Niissä mentaalinen pyrittiin redusoimaan tai eliminoimaan fysikaalisten ominaisuuksien avulla. Funktionalismi sen sijaan on yleensä ei-reduktiivista fysikalismia. Ei-reduktionismissa mentaalinen siis ajatellaan ominaisuudeksi, joka ei ole redusoitavissa fysikaalisiin ominaisuuksiin, toisin sanoen mentaalista ei voida selittää pelkästään fysikaalisilla käsitteillä. Mikä tähän on sitten syynä? Miksi funktionalismi ei ole reduktionismia, ainakaan välttämättä?

Funktionalismissa jokin fysikaalinen systeemi toteuttaa tietyn mentaalisen ominaisuuden, jos tämä systeemi realisoi tietyn funktionaalisen kuvauksen. Millainen tuo tietty funktionaalinen kuvaus sitten on? Se on määritelmällisesti abstrakti kuvaus tuosta mentaalisesta ominaisuudesta, siis abstrakti entiteetti. Funktionaalinen kuvaus ei siis ole fysikaalinen entiteetti, se ei ole riippuvainen mistään tietystä fysikaalisesta ominaisuudesta. Riittää kun jokin reaalinen systeemi toteuttaa halutun funktion, niin syntyy myös tuo mentaalinen ominaisuus. Teoriassa edes tuon realisoivan systeemin ei tarvitse olla fysikaalinen, kunhan jonkinlainen kausaalisia voimia omaava entiteetti toteuttaa funktion. Huvittavin esimerkki tästä tulee Aaron Slomanilta (1986), jonka mukaan tällaisen systeemin voisi toteuttaa vaikka parvi immateriaalisia enkeleitä tanssimassa nuppineulan päässä. Funktionalismi on siis yhteensopiva myös dualismin kanssa. Näin ollen funktionalismi ei yleensä ole reduktiivista fysikalismia, vaan ei-reduktiivista fysikalismia. Funktionalismin myötä reduktiivinen fysikalismi on menettänyt suuresti suosiotaan mielenfilosofiassa ja ei-reduktiivisesta on tullut hallitsevampi näkemys.

Funktionalismin oivallus oli siis se, että mentaalisuus voisi olla riippuvainen vain abstraktista funktiosta, jonka taas mikä tahansa systeemi voi toteuttaa. Tämä avaa myös sen mahdollisuuden, että tietokone, joka luonnostaan on väline, joka voidaan ohjelmoida toteuttamaan haluttu funktio, voisi olla aidosti tietoinen.

Funktionalismikaan ei tietenkään ole täysin ongelmaton, päinvastoin. Sitä vastaan on olemassa paljon argumentteja, joista monet erittäin haastavia. Seuraavissa luvuissa lähden tarkemmin tutkimaan funktionalismia ja erityisesti sen tietokoneille sovitettua versiota, komputationalismia. Sen erityispiirteitä, ongelmia, sekä ratkaisuehdotusta tietoisen koneen ongelmaan. Komputationalismi erottuu muusta funktionalismista esittelemällä mielen ja materian tutkimisen apuvälineeksi aivan erityisen laitteen: Turingin koneen.

4 KOMPUTATIONALISMI

4.1 Turingin kone

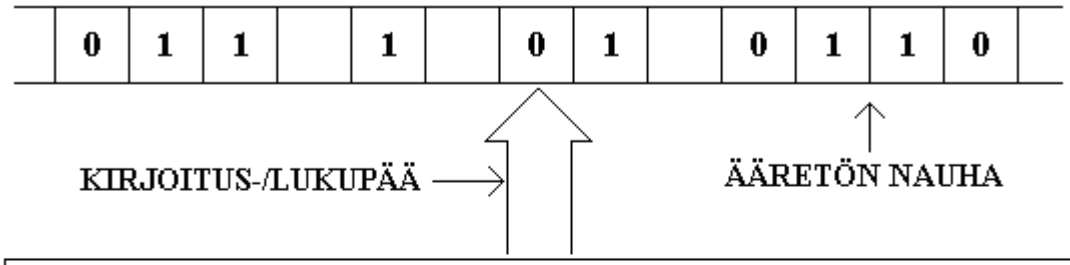
Kun kysytään voiko kone olla tietoinen, täytyy tehdä muutama huomio. Mikä on kone? Koneella tarkoitetaan usein jotain mekaanista systeemiä, jonka ihminen on rakentanut. Kuitenkaan sillä, että jokin systeemi on ihmisen tekemä, ei pitäisi olla mitään merkitystä sen luonteeseen olla kone. Jos jossain päin universumia kasvaa televisioita puissa, lasketaan nämä luultavasti kuitenkin koneiksi. Niinpä koneitten määritelmää voidaan joutua laajentamaan koskemaan kaikkia kausaalisia fysikaalisia systeemeitä (Harnad, 2003, 69). Näin ollen myös kaikki biologiset organismit, mukaan lukien ihmiset, voidaan ajatella koneiksi. Tällöin kysymys tietoisista koneista muuttuu kuitenkin triviaaliksi, kuten myös kysymys, voiko ihminen tehdä tietoisia koneita. Kun kysymme, onko komputationalismi totta, tai voiko tietokone olla tietoinen, puhumme kuitenkin eri asiasta. Tällöin puhumme siitä, voiko juuri tietynlainen, komputationaalinen kone olla tietoinen.

Putnam ehdotti artikkelissaan (1967) erityistä konetta malliksi funktionalismilleen. Tämä on se kone, josta me olemme kiinnostuneita. Alan Turing oli määritellyt kyseisen koneen vuonna 1936 artikkelissaan On Computable Numbers with an Application to the Entscheidungsproblem. Siinä hän oli kuvannut abstrakteja koneita, joiden erityispiirre on se, että niillä voidaan suorittaa mikä tahansa laskennallinen funktio tai matemaattinen algoritmi. Näitä on nimitetty myöhemmin Turingin koneiksi. Ne ovat matemaattisesti idealisoituja koneita, ja Turing (1969) kutsuikin niitä loogisiksi laskentakoneiksi (logical computing machines) erotukseksi konkreettisista digitaalisista tietokoneista. Jos jonkin ongelman ratkaisemiseksi on siis olemassa matemaattinen algoritmi, Turingin koneella voidaan suorittaa tämä algoritmi.

Yleisesti hyväksyttyä väitettä, että Turingin koneella on edellä kuvattu ominaisuus, kutsutaan Church-Turingin teesiksi. Se voidaan muodollisemmin kuvata seuraavasti: laskennallisia funktioita ovat kaikki ja ainoastaan ne funktiot, jotka ovat kuvattavissa Turingin koneina (Cotogno, 2003, 184). Sitä ei ole mahdollista osoittaa todeksi, mutta monet asiat tukevat sitä. Ensinnäkin kaikki tunnetut algoritmit ovat kuvattavissa

Turingin koneina, toisekseen monet muut itsenäisesti kehitetyt kuvaukset laskennallisista funktioista, kuten Gödelin (1934) ja Kleenen (1936) rekursiivinen algoritmi ja Churchin (1936) lambdakalkyyli antavat saman funktioiden joukon kuin Turingin kuvaus (Hauser, 1993). Lisäksi se on intuitiivisesti uskottava väite.

Turingin koneet ovat diskreettejä tilakoneita. Nämä ovat koneita, jotka siirtyvät diskreetein askelin tilasta toiseen. Turingin kone koostuu nauhasta, jonka pituutta voidaan tarvittaessa laajentaa rajattomasti, kirjoitus- ja lukupäästä sekä tilataulukosta, joka kertoo yksiselitteisesti, mitä kirjoitus-/lukupään tulee tehdä missäkin tilassa, riippuen syöttötiedosta (ks. Kuva 1). Kirjoitus-/lukupää lukee nauhalta syöttötiedot ja kirjoittaa nauhalle välitulokset ja lopulta lopputuloksen tilataulukon sääntöjen mukaisesti. Kaikki tämä data koostuu ykkösistä ja nollista ja mahdollisista tyhjistä väleistä. Tarvitaan siis sopiva tilataulukko (säännöstö) sekä alkuarvot, jotta Turingin kone voi suorittaa laskennan. Oletusarvoisesti eri tehtävien ratkaisut tarvitsevat omat tilataulukkonsa. Turing (1936) kuitenkin osoitti, että on olemassa Turingin koneita, jotka pystyvät yksittäin suorittamaan minkä tahansa laskennallisen funktion. Ne pystyvät siis imitoimaan minkä tahansa diskreetin tilakoneen toimintaa. Näitä kutsutaan universaaleiksi Turingin koneiksi. Tällä on se merkittävä seuraus, että eri tehtäviä varten ei tarvitse suunnitella erillisiä koneita, vaan sama kone voidaan ohjelmoida suorittamaan kaikki tehtävät.



KONTROLLERI

Tilataulukko

	Nykyinen tila	B

	Tila A:			Tila B:			...	Tila N:		
Luettu symboli	Kirjoita symboli	Liikuta nauhaa	Seuraava tila	Kirjoita symboli	Liikuta nauhaa	Seuraava tila	... ...	Kirjoita symboli	Liikuta nauhaa	Seuraava tila
tyhjä	1	Oikea	A	0	Vasen	D	...	0	Oikea	D
0	0	Vasen	B		Vasen	E	...	0	Vasen	B
1	1	Oikea	C	1	Oikea	F	...	1	Vasen	C

Kuva 1. Esimerkki Turingin koneesta.

Turingin koneet ovat siis abstrakteja, idealisoituja koneita. Tämä on funktionalismille olennainen asia. Siinähän mentaalisuus selitettiin juuri abstraktien funktioiden avulla. Abstraktisuus ja ideaalisuus näkyvät Turingin koneissa usealla tavalla. Ensinnäkin niissä on periaatteessa ääretön nauha, eli muisti. Tällaista ei tietenkään todellisuudessa ole mahdollisuutta realisoida, mutta käytännössä oikeissa koneissa äärellinen nauha käy aivan yhtä hyvin, koska ääretöntä muistia vaativaa ohjelmaa ei ole mielekästä toteuttaa. Toisekseen, Turingin koneet ovat diskreettejä tilakoneita. Tarkkaan ottaen sellaisiakaan koneita ei voida realisoida, ainakaan nykyisin keinoin, koska todellisuudessa nykyisen käsityksen mukaan kaikki liikkuu fysikaalisessa maailmassa ei-diskreetisti, lukuunottamatta ehkä joitain kvanttitason ilmiöitä. Toki on olemassa monia koneita, jotka voidaan hyvin ajatella diskreeteiksi tilakoneiksi, kuten valokatkaisinta, jolla on kaksi erillistä tilaa: päällä ja poissa (Turing, 1950), aidosti diskreettejä ne eivät kuitenkaan ole. Kolmanneksi, Turingin koneet ovat abstrakteja koneita. Ne voidaan realisoida lähes äärettömän monella tapaa. Se onko ykköset ja nollat realisoitu 5 ja 0 voltin jännitteillä, vai päivänkakkaroina ja voikukkina, on merkityksetöntä. Tämä ominaisuus on se tärkein asia funktionalismille. Moninainen realisoitavuus -argumenttihan vaatii, että mentaalinen funktio voidaan realisoida monin eri tavoin. Kuitenkin juuri se, että mentaalinen funktio voidaan realisoida vaikka päivänkakkaroina ja voikukkina, on nostanut tekoälyn filosofiasta keskusteltaessa paljon mielipiteitä ja tulemme palaamaan tähän aiheeseen myöhemmin.

Joka tapauksessa, vaikka moderni tietokone on sisäiseltä rakenteeltaan hyvin erilainen verrattuna idealisoituun Turingin koneeseen, voidaan käytännössä katsoa kaikkien nykyisten digitaalisten tietokoneiden kykenevän realisoimaan universaalin Turingin koneen.

Turingin kone on abstraktisuutensa ja universalisuutensa myötä sopiva malli funktionalistiselle mielelle. Abstraktisuus tukee sitä, että koneen voi toteuttaa mikä tahansa oikean rakenteen omaava kausaalinen systeemi. Universaalisuus taas mahdollistaa sen, että kone voi toteuttaa minkä tahansa funktion, myös oletetun mentaalisen funktion. Komputationalismin pääteesi voidaankin lausua näin: edellytys, että jollakin systeemillä voisi olla mentaalisia tiloja on se, että sen tulee olla fysikaalisesti realisoitu riittävän kompleksisuuden omaava Turingin kone. Mentaaliset

tilat voidaan tällöin identifioida tilataulukon sisäisiin tiloihin. Eli ihmisen mentaaliset tilat ovat identtisiä tietyn Turingin koneen sisäisen tilan kanssa. Oleellista ihmisen tietoisuudelle ei siis ole se, että ihmisaivot, jotka tuottavat tietoisuuden, ovat monimutkainen biologinen organismi, koostuen massiivisesta määrästä neuroneita, joiden välillä liikennöi monimutkainen sähköimpulssiverkko. Oleellista on se, että aivot ovat Turingin kone.

4.2 Turingin kone -mallin ongelmia

Komputationalismia voidaan kuitenkin helposti syyttää samasta asiasta kuin identiteettiteoriaa. Identiteettiteorian mukaanhan kaksi systeemiä voivat olla samassa mentaalisessa tilassa vain jos ne ovat samassa fysikaalisessa tilassa. Naiivin komputationalismin mukaan taas kaksi systeemiä ovat samassa mentaalisessa tilassa, jos ne ovat samassa Turingin koneen tilataulukon tilassa. Mutta mitä tämä käytännössä tarkoittaa? Erilaisia Turingin koneita voi olla ääretön määrä. Eri Turingin koneiden tilojen vertailu on selvästikin mahdotonta, joten jotta kaksi systeemiä voisivat olla samassa mentaalisessa tilassa, niiden täytyisi toteuttaa täysin sama Turingin kone, ja olla lisäksi kyseisen Turingin koneen tilataulukon mukaisesti samassa tilassa. Kuitenkaan esimerkiksi kahden ihmisen aivot eivät hyvin suurella todennäköisyydellä koskaan toteuta juuri samaa Turingin konetta. Tämä tarkoittaa sitä, että kaksi ihmistä eivät koskaan voisi olla samassa mentaalisessa tilassa. Ja juuri tämä oli identiteettiteorian suurin ongelma. (Kim, 2005, 138)

Ongelma on kuitenkin siinä, että systeemin tila ajatellaan yhdeksi monadiseksi tai holistiseksi tilaksi. Tämä tarkoittaa sitä, että perinteinen tilakone voi olla vain yhdessä tilassa kerrallaan. Ongelmaa voidaan kiertää laajentamalla tilakoneen määritelmää siten, että tilakoneen kokonaistila voi koostua useista osatiloista, jolloin kokonaistila on näiden tilojen konjunktio. Tällöin voimme väittää, että siihen, että kaksi systeemiä ovat samassa mentaalisessa tilassa, riittää jos vain osa niiden funktionaalisesta organisaatiosta toteuttaa tietyn Turingin koneen, joka on kyseisessä tilassa. Systeemien ei tarvitse siis olla täysin identtisiä rakenteeltaan, eikä samassa tilassa, vaan riittää että vain osa niiden funktionaalisesta organisaatiosta toteuttaa kyseisen mentaalisen Turingin kone -tilan. Tällöin siis sama systeemi voi olla yhtä aikaa sekä useassa

abstraktissa Turingin koneen tilassa, että myös useassa mentaalisessa tilassa, joka tuntuu esimerkiksi ihmisten tapauksessa luonnolliselta. Ei ole kuitenkaan täysin selvää, voidaanko ihmismielen tilat rajata tällä tavoin pieniin moduuleihin, vai onko mielen holistisuus redusoimattomissa. (Kim, 2005, 141)

Naiivin Turingin kone -mallin vajaavuus voidaan huomata kuitenkin myös seuraavassa ongelmassa. Moninainen realisoitavuus -periaate voidaan nähdä nimittäin myös yhden ja saman Turingin koneen sisällä. Jos ajatellaan esimerkiksi tavallista tiedonkäsittelyongelmaa, lajittelua. Tiedämme, että lajittelu voidaan toteuttaa useilla eri algoritmeilla, esimerkiksi lisäys-, limitys- tai pikalajittelua käyttäen. Ne ovat kaikki eri funktioita, vaikka toteuttavatkin saman syöte/tulos -käyttäytymisen. Naiivin komputationalismin mukaan vaikka kaksi eri funktiota tuottaisivat saman syöte/tulos -käyttäytymisen, ne eivät voisi silti toteuttaa samaa mentaalista ominaisuutta, koska niiden sisäiset tilat ovat erilaisia. Tätä emme tietenkään halua.

Edellä kuvatut ongelmat ovat vain osa monista ongelmista, joita Turingin kone -mallilla on mielen selittämisessä. Ja ehkä kaikkein yleisin ongelma, joka heti huomataan vertailtaessa Turingin konetta ihmisaivoihin, on se, että Turingin kone on määritelmällisesti yksiprosessorinen sarjakäsittelynä toimiva yhden tilan kone. Sen sijaan ihmisaivojen miljoonat neuronit ovat kaikki itsenäisiä rinnakkain toimivia prosessoreja, ja näin ollen ne käsittelevät tietoa massiivisesti rinnakkain. Tämä on usein koettu ongelmalliseksi, jotta Turingin kone -malli voisi selittää tietoisuuden.

Turingin koneella on kuitenkin se etu, että se pystyy käytännössä simuloimaan täydellisesti myös rinnakkaislaskentaa. Sillä koska Turingin kone voi ratkaista minkä tahansa algoritmisesti ratkeavan ongelman, voidaan se myös ohjelmoida mallintamaan useaa prosessoria ja näiden välistä tiedonsiirtoa. Näin ollen esimerkiksi neuronien toiminta voidaan simuloida myös yksiprosessorisella koneella, jolloin neuronit ajatellaan itsenäisiksi *virtuaalikoneiksi* Turingin kone -mallin sisällä. Tällöin siis ulkoapäin katsottuna on vain yksi prosessori, joka suorittaa kyseistä ohjelmaa. Kuitenkin Turingin kone -mallin sisältä katsottuna Turingin kone sisältää useita muita prosessoreita, jotka siis ovat mallin sisäisiä Turingin koneita. Moninainen realisoitavuus -periaatteen mukaisestihan ei ole merkitystä, mikä kausaalinen systeemi toteuttaa

Turingin koneen: yleensä se on jokin fysikaalinen systeemi, mutta se voi yhtä hyvin olla myös toinen abstrakti Turingin kone. Tästä seuraa se, että periaatteessa yksi Turingin kone voi toteuttaa aidosti myös useita Turingin koneita ja täten rinnakkaislaskentaa, samalla tavoin kuin aivojen neuronit toteuttavat useita Turingin koneita.

Turingin kone usein ajatellaan kuitenkin jokseenkin rajoittuneeksi, joten mentaalisen malliksi on otettu usein muita matemaattisia abstraktioita, jotka eivät ole yhtä rajoittuneita kuin Turingin kone. Esimerkiksi neuroverkot ovat nykyään usein käytetty malli mentaalisen kuvaukseksi. En kuitenkaan nyt lähde tarkemmin analysoimaan eri matemaattisia malleja tietoisuudelle, vaan jatkossa, jollen toisin mainitse, teen hyvin yleisen oletuksen, että komputationalismissa mallin mentaaliselle ei tarvitse olla Turingin kone, vaan riittää, että se on jokin matemaattisesti määriteltävissä oleva formalismi. Sillä ajatellaan hetki miten komputationalismi suhteutuu funktionalismiin. Funktionalismissa idea oli, että fysikaalinen systeemi toteuttaa mentaalisen ominaisuuden, jos se toteuttaa funktionaalisen kuvauksen tuosta mentaalisesta ominaisuudesta. Turingin kone on siis vain yksi, varsin yksinkertainen, laite, jolla tuo funktionaalinen kuvaus on mahdollista formalisoida. Formalismin ei kuitenkaan funktionalismin mukaan tarvitse olla Turingin kone, vaan se voi olla mikä tahansa formalismi, joka kykenee tuon kuvauksen esittämään.

4.3 Komputaatio

Komputationalismin perusidea on siis se, että mentaalinen tila on jokin fysikaalinen tila, joka toteuttaa jonkin matemaattisen formalismin avulla kuvatun funktionaalisen ominaisuuden. Jos taas pohdimme mistä komputationalismi on saanut nimensä, se tulee selvästi sanasta komputaatio. Ja usein käytetäänkin komputationalismin luonnehdinnassa väitettä, että sen mukaan mieli on komputaatiota. Mutta mitä sitten on komputaatio? Komputaatio on jälleen yksi käsite, jolle löytyy useita erilaisia määritelmiä. Yleensä komputaatio voidaan määritellään väljästi näin: komputaatio on abstraktien symbolien formaalia manipulointia formaalien sääntöjen mukaisesti (Boden, 1988). Tarkempi kuvaus voisi olla, että komputaatioita ovat ne operaatiot, joita Turingin kone suorittaa. Mutta kuten huomaamme edellisen kappaleen perusteella, tämä määritelmä rajaa ulkopuolelle myös muita operaatioita, jotka yleensä ajatellaan

komputaatioiksi. Joten väljempi määritelmä riittäköön meille. Yleisesti ottaen komputaatio kuitenkin yleensä määritellään suhteessa johonkin formalismiin, oli se sitten Turingin kone, Pascalin ohjelma, soluautomaatti, tai neuroverkko (Chalmers, 1994).

Yksi komputaatio on siis yksi sellainen abstrakti operaatio, jonka abstrakti matemaattinen formalismi suorittaa. Aiemmin olen puhunut pelkästään systeemin abstrakteista tiloista. Ne liittyvät kuitenkin hyvin tiukasti toisiinsa. Komputaatio on se operaatio, jolla abstrakti tilakone liikkuu tilasta toiseen, ja tilat ovat niitä välittäjiä, jotka ohjaavat abstraktien operaatioiden suorittamista. Ne ovat ikään kuin saman asian kaksi puolta. Tutkielman alussa puhuin siitä miten tietoisuus voidaan nähdä joko prosessina, kuten ajattelu yleensä nähdään, tai tilana, kuten esimerkiksi mentaalinen tila kipu. Tämä sama asia voidaan nähdä analogiana komputaation ja abstraktien tilojen kanssa. Käytänkin seuraavassa kappaleessa komputaatio -termiä, vaikka sama asia voitaisiin kääntää myös lauseisiin, joissa käytettäisiin abstrakteja tiloja komputaation sijasta.

4.4 Komputaation realisaatio

Komputationalismi väittää siis, että tietoisuus on komputaatiota. Komputaatio on määritelmän mukaisesti abstrakti operaatio. Tietoisista systeemeistä puhuttaessa on kuitenkin kyse fysikaalisista systeemeistä, jotka suorittavat fysikaalisia operaatioita. Milloin fysikaalinen systeemi voi siis tarkasti ottaen suorittaa komputaatiota tai toteuttaa tietyn formaalin tilan? Ei riitä, että vain sanotaan, että sama komputaatio voidaan suorittaa äärettömän monella tapaa. Komputationalismin täytyy myös kyetä määrittelemään linkki, joka yhdistää abstraktit operaatiot fysikaalisiin operaatioihin. Moninainen realisoitavuus -periaatteen hyväksyminen edellyttää siis, että täytyy kyetä luomaan teoria myös siitä milloin tietty fysikaalinen systeemi implementoi tietyn komputaation. Tämä ei ole aivan ongelmaton tehtävä.

Kuten on todettu, ei ole merkitystä sillä miten Turingin koneen ykköset ja nollat ovat realisoitu. Komputaatio voidaan realisoida esimerkiksi kuten tavallisissa digitaalisissa tietokoneissa sähkövirtojen ja transistorien avulla, tai sitten komputaatio voidaan realisoida mitä mielikuvituksellisimmilla tavoilla. Esimerkiksi Pylyshyn sanoo, että

komputationaalinen prosessi voitaisiin realisoida kouluttamalla joukko puluja nokkimaan Turingin koneen sääntöjen mukaisesti (Pylyshyn, 1985, 57).

Näiden näkemysten taustalla on perinteinen näkemys siitä, milloin fysikaalinen systeemi implementoi tietyn komputaation. Se voidaan ilmaista lyhyesti näin: Fysikaalinen systeemi implementoi annetun komputaation silloin, kun fysikaalisen systeemin kausaalinen rakenne heijastaa annetun komputaation formaalia rakennetta. Tai tarkemmin: Fysikaalinen systeemi implementoi annetun komputaation silloin kun (1) systeemin fysikaaliset tilat ovat jaettavissa tilatyyppeihin ja (2) on olemassa yhden suhde yhteen kuvaus komputaation formaaleista tiloista fysikaalisiin tilatyyppeihin siten, että formaalien tilojen väliset abstraktit tilasiirtymäsuhteet ovat kuvattavissa fysikaalisten tilatyyppien välisiksi kausaalisiksi tilasiirtymäsuhteiksi. (Chalmers, 1994)

Tästä seuraa kuitenkin ongelmia. Vanhemmalla iällään komputationalismin vastustajaksi kääntynyt Putnam (1988) ja John Searle (1990) ovat nimittäin väittäneet, että tällaisen näkemyksen perusteella mikä tahansa fysikaalinen systeemi voidaan katsoa implementoivan minkä tahansa komputaation, jos systeemiä tulkitaan sopivasti. Searle esimerkiksi väittää, että hänen takanaan oleva seinä implementoi parhaillaan Wordstar-ohjelman. Tämän hän perustelee sillä, että seinästä on löydettävissä molekyylien liikesarja, joka on isomorfinen Wordstar-ohjelman formaalin rakenteen kanssa, ja näin ollen toteuttaa ylläolevat ehdot komputaation implementaatiolle. Tähän Searle vielä lisää, että tarpeeksi suuri seinä siis implementoi minkä tahansa ohjelman, mukaanlukien sen ohjelman, jonka aivoni oletetaan suorittavan. Jos Searlen ja Putnamin väitteet ovat tosia, meidän täytyisi siis hyväksyä äärimmäinen panpsykismin muoto, joka väittää, että kaikki fysikaaliset systeemit ovat tietoisia, tai sitten meidän pitäisi hylätä ylläoleva yleinen ehto komputaation implementaatiosta ja määritellä tiukemmat ehdot sille.

Tarkastelen seuraavaksi ehdotusta, jonka Chalmers (1994) antaa tämän ongelman ratkaisuksi. Hän on samaa mieltä kanssamme siitä, että Turingin koneen tapainen yksinkertainen äärellinen tilakone -malli on riittämätön selittämään mentaalisuutta sen monadisten tilojen vuoksi. Chalmersin mukaan useimmilla komputationaalisilla formalismeilla on nimenomaan kombinatorinen rakenne. Tällä tarkoitettiin sitä, että

niiden tila on usean alitilan yhdistelmä. Esimerkkinä voidaan ajatella vaikka tavallista digitaalista tietokonetta: se ei ole vain yhdessä tilassa S, vaan sen kokonaistila koostuu hyvin monien elementtien, kuten muistien, rekistereiden yms. tilojen kombinaatiosta. Chalmers määritteleekin kombinatorisen tilakoneen: se eroaa äärellisestä tilakoneesta vain siten, että siinä systeemin tila ei ole yksilöity monadisella symbolilla S, vaan vektorilla [S_1, S_2, S_3, ...]. Tämän vektorin elementit voidaan ajatella kokonaistilan komponenteiksi. Nyt fysikaalinen systeemi implementoi annetun kombinatorisen tilakoneen, jos (1) on olemassa vektorisaatio fysikaalisen systeemin tiloista ja (2) on olemassa kuvaus tämän vektorin elementeistä vastaaviin kombinatorisen tilakoneen vektorin elementteihin, siten että tilasiirtymäsuhteet ovat isomorfisia ilmeisellä tavalla.

Chalmers myös huomauttaa, että tilasiirtymien täytyy olla luotettavia: tilojen välillä täytyy olla kontrafaktuaaleja tukeva side. Jos on annettu formaali tilasiirtymä A->B, täytyy olla niin, että jos fysikaalinen systeemi on tilassa A, niin sen täytyy myös luotettavasti siirtyä tilaan B. Ja tämän konditionaalin täytyy päteä kaikkiin siirtymiin tilataulukossa, ei vaan niihin, joihin tilakone tietyn ajan kuluessa joutuu. Ajatellaan, että tilataulukossa olisi siirtymä Q->R, mutta tilakone ei koskaan saavuttaisi tilaa Q, tästä huolimatta täytyisi päteä, että jos fysikaalinen systeemi olisikin joutunut tilaan Q, se olisi siirtynyt seuraavaksi tilaan R.

Kuten Chalmers huomauttaa, voisi luulla, että tilakoneen vaihtaminen kombinatoriseen tilakoneeseen ei tuo juuri mitään apua implementaation ongelmaan. Ainakaan se ei ole laskennallisesti yhtään vahvempi kuin äärellinen tilakonekaan. Sillä on kuitenkin ainakin yksi suuri etu: implementaatioehdot ovat paljon tiukemmat kuin äärellisen tilakoneen ehdot. Jotta fysikaalinen systeemi implementoisi kombinatorisen tilakoneen, vaaditaan että se toteuttaa monimutkaisia kausaalisia interaktiota useiden itsenäisten osiensa välillä. Kombinatorinen tilakone pystyy siis kuvaamaan systeemin kausaalisen organisaation huomattavasti suuremmalla tarkkuudella kuin äärellinen tilakone.

Näiden määritelmien pohjalta Chalmers pyrkii nyt vastaamaan kysymyksiin komputaation implementaatiosta:

(1) Implementoiko jokainen systeemi jonkun komputaation? Kyllä, esimerkiksi jokainen fysikaalinen systeemi implementoi yksinkertaisen äärellisen tilakoneen, jolla on yksi sisäinen tila.

(2) Implementoiko jokainen systeemi minkä tahansa komputaation? Ei. Otetaan esimerkiksi kombinatorinen tilakone, jonka tilavektori koostuu 1000 elementistä, jotka kaikki voivat olla 10 eri tilassa, ja näiden elementtien välille vielä monimutkaiset tilasiirtymät. Tällaisella systeemillä on tarpeeksi monimutkaiset ehdot, että äärimmäisen harvat fysikaaliset systeemit läpäisevät ne. Vaatimus, että tilasiirtymien täytyy olla luotettavat on ratkaiseva tekijä, joka rajoittaa kelpaavien systeemien määrää. Ylläkuvatulla systeemillä on jo ainakin 10^{1000} tekijää tilasiirtymissä. Jokaisesta tekijästä on lisäksi siirtymä ainakin yhteen 10^{1000}:sta kohteesta. Tästä seuraa se, että todennäköisyys, että satunnainen systeemi toteuttaisi tämän kombinatorisen tilakoneen on luokkaa $1/(10^{1000})^{10^{1000}}$. Ei ole syytä olettaa, että satunnainen systeemi, kuten Searlen seinä toteuttaisi tarvittavat ehdot. Emme tosin tiedä fysikaalisen materian fundamentaalisia rakennusosia, ja siis emme voi tietää, jos esimerkiksi protonilla olisikin äärettömän rikas sisäinen rakenne ja näin teoriassa voisi toteuttaa ylläolevan systeemin. Mutta jos joku väittää, että näin olisi, tarvitsisi hän todella vahvoja argumentteja. (Chalmers, 1994).

(3) Voiko systeemi implementoida useita komputaatioita samaan aikaan? Kyllä, jokainen monimutkainen systeemi implementoi useita yksinkertaisia komputaatioita.

Kombinatorisen tilakoneen avulla Chalmers kykenee vastaamaan varsin uskottavasti Putnamin ja Searlen väitteeseen siitä, että komputationalismista seuraisi se, että kaikki fysikaaliset systeemit olisivat samalla tavalla tietoisia kuin ihmiset. Kaikki fysikaaliset systeemit periaatteessa ovat kykeneviä tietoisuuteen, mutta vain silloin, jos ne toteuttaisivat tarpeeksi kompleksisen ja funktionaalisesti relevantin komputaation.

Neljäs kysymys, joka ei liity implementaatioon, mutta mihin Chalmers silti haluaa vastata, on seuraava: Kun selitämme, miten annetun komputaation ja kombinatorisen tilakoneen implementaatio tapahtuu, mikään tässä selvityksessä ei viittaa semantiikkaan

tai tilakoneen sisäisten tilojen representationaalisiin sisältöihin, toisin sanoen merkityksiin, joita tilat kuvaavat. Miksei ylläoleva selitys kerro siis sitä, mistä tilakoneen sisäinen tila saa juuri sen tietyn merkityksen? Miksei selitys kerro esimerkiksi miksi tietty tila vastaa tuntemusta kipu? Tähän on kuitenkin selkeä vastaus. Näin asian tuleekin Chalmersin mukaan olla: komputaatiot on määritelty pelkästään syntaktisesti, ei semanttisesti. Jos sisällyttäisimme semanttisia aineksia implementaation ehtoihin, vaarantaisi se komputaation roolin kognitiotieteissä, sillä semantiikka on aivan liian heikosti ymmärretty käsite.

Semantiikka tarvitsee siis täysin oman selvityksensä. Teoriat siitä mistä tilakoneen tilat saavat merkityksensä ja sisältönsä, ovatkin yksi tärkeimmistä aiheista, joita käsitellään tutkielman seuraavissa luvuissa. Aloitan kuitenkin seuraavaksi esittelemällä joitakin vasta-argumentteja, joiden mukaan komputationalismi ei ylipäänsä voi olla oikea teoria mielen ja materian suhteesta.

5 VASTA-ARGUMENTTEJA KOMPUTATIONALISMILLE

Varsinaisen komputationalismin muotoili siis Hilary Putnam 1960-luvulla. Kuitenkin varsin suurta osaa länsimaisen filosofian historiasta on leimannut jonkinlainen usko siihen, että ihmisen älykkyys on formalisoitavissa. Tästä on hyvin pieni harppaus siihen ajatukseen, että myös ihmisälyn kaltainen tekoäly olisi luotavissa. Viimeistään Turingin esittelemä behavioristinen argumentti (Turingin testi) 1950, vakuutti laajat piirit tekoälyn onnistumisesta. Tutkijat, jotka Turingin hengessä uskoivat tekoälyn onnistumiseen, saivat olla lähes rauhassa aina 1960-luvun loppupuolelle saakka, mikä ei ollut yllätys Yhdysvalloissa tuolloin vallinneessa behavioristisessa ilmapiirissä. Suurin osa tekoälytutkimuksesta nimittäin tapahtui tuolloin juuri Yhdysvalloissa. Hubert Dreyfus oli kuitenkin ensimmäinen akateeminen filosofi, joka otti merkittävällä tavalla osaa keskusteluun tekoälyn filosofiasta (Anderson, 1989, 43). Hän julkaisi vuonna 1965 artikkelin Alchemy and Artificial Intelligence ja vuonna 1972 tekoälyn kritiikin klassikon What Computers Can't Do. Teosten nimistä voi päätellä hänen mielipiteensä tekoälyn mahdollisuuksista. Hänen tarkoituksensa oli kumota täydellisesti se optimismi joka vallitsi tekoälytutkijoiden keskuudessa. Hän pyrki itseasiassa osoittamaan, että koko länsimaisessa ajattelussa on perustavanlaatuisia virheitä, ja että usko tekoälyn onnistumiseen oli suoraan seurausta näistä virheistä.

5.1 Dreyfusin tekoälyn kritiikki

Dreyfusin mukaan länsimaista ajattelua hallitsi neljä hyvin vahvaa oletusta, joiden hyväksymiseen ei ole riittäviä perusteita. Ajatus tietoisen koneen mahdollisuudesta seuraa kuitenkin Dreyfusin mukaan suoraan näistä vääristä oletuksista. Hän kutsui näitä oletuksia seuraavin nimin: biologinen, psykologinen, epistemologinen ja ontologinen oletus.

5.1.1 Biologinen oletus

Ensimmäinen näistä, eli biologinen oletus, viittaa siihen, että oletamme jotain merkittävää tietoisuutemme biologisesta perustasta. Oletuksen mukaan ihmisaivot

toimivat kuten digitaalinen tietokone, vaikka ovatkin perustaltaan biologista, orgaanista materiaa. Siitä asti kun neurofysiologia löysi, että aivojen neuronit lähettävät jokseenkin kaikki-tai-ei-mitään periaatteella sähköisiä sykäyksiä toisiin neuroneihin, on nämä sykäykset rinnastettu biteiksi informaatiota tietokoneessa (Dreyfus, 1972, 159). Dreyfusin mukaan kyseinen oletus on kuitenkin vain empiirinen hypoteesi, jolle ei ole selviä todisteita, ja joka on muutenkin aikansa elänyt. Hänen mukaansa useat tutkimukset päinvastoin osoittavat, että aivot toimivat ainakin osittain analogisesti. Ja jos näin on, ei ole syytä uskoa, että aivojen toiminta olisi mahdollista kuvata digitaalisella formalismilla tai millään formalismilla ylipäätään.

Usein tehdään kyllä oletus, että aivot toimivat kuten digitaalinen tietokone, varsinkin tekoälytutkijoiden keskuudessa. Tuota oletusta ei ole kuitenkaan välttämätöntä tehdä tekoälyn luomisen kannalta. Esimerkiksi Kurzweil (2002; 428, 442) vastaa seuraavasti: On totta, että ihmisaivot käyttävät digitaalisesti kontrolloituja analogisia metodeita, mutta myös me voimme käyttää niitä koneissa, jotka olemme itse valmistaneet. Esimerkiksi kehittyneet neuroverkot käyttävät jo nyt hyvin tarkkoja malleja ihmisen neuroneista, sisältäen epälineaareja, analogisia aktivaatiofunktioita. Analogisia metodeja voidaan helposti luoda käyttäen perinteisiä transistoreja, jotka ovat olennaisesti analogisia laitteita. Digitaalisuus on vain keinotekoisesti tuotettu ominaisuus, joka saadaan aikaiseksi asettamalle transistorien syötearvoille jokin raja-arvo.

Toisaalta analogiset metodit eivät kykene tekemään mitään sellaista asiaa, mitä digitaaliset metodit eivät kykenisi tekemään. Digitaalinen prosessi voi emuloida analogista prosessia mille hyvänsä tarkkuudelle saakka. Luonnossa on aina läsnä taustakohinaa, joten voimme käytännössä approksimoida täydellisesti analogisia prosesseja käyttäen digitaalisia metodeja.

Toisaalta digitaalisuus/analogisuus erottelun mielekkyyttä voidaan kyseenalaistaa myös sillä, että ne ovat molemmat vain tapoja suorittaa informaation prosessointia. Fysiikka asettaa kuitenkin rajat myös informaation määrälle ja sen prosessoinnille. Analogisen prosessin ylivertaisuuden taustalla on pidetty ehkä sitä, että se käsittelisi tietoa jollakin tapaa äärettömällä tarkkuudella. Fysiikan laeista kuitenkin seuraa, että jokaisella

fysikaalisella systeemillä on tietty raja, kuinka monta bittiä informaatiota se voi sisältää (Cotogno, 2003, 202). Tällöin ei tunnu merkitykselliseltä se, onko tuo informaatio esitetty digitaalisesti vai analogisesti. Analogisuus-argumentti ei vaikuta siis pätevältä.

On tosin esitetty, että aivot eivät voi olla digitaalinen tietokone myös muista syistä kuin oletetun analogisuuden takia. Yleisin tällainen syy nykyään on useiden tutkijoiden esittämä väite, jonka mukaan aivot eivät voi olla digitaalinen tietokone, koska aivojen prosesseissa on mukana kvanttimekaanisia ilmiöitä, jotka eivät ole digitaalisia. Mutta tätä argumenttia tutkimme lisää kappaleessa 5.3.

5.1.2 Psykologinen oletus

Biologinen oletus on joka tapauksessa helposti ratkaistavissa, se on täysin empiirisesti testattavissa oleva hypoteesi. Dreyfusin kritisoima toinen oletus, eli psykologinen oletus, on kuitenkin huomattavasti hankalampi tapaus. Sen mukaan, ei aivot, kuten biologisessa oletuksessa, vaan mieli, toimii kuten digitaalinen tietokone. Tämä on juuri komputationalismin olennaisin teesi: mieli on komputaatiota. Tämä ei ole empiirisesti testattavissa oleva väite, se on pitkälti filosofinen kysymys.

Psykologisessa oletuksessa on siis kyse siitä, mistä jo funktionalismin kohdalla keskusteltiin. Sen mukaan on löydettävissä kolmas taso fysikaalisen ja fenomenaalisten tasojen välillä. Sitä voidaan kutsua funktionaaliseksi, komputationaaliseksi tai informaatioprosessointitasoksi. Psykologisen oletuksen mukaan mieli ja fenomenaaliset ilmiöt syntyvät silloin kun tuolla komputationaalisella tasolla tapahtuu tietty tapahtuma. Mitä perusteita tälle oletukselle sitten voisi olla? Aikaisemmissa funktionalismia ja komputationalismia käsittelevissä kappaleissa tämä oletus oikeastaan vain esiteltiin, mutta mitään perusteita sille ei annettu. Ainoastaan todettiin, että se selviytyy esimerkiksi identiteettiteoriaa paremmin moninainen realisoitavuus -periaatteen toteuttamisesta.

Dreyfus pitää ehkä suurimpana syynä sille, että psykologinen oletus usein hyväksytään, yhtä toista myös yleensä hyväksyttyä oletusta. Hän siteeraa Milleriä ym. (1960, 16) todisteena tälle: "Minkä tahansa kuvauksen, joka täydellisesti selittää ihmisen

käyttäytymisen, tulee olla joukko sääntöjä. Toisin sanoen, siinä tulee olla suunnitelman piirteet, jotka kykenevät ohjaamaan tekoja, jotka on kuvattu". Tämä tarkoittaa siis sitä, että mikäli on olemassa psykologinen teoria, joka onnistuu selittämään ihmisen käyttäytymisen, täytyy tuon teorian olla sääntöjen joukko. Teorian täytyy kyetä ohjaamaan ihmisen käyttäytymistä eri tilanteissa.

Oletukselle, että psykologisen teorian täytyy olla tällainen, on pitkä historia. Kant analysoi kaiken kokemisen ja havaitsemisen sääntöjen avulla. Platon taas ajatteli, että kaikilla teoilla, jotka ovat harkittuja, eivätkä satunnaisia, on rationaalinen rakenne. Tämä rakenne on ilmaistavissa jonkin teorian avulla, joka taas on joukko sääntöjä. Hänen mukaansa nämä säännöt ovat valmiina ihmisen mielessä. Ja vaikka ihminen ei välttämättä tietoisesti seuraa näitä sääntöjä, on teoilla, jotka ihminen tekee, kuitenkin tällainen rationaalinen rakenne.

Jos edellä esitetty oletus psykologisen teorian luonteesta hyväksytään, seuraa tästä, että psykologisen teorian täytyy olla ilmaistavissa tietokoneohjelmana, koska tietokoneohjelma on nimenomaan joukko sääntöjä. Toisaalta koska ihminen tietoisesti kykenee ohjaamaan käyttäytymistään, niin tästä seuraa, että ihmisen tietoisuus on vahvasti linkitetty tuohon sääntöjen joukkoon, eli tietokoneohjelmaan. Tietoisuus on siis tietokoneohjelman suoritusta. Ja juuri tämä on komputationalismin perusteesi, ja psykologinen oletus.

Dreyfus myöntää, että psykologiselle oletukselle on kyllä joitain perusteita. Ihminen on fysikaalinen objekti. Modernin fysiikan menestys on osoittanut, että fysikaalisen objektin käyttäytymisen täydellinen selitys voidaan antaa tarkkojen fysikaalisten lakien avulla. Nämä lait taas voidaan ohjelmoida tietokoneelle, joka tällöin, ainakin periaatteessa, voi täydellisesti simuloida fysikaalisen objektin käyttäytymisen. Tästä seuraa, että myös neurofysiologinen kuvaus ihmisen käyttäytymisestä on ainakin periaatteessa simuloitavissa digitaalisessa tietokoneessa (Dreyfus, 1972, 177). Sen myötä, että periaatteessa kaikki fysikaaliset lait ovat simuloitavissa tietokoneella, on ehdotettu myös mielenkiintoista teoriaa, jonka mukaan kaikki fysikaaliset objektit eivät pelkästään ole simuloitavissa tietokoneen avulla, vaan itseasiassa koko universumi

voidaan itsessään käsittää jättimäisenä tietokoneena. Jätän kuitenkin tämän teorian tutkimisen tämän tutkielman puitteissa sivuun.

Dreyfus joka tapauksessa väittää, että vaikka edellä kuvattu oletus psykologisen teorian luonteesta kykenisi selittämään ihmisen käyttäytymisen, niin siitä ei seuraa, että se kykenisi selittämään myös ihmisen fenomenaaliset kokemukset, siis kvaliat. Hän ei suostu hyväksymään, että käyttäytymisen ja tietoisuuden välillä vallitsisi välttämätöntä suhdetta. Hänen mukaansa fenomenaaliset kokemukset eivät ole selitettävissä käyttäytymisen avulla. Itse olen tästä samaa mieltä Dreyfusin kanssa. Fenomenaaliset kokemukset eivät ole selitettävissä käyttäytymisen avulla. Dreyfus ei kuitenkaan kykene myöskään kumoamaan psykologista oletusta, vaikka edellisessä kappaleessa kuvattu argumentti ei sitä täysin pystykään puolustamaan.

Edellisen argumentin hylkääminen tosin saattaa olla hyppy kohti dualismia. Usein nimittäin komputationalistit esittävät kysymyksen, että jos kerran kaikki fysikaaliset tapahtumat ovat selitettävissä sääntöjen, siis komputaatioiden avulla, niin mitä muutakaan mentaaliset tapahtumat voisivat olla kuin komputaatiota. Esimerkiksi Penrose (1995) on kuitenkin yrittänyt puolustaa fysikalismia, joka on ei-komputationalistista. Myös tästä lisää luvussa 5.3.

5.1.3 Epistemologinen oletus

Dreyfusin kritisoima kolmas oletus on epistemologinen oletus. Se on hyvin samantyyppinen kuin psykologinen oletus, mutta hieman heikompi. Psykologisen oletuksen mukaanhan käyttäytyminen on *selitettävissä* formaalien sääntöjen avulla ja noiden samojen sääntöjen perusteella on mahdollista luoda käyttäytymistä uudelleen. Epistemologisen oletus sen sijaan olettaa ainoastaan, että kaikki ei-satunnainen käyttäytyminen on *formalisoitavissa* säännöiksi ja noita sääntöjä käyttäen voidaan ohjelmoida tietokone suorittamaan tiettyä käyttäytymistä. Se ei siis oleta, että ihmisen käyttäytyminen voidaan *selittää* sääntöjen avulla, vaan ainoastaan sen, että se on *formalisoitavissa* sääntöihin.

Mikä on se merkittävä ero, joka on siinä, että käyttäytyminen voidaan joko selittää tai vain formalisoida sääntöjen avulla? Se on siinä, että jos selitämme ihmisen käyttäytymisen sääntöjen avulla, hyväksymme, että se on vain sääntöjen noudattamista. Jos taas vain sanomme, että ihmisen käyttäytyminen on formalisoitavissa, siitä seuraa, että voimme simuloida käyttäytymistä, mutta käyttäytymisen todellisesta luonteesta emme voi sanoa mitään varmaa. Ihmisen käyttäytyminen saattaa tällöin olla aidosti jotain muuta, kuin sääntöjen noudattamista. Se miten tietoisuus liittyy käyttäytymiseen tässä tapauksessa on siinä, että psykologisen oletuksen hyväksyessämme, hyväksymme myös, että tietoisuus on sääntöjen noudattamista, eli komputaatiota. Vain epistemologisen oletuksen hyväksyessämme, emme voi sanoa tietoisuuden luonteesta mitään varmaa. Jos haluamme tällöin väittää, että tietoisuus on komputaatiota, täytyy meidän kyetä esittämään erillisiä argumentteja sen puolesta, että tietoisuuden syntyyn todellakin riittää pelkästään oikeanlainen komputaatio.

On siis huomattava, että jos hyväksymme psykologisen oletuksen, hyväksymme myös epistemologisen oletuksen. Toisin sanoen, jos emme vakuutu Dreyfusin argumenteista psykologista oletusta kohtaan, emme luultavasti vakuutu myöskään Dreyfusin epistemologista oletusta kohtaan esittämistä argumenteista. Mutta, jos emme vakuutu psykologista oletusta tukevista argumenteista, voimme silti säilyttää uskomme komputationalismiin, ehkä kuitenkin hieman heikompaan versioon siitä, jos kykenemme löytämään uskottavia argumentteja epistemologisen oletuksen puolesta.

Dreyfusin argumentti epistemologista oletusta vastaan perustuu väitteeseen, että tietokoneet eivät kykene samanlaiseen kielelliseen pätevyyteen kuin ihmiset. Ja koska tietokoneet eivät ole kielellisesti yhtä päteviä kuin ihmiset, ei ihmisten käyttäytyminen ole formalisoitavissa, ja tästä seuraa, että myöskään samankaltainen tietoisuus ei ole luotavissa tietokoneelle, kuin mikä ihmisellä on.

Tässä tuntuu kuitenkin heti alkuunsa olevan ristiriita. Eikö Dreyfus (1972, 177) ollut myöntävinään, että kaikkien fysikaalisten objektien käyttäytyminen on formalisoitavissa ja siis simuloitavissa. Nyt hän tuntuu kuitenkin väittävän jotain muuta. Hänen mielestään on niin, että vaikka periaatteessa ihmisaivot ovat simuloitavissa digitaalisella tietokoneelle, niin käytännössä vaadittava laskutoimitusten määrä on niin valtava, että

edes planeetan kokoinen tietokone ei kykenisi niitä suorittamaan (Dreyfus, 1972, 196). Tämä tarkoittaa sitä, että ihmisaivojen prosessit ovat heikosti ei-laskennallisia.

On totta, että kaoottisessa luonnossa esiintyy heikosti ei-laskennallisia ilmiöitä, mutta nämä ilmiöt perustuvat siihen, että luonto on käytännössä avoin ja ääretön systeemi, jossa on ääretön määrä muuttujia. On perusteltua pitää kuitenkin ihmisaivoja oleellisesti suljettuna systeeminä, jolloin muuttujien määrä on äärellinen. Dreyfus kuitenkin selvästi tuntuu olettavan, että aivojen simulointiin ei riitä pelkästään neuronien, synapsien ja niiden välisten sähköimpulssien mallintaminen, vaan simulointi vaatii, että mallinnetaan myös epämääräinen joukko muita fysikaalisia ominaisuuksia. Komputationalismi olettaa kuitenkin, että neuroneita matalammalle tasolle ei tarvitse mennä, toisin sanoen, neuroneita matalammalla tasolla ei ole funktionaalisesti merkittäviä ominaisuuksia, jotka tulisi ottaa huomioon. Tämä on joka tapauksessa empiirisesti ratkaistavissa oleva kysymys. Onhan esimerkiksi ehdotettu (Penrose, 1995), että kvanttimekaanisella tasolla voisi olla joitain funktionaalisesti olennaisia tapahtumia (ks. 5.3). Mitään todellista näyttöä tämän puolesta ei kuitenkaan ole.

Palataan kuitenkin takaisin Dreyfusin väitteeseen tietokoneen kielellisestä kykenemättömyydestä. Hän perustaa väitteensä siihen, että koska tietokone kykenee käyttäytymään vain tiukkojen sääntöjen mukaan, tai vaihtoehtoisesti täysin satunnaisesti, niin tästä seuraa, että kun tietokone yrittää ymmärtää lauseiden merkityksiä, se kykee luokittelemaan niitä vain joko täysin sääntöjen mukaisesti tai satunnaisesti. Ongelmaksi tietokoneelle muodostuvat Dreyfusin mukaan moniselitteiset lauseet. Tietokone ei kykene päättelemään mikä mahdollisista merkityksistä on se oikea. Ihminen kykenee kuitenkin Dreyfusin mukaan tekemään kontekstin perusteella niin sanotun valistuneen arvauksen, jolloin ihminen kykenee moniselitteisyydestä huolimatta ymmärtämään lauseen oikean merkityksen. Dreyfus väittää, että tietokoneelle tämä on mahdoton tehtävä. (Dreyfus, 1972, 199)

Dreyfus ei tunnu ymmärtävän, että myös tietokoneella voi olla jonkinlainen käsitys vallitsevasta kontekstista ja näin ollen tehdä valistuneen arvauksen. Vaikka nykyisille tietokoneille tämä onkin lähes mahdoton tehtävä, ei ole perusteltua väittää, ettei se olisi lainkaan mahdollista. Myöskään me ihmiset emme aina ymmärrä kaikkia moniselitteisiä

lauseita. Se ei kuitenkaan tarkoita, ettemmekö ymmärtäisi kuitenkin jotain ja ettemmekö olisi aidosti tietoisia. Onko siis perusteltua sanoa, että koneet eivät koskaan kykene ymmärtämään kieltä, vain sillä perusteella, että nykyiset tietokoneet ovat hyvin kaukana ihmisen kyvykkyydestä ymmärtää kieltä? On kyllä totta, että riittävän suuren kontekstin ohjelmoiminen käsin tietokoneeseen voi olla mahdotonta. Mutta jos tietokone esimerkiksi liitetään aistinsensorien ja keinoraajojen avulla maailmaan, se voisi kasvaa ihmisten parissa ja oppia kuten pieni lapsi ympäristöstään, ja hitaasti oppia ymmärtämään myös konteksti, jossa ihminen käyttää kieltään, ja näin ymmärtämään samalla tavoin kieltä kuin ihminen.

Merkityksiä tulemme tutkimaan lisää kappaleessa 6. Dreyfus esimerkiksi olettaa edellämainitussa tapauksessa, että tapa, jolla tietokone käsittelee sanoja ja merkityksiä on aina hyvin yksinkertainen ja suoraviivainen. Hän olettaa, että merkitykset ovat aina yksittäisiä symboleita tietokoneen muistissa. Ja tietokone käsittelee näitä symboleita itsenäisinä ja atomisina entiteetteinä. Tämä on fysikaalinen symbolisysteemi -hypoteesin mukainen tapa ajatella merkityksiä. Se on usein varsin ongelmallinen hypoteesi. Toinen nykyään suositumpi tapa ajatella merkityksiä, on konnektionistinen tapa. Siinä tietokoneella mallinnetaan neuroverkkoa. Tällöin merkitykset eivät ole atomisina symboleina neuroverkossa, vaan niiden realisaatio on hajautettu. Tällä on etuna huomattavasti parempi mukautuvuus edellisen kaltaisissa tilanteissa.

5.1.4 Ontologinen oletus

Dreyfus kritisoima neljäs ja viimeinen oletus on ontologinen oletus. Se liittyy kiinteästi käsitteeseen konteksti, josta puhuttiin edellisessä kappaleessa. Ontologisen oletuksen mukaan maailma voidaan täysin analysoida kontekstittomien ja atomisten faktojen avulla (Dreyfus, 1972, 205). Tämä liittyy selvästi myös psykologisessa oletuksessa käsiteltyyn teoriaan, että kaikki fysikaaliset tapahtumat ovat selitettävissä sääntöjen avulla. Dreyfusin mukaan ontologinen oletus on välttämätöntä hyväksyä, jos hyväksyy epistemologisen oletuksen. Tietokoneellehan ei Dreyfusin mukaan kyetä luomaan kontekstia ympäröivästä maailmasta, jonka perusteella ihminen taas kykenee ymmärtämään kieltä sekä ylipäätään ympäröivää maailmaa. Nyt, jotta tietokone kykenisi aidosti ymmärtämään ilman tuota kontekstia, tulisi maailman olla

analysoitavissa kontekstittomien ja atomisten faktojen avulla, jolloin tietokone kykenisi ymmärtämään käyttämänsä faktat yksinkertaisesti siitä syytä, että ne olisivat itse itsensä selittäviä, ilman että ne tarvitsisivat selityksekseen ympäröivää kontekstia.

Ontologisella oletuksella on toki pitkä historia. Vaatimus, että tiedon täytyy olla ilmaistavissa sääntöjen ja määritysten avulla yksiselitteisesti, oli läsnä jo Platonin teorioissa. Leibniz ilmaisi asian eksplisiittisesti. Hänen mukaansa ymmärtämisessä analysoimme konseptit yksinkertaisempiin elementteihin, ja nämä edelleen yksinkertaisempiin elementteihin. Jotta vältymme ikuiselta regressiolta, täytyy olla lopulliset yksinkertaiset elementit, joiden avulla kaikki kompleksit konseptit voidaan ymmärtää. Näitä hän kutsui ihmismielen aakkosiksi. Ja jotta näillä aakkosilla olisi yhteys maailmaan, niillä täytyy olla itsessään jonkinlaisia ominaisuuksia: yhteys, ryhmitys ja järjestys, mitkä löytyvät myös vastaavista fysikaalisista objekteista. Nämä sisäiset ominaisuudet antavat merkityksen käsitteille. Ei pelkästään rationalistinen, mutta myös empiristinen traditio länsimaisessa filosofiassa on vahvasti kannattanut ideaa tällaisista tiedon diskreeteistä elementeistä. Humelle kaikki kokemus koostui impressioista, jotka olivat itsenäisiä, determinoituja osia kokemuksesta. Rationalismi ja empirismi yhtyivät Russellin loogisessa atomismissa, ja puhtaimman muotonsa ontologinen oletus sai Wittgensteinin Tractatuksessa, jossa maailma on määritelty käyttäen atomisia faktoja, jotka voidaan ilmaista loogisesti itsenäisten propositioiden avulla. (Dreyfus, 1972, 211) Uusimmista teorioista Jerry Fodorin ajattelun kieli on tällainen. Siinä ihmisellä on synnynnäinen mentaalinen kieli, *mentalese*, joka koostuu mentaalisista representaatioista. Tietoisuus syntyy tämän kielen prosessoinnista.

Wittgenstein itse kuitenkin Tractatuksen jälkeen sekä esimerkiksi Heidegger ovat kritisoineet 1900-luvulla hyvin paljon kyseistä länsimaisen filosofian traditiota. Ja Dreyfus on jatkanut kritisointia. Heidän mukaansa ymmärrys ei ole analysoitavissa pienempiin osiin, ja lopulta itsessään merkityksellisiin entiteetteihin, vaan se on luonteeltaan holistista. Ymmärrys tapahtuu päinvastoin kuin perinteisessä näkemyksessä siten, että yksittäiset käsitteet saavat merkityksensä kontekstin ja siis ympäröivän maailman perusteella, eivätkä käsitteen sisäisten merkitysten avulla, niinkuin ontologisessa oletuksessa. Tälle näkemykselle tuntuu toki olevan jotain perustetta. Käsitteiden sisäiset merkitykset kuulostavat jokseenkin epäilyttäviltä, sen

sijaan monet käsitteet tuntuvat saavan merkityksensä juuri suhteestaan muihin käsitteisiin.

Joka tapauksessa, jos emme usko Dreyfusin väitettä, että tietokoneelle ei voida luoda tietoa vallitsevasta kontekstista, ei meidän tällöin tarvitse edes hyväksyä ontologista oletusta. Voimme tällöin olettaa, että myös tietokone voi ymmärtää käsitteiden merkityksiä holistisin periaattein, samalla tavoin kuin ihminen. Nousee kuitenkin kysymys, miten tuo vaadittava konteksti on mahdollista luoda tietokoneelle, tai miten itseasiassa ihminen kykenee tietoisuuden kontekstista saavuttamaan.

Dreyfusin mielestä siis koska tekoälytutkimus hyväksyy ontologisen oletuksen, niin tästä seuraa, että tietokoneeseen voitaisiin luoda jonkinlainen konteksti vain lisäten yksittäisiä faktoja valtava määrä tietokoneen tietokantaan. Tietokoneella ei ole tällöin kuitenkaan keinoa löytää näistä faktoista niitä relevantteja faktoja vallitsevaan tilanteeseen nähden, joten tietokoneella ei voi olla ymmärrystä vallitsevasta kontekstista.

Toisaalta Dreyfus myöntää, että myös holistisella käsityksellä merkitysten ymmärtämisessä on samanlainen ongelma. Kun selvitämme faktojen merkityksiä kontekstin perusteella, täytyy meillä olla jonkinlainen selitys sille mikä on oikea konteksti. Yksi mahdollinen vastaus tähän olisi, että löydämme oikean kontekstin laajemman kontekstin perusteella. Tämä ajautuu kuitenkin ilmiselvästi täysin vastaavalla tavalla ikuiseen regressioon, kuin ontologinen oletuskin.

Dreyfusin ratkaisu ongelmaan on se, että ihmisille on olemassa myös kolmas vaihtoehto. Hänen mukaansa edellemainittujen regressiivisten kontekstien sijasta ihminen kykenee tunnistamaan nykyisen kontekstin jatkumona edellisestä kontekstista. Voimme siis löytää relevantin kontekstin sen perusteella, mikä hetkeä aikaisemmin oli relevantti. Tässäkin on helppo kuitenkin huomata ongelma: ihmisenkin historia päättyy jonnekin. Mistä siis ihminen vastasyntyneenä löytää sen oikean kontekstin? Dreyfusin vastaus tähän on, että ihminen on geneettisesti luotu reagoimaan vauvana ympäristön signaaleihin tietyllä tapaa. Ja näin lapsi kasvaessaan voi näiden geneettisten refleksien avulla tulla tietoiseksi vallitsevasta kontekstista. Nyt voimme kuitenkin taas kysyä, että

emmekö myös me kykenisi ohjelmoimaan tietokoneen siten, että myös sillä olisi samankaltaisia alkukantaisia refleksejä, joiden avulla se voisi hitaasti oppia ymmärtämään paremmin vallitsevaa kontekstia. Dreyfus on varovaisen optimistinen tämän lähestymistavan suhteen. Hänen mukaansa kukaan ei ole kuitenkaan vakavasti yrittänyt kyseistä lähestymistapaa tekoälyn luomiseen. Ja myös tähän jää pieni ongelma: miten näistä määrätyistä alkukantaisista reflekseistä ympäristön signaaleihin voi kehittyä joustava ymmärrys kontekstista. Tämä ongelma on kuitenkin läsnä myös ihmisen tapauksessa.

Itseasiassa hänen päällimmäinen kritisismin kohteensa kirjansa julkaisemisen aikoihin olikin silloin hallinneiden tekoälyohjelmien tutkimusmetodit. Myöhemmin hän on suhtautunut jokseenkin optimistisesti esimerkiksi konnektionistisiin lähestymistapoihin tekoälyn luomiseen, joissa tietokoneet saavat ymmärryksen merkityksistä juuri holistisin periaattein, eivätkä ontologisen oletuksen mukaisesti.

Dreyfusin epäilyt siitä, että länsimaisen filosofian traditiossa käytetyt oletukset saattavat olla virheellisiä, ovat toki oikeutettuja. Mutta hän ei kuitenkaan pysty vakuuttamaan, että ne täysin olisivat virheellisiä. Ainakaan hän ei pysty vakuuttamaan, että hänen argumenttinsa olisivat uhka tekoälylle. Komputationalismi hyväksyy ainakin jonkinasteisesti kyllä psykologisen ja epistemologisen oletuksen, mutta Dreyfusin argumentit juuri niiden kumoamiseen ovat ehkä kaikkein heikoimmat. Sen sijaan komputationalismilla ei juurikaan ole sidoksia biologiseen eikä ontologiseen oletukseen, ja juuri nämä ovat ne oletukset, joita vastaan Dreyfusin argumentit parhaiten purevat. Keskustelun arvoisia nämä argumentit kyllä ovat, ja tavallaan ne ovat pohjana myöhemmälle tekoälyn kritiikille. Kuten tulemme huomaamaan, myös seuraavissa kappaleissa käsiteltävät argumentit ovat tavallaan jonkinlaisia, hieman konkreettisempia, variaatioita näistä argumenteista.

5.2 Argumentti matematiikasta

Tekoälyä vastaan on kehitetty useita argumentteja, jotka saavat voimansa matemaattisen logiikan tuloksista. Nämä argumentit viittaavat luvussa 5.1.2 käsiteltyyn psykologiseen oletukseen. Psykologisessa oletuksessa väitettiin, että mentaaliset prosessit ovat

komputationaalisia, toisin sanoen laskennallisia. Nämä matemaattiset ja kvanttimekaaniset argumentit sen sijaan pyrkivät osoittamaan, että mentaaliset prosessit eivät voi olla laskennallisia. Näissä prosesseissa on argumenttien mukaan jokin ei-laskennallinen piirre, mistä seuraa, että mentaalisuus ei ole luotavissa pelkästään komputationaalisia keinoja käyttäen.

Tunnetuin matemaattisen logiikan tulos, jota on käytetty komputationalismia vastaan on Gödelin (1934) epätäydellisyysteoreema. Sen mukaan missä tahansa sellaisessa aksiomaattisessa systeemissä, joka kykenee generoimaan luonnolliset luvut, on välttämättä propositioita, joita ei voida todistaa oikeaksi tai vääräksi, ellei mahdollisesti kyseinen aksiomaattinen systeemi itsessään ole inkonsistentti. Tällaiset propositiot ovat itseasiassa yhtä yleisiä kuin todistettavissa olevat propositiot. Myös Turing (1936) päätyi samanlaiseen tulokseen tutkiessaan komputaatiota. Hän osoitti, että on olemassa ongelmia, jotka ovat hyvin määriteltyjä, ja joihin on olemassa jokin ratkaisu, mutta silti kyseinen ratkaisu ei ole löydettävissä Turingin konetta käyttäen. Ja kun muistamme, että Turingin kone kykenee mallintamaan minkä tahansa laskennallisen prosessin, seuraa tästä, että on olemassa ongelmia, joihin ei ole löydettävissä ratkaisua laskennallisia keinoja käyttäen. Myös Church (1936) kehitti samoihin aikoihin vastaavan teoreeman aritmetiikan alalla. Nämä kolme tulosta olivat ensimmäiset todisteet siitä, että logiikalla, matematiikalla ja komputaatiolla on tarkat rajat, joita ne eivät kykene ylittämään (Kurzweil, 2002).

Nyt argumentit, jotka yrittävät kieltää sen, että mentaalisuus olisi luotavissa pelkästään komputationaalisia keinoja käyttäen, perustuvat siihen, että vaikka komputaatiolla on rajansa, ihminen kykenee nuo rajat jollakin tapaa ylittämään. Ihminen kykenee näkemään näitten propositioiden totuusarvon, vaikka algoritmia niiden osoittamiseksi ei olekaan. Näin ovat argumentoineet esimerkiksi Lucas (1963) ja Penrose (1994).

Millä tavoin ihminen sitten kykenee näkemään näiden propositioiden totuusarvot? Nämä perustelut ovat hyvin monimutkaisia ja hyvin paljon kritisoituja. En lähde tämän tutkielman puitteissa niitä tarkemmin tutkimaan. Näiden argumenttien kriitikoiden vastaus on usein kuitenkin se, että ihmiset eivät aidosti ole yhtään sen kyvykkäämpiä ratkaisemaan näiden propositioiden arvoja, kuin tietokoneetkaan. Ihmiset voivat tehdä

joissain tapauksissa valistuneita arvauksia ja käyttää heuristisia metodeja, jotka toimivat satunnaisesti. Nämä ovat kuitenkin algoritmisia prosesseja ja myös tietokoneet voivat käyttää niitä (Kurzweil, 2005).

Chalmers (1995) väittää, että oletus, että ihminen kykenee näkemään satunnaisen Gödelin lauseen totuuden, vaatii sen, että kykenemmme määrittämään onko mikä tahansa satunnainen formaali systeemi konsistentti vai inkonsistentti. Ei ole kuitenkaan syytä uskoa, että meillä olisi tämä kyky yleisesti. Emme voi tietää onko mielemme konsistentti ja tästä seuraa, ettemme voi aidosti tietää näiden propositioiden totuusarvoa. Chalmersin mukaan kaikki Gödeliläiset argumentit perustuvat jollakin tapaa tähän oletettuun tietoon, että mielemme on konsistentti, vaikka itseasiassa tämä tieto johtaa ristiriitaan.

Lisäksi Gödelin epätäydellisyysteoreeman tulkinnasta on kiistelty paljon. Teoreema on esimerkiksi relevantti vain äärettömien matemaattisten systeemien yhteydessä. Ihmisaivot, tietokoneiden tavoin, eivät kuitenkaan ole äärettömiä systeemeitä, niissä ei voi olla ääretöntä määrää eri tiloja. Vaikka ihmismieli kykeneekin muodostamaan äärettömän käsitteen ja käyttämään sitä päättelyssään, ei mieli itsessään silti ole ääretön. Voiko äärettömiä systeemejä koskevaa teoreemaa siis käyttää hyväksi vertailtaessa ihmismielen ja tietokoneen matemaattisia kykyjä?

Matemaattiset argumentit ovat joka tapauksessa jollakin tapaa epäuskottavia. Se, että ihmismielen tietoisuuden perimmäistä luonnetta arvioidaan sellaisen jokseenkin toisarvoisen seikan kuin matemaattisen kyvyn perusteella, tuntuu epäilyttävältä. Myös ihmiset tekevät usein matemaattisessa päättelyssään virheitä, ja varsin usein tietokone onkin parempi matemaattisessa päättelyssään kuin ihminen.

5.3 Argumentti kvanttimekaniikasta

Gödeliläiset argumentit yrittävät siis osoittaa, että ihmismielessä on jokin ei-laskennallinen piirre. Loogisesti tämä on toki mahdollista. Oletetaan siis hetki, että gödeliläinen argumentti olisi pätevä, ja että aivot toimisivat jonkin ei-laskennallisen prosessin mukaisesti. Mikä voisi olla se fysikaalinen prosessi, joka saa aikaiseksi tämän

ei-laskennallisuuden? Tämä argumentti viittaa siis biologiseen oletukseen (ks. 5.1.1). Dreyfusin ehdotus tällaiseksi prosessiksi oli analogiset prosessit. Tämä ei tuntunut kuitenkaan uskottavalta, joten olisiko olemassa muita ehdotuksia ei-laskennalliseksi prosessiksi. Ja nyt, toisin kuin kappaleessa 5.1.3, meidän täytyy puhua vahvasti ei-laskennallisista prosesseista. Siis prosesseista jotka eivät ole edes periaatteessa simuloitavissa. Loogisesti on kyllä luotavissa formaaleja maailmoja, jotka toimivat ei-laskennallisesti (esim. Penrose, 1994, 30), mutta onko tässä aktuaalisessa maailmassa, jossa elämme, aidosti olemassa ei-laskennallisia prosesseja.

Näkemyksestä, jonka mukaan tällaisia prosesseja ei ole olemassa, käytetään nimitystä fysikaalinen Churchin teesi. Se on siis vahvempi versio Church-Turingin teesistä. Church-Turingin teesin mukaanhan kaikki laskennalliset funktiot voitiin suorittaa Turingin koneella, mutta se ei ottanut kantaa siihen onko olemassa ei-laskennallisia funktioita tai esiintyykö luonnossa ei-laskennallisia prosesseja. On osoitettu, että on olemassa ei-laskennallisia funktioita, esimerkiksi pysäytysongelma (Turing, 1936). Mutta se, onko luonnossa sellaisia prosesseja, jotka käyttäytyvät ei-laskennallisesti, on avoin kysymys. Fysikaalisen Churchin teesin mukaan tällaisia prosesseja ei ole, sen mukaan siis kaikki fysikaaliset systeemit voidaan mallintaa Turingin koneena. Tämä teesi on kuitenkin empiirinen. Jos luonnosta löytyisi jokin ei-laskennallinen prosessi, kumoaisi se teesin. Tällaisesta ei-laskennallisesta prosessista käytetään myös nimitystä hyperkomputaatio, siis komputaation rajat ylittävä. (Cotogno, 2003, 181.)

Mutta onko siis tällaista prosessia aidosti olemassa? Ainoa uskottava vaihtoehto tällaiseksi prosessiksi tuntuu löytyvän kvanttimekaniikan alueelta. Nimittäin ainoa tunnettu fysikaalinen tapahtuma, jonka uskotaan olevan ei-algoritminen, on kvanttimekaniikan kvanttihyppy, eli aaltofunktion romahdus tai tilavektorin reduktio. Kvanttimekaniikassa fysikaalisten systeemien tila ilmaistaan aaltofunktiona tai tilavektorina. Systeemin tila on tavallaan yksinkertaisten tilojen kombinaatio, superpositio. Vasta kvanttimekaanisessa mittauksessa systeemin aaltofunktiolle oletetaan tapahtuvan romahdus. Tässä tapahtumassa systeemin tila vakiintuu yhdeksi ominaistilaksi. Se mihin ominaistilaan systeemi päätyy ei ole laskennallinen, sille on mahdollista antaa vain todennäköisyyksiä. Tämä tekee aaltofunktion romahduksesta erityislaatuisen ilmiön muiden fysikaalisten ilmiöiden joukossa. On toki paljon

keskustelua siitä, onko tämä oikeasti todellinen ilmiö, vai liittyykö sen taustalle kenties vielä tuntemattomia fysikaalisia ilmiöitä, jotka saattaisivat olla täysin laskennallisia. Kuitenkin esimerkiksi Penrose (1994, 377) olettaa teoriassaan, että aaltofunktion romahdus on todellinen, ei-laskennallinen ilmiö, ja sen lisäksi hän uskoo, että myös tulevaisuuden kvanttigravitaatioteoriassa, eli kaiken teoriassa, täytyy olla ei-algoritmisia piirteitä.

Juuri Penrose (esim. 1994) ja Hameroff (1994) ovat ehdottaneet, että nämä kvanttimekaniikassa mahdollisesti ilmenevät ei-laskennallisuudet ovat keskeisessä asemassa tietoisuuden synnyssä. He ovat väittäneet, että aivojen neuroneissa on hienorakenteita, mikrotubuleita, joissa tapahtuu eksoottista komputaation muotoa, kvanttikomputaatiota. Kvanttikomputaatio käyttää laskennassaan kubitteja, jotka ovat tavallaan edellisessä kappaleessa kuvattuja pienimpiä mahdollisia fysikaalisia systeemeitä. Ne kykenevät esittämään kaikki mahdolliset ratkaisut ongelmiin yhtäaikaisesti, jolloin niitä voidaan pitää äärimmäisenä rinnakkaisprosessoinnin muotona. Penrose ehdottaa, että mikrotubulit ja niiden kvanttikomputaatio mahdollistavat tietoisuuden ihmisellä.

Esimerkiksi Koch & Hepp (2006) ovat kuitenkin kritisoineet tätä ehdotusta. Heidän mielestään aivot ovat märkä ja kuuma paikka, eikä siis kovinkaan ystävällinen ympäristö kvanttikomputaation vaatimalle kvanttikoherenssille tai kvanttikietoutumiselle. Kvanttikomputaatio nimittäin vaatii, että kvantti-ilmiöt tapahtuvat koherentisti, ikäänkuin yhteisöllisesti. Tällöin kvanttisysteemin tila on jollakin tavalla riippuvainen muista kvanttisysteemeistä, vaikkei niillä ole spatiaalista yhteyttä toisiinsa. Yleensä näin tapahtuu kuitenkin vain hyvin äärimmäisissä olosuhteissa, kuten suprajohteissa ja kvanttinesteissä. Aivoissa tapahtuu jatkuvasti kyllä valtava määrä kvanttitason ilmiöitä, mutta jotta tietoisuudelle relevantit kvantti-ilmiöt voisivat saavuttaa kvanttikoherenssin ja näin toimia osana kvanttikomputaatiota, täytyisi näiden tietoisuudelle relevanttien kvantti-ilmiöiden olla vahvasti eristettyinä muista aivojen ilmiöistä, eikä tämä tunnu mahdolliselta. Ei ole perusteita uskoa, että kvanttitilat voisivat pysyä yllä tarpeeksi kauan, jotta Penrosen teoria voisi toimia.

Kurzweil (2005, 451) toteaa myös, että vaikka aivoissa tapahtuisikin kvanttikomputaatiota, tämä ei huomattavasti muuta käsitystä ihmisaivoissa tapahtuvasta komputaatiosta. Ei ole mitään sellaista syytä, joka rajoittaisi kvanttikomputaation vain biologisiin organismeihin. Jos kvanttikomputaatiota tapahtuu biologisissa organismeissa, voidaan sitä keinotekoisesti tuottaa myös ei-biologisiin systeemeihin. Jo nyt on kyetty valmistamaan kokeellisia muutaman kubitin kvanttitietokoneita. Ja itseasiassa jopa tavallinen transistori käyttää kvanttiefektiä elektronitunneloinnissa.

Kvanttikomputaatio ei joka tapauksessa uhkaa komputationalismia, sillä on tunnettua, että myös klassiset tietokoneet kykenevät simuloimaan kvanttikomputaatiota. Mistä tämä sitten johtuu? Kvanttikomputaatio perustuu siis siihen, että kvanttikietoutumisen ja aaltofunktion romahduksen ansiosta kvanttitietokone pysty ratkaisemaan joitain ongelmia tehokkaammin kuin klassinen tietokone. Tämä johtuu kvanttikietoutumisesta, joka mahdollistaa valtavan rinnakkaisprosessoinnin. Aaltofunktion romahdus on osana tätä prosessia siinä, että se antaa satunnaisluvun, jonka perusteella kvanttikomputaatio suoritetaan. Kvanttitietokoneen suorittamat algoritmit ovat siis satunnaistettuja algoritmeja, jotka ratkaisevat ongelman tietyllä todennäköisyydellä. Kvanttitietokone voidaankin rinnastaa klassiseen satunnaistettuun Turingin koneeseen, ainoastaan sillä periaatteellisella erolla, että kvanttikomputaatio käyttää aidosti satunnaisia lukuja, kun taas klassiset koneet käyttävät pseudo-satunnaislukuja, jotka ovat siis algoritmisten metodien luomia. Pseudo-satunnaisuuslukujen luonti voidaan kuitenkin tehdä niin lähelle aitoa satunnaisuutta kuin halutaan. Knill (1996) toteaakin, että kvanttisatunnaisuus ei lisää laskennallista tehokkuutta verrattuna klassiseen kolikon heittoon. Kvanttikomputaation ainoa todellinen etu on siis se, että se saattaa esitellä laskennalle uuden kompleksisuusluokan, ja sillä voidaan ratkaista joitain ongelmia tehokkaammin kuin klassisella tietokoneella. Church-Turingin teesiä se ei kuitenkaan kumoa, se ei kykene ratkaisemaan mitään sellaista ongelmaa, mitä klassinen tietokone ei kykene.

Kvanttimekaaninen argumentti ei näytä nykyvalossa kovinkaan vahvalta, ainakaan se ei kykene horjuttamaan komputationalismia. Korkeintaan olisi tarpeellista laajentaa komputationalismin määritelmää siten, että se hyväksyisi myös kvanttikomputationaaliset prosessit tietoisuuden lähteeksi. Monet tutkivat uskovat joka

tapauksessa, että ihmisen käyttäytymisen täydelliseen selittämiseen riittävät neurofysiologiset tulokset, eikä kvanttimekaanisia ilmiöitä tarvitse ottaa huomioon.

Kuitenkaan kvanttimekaanista argumenttia ei kannata täysin unohtaa. Vaikka neurofysiologiset tulokset riittäisivätkin ihmisen käyttäytymisen selittämiseen, kvanttimekaanisilla ilmiöillä saattaisi silti olla jonkinlainen rooli tietoisuuden synnyssä. Monet tutkijat ovat ehdottaneet kvanttimekaanisia ilmiöitä ratkaisuksi kvalioiden, mielen ykseyden ja vapaan tahdon ongelmaan. Kuten aiemmin on todettu, fysikalismilla on suuria ongelmia selittää nimenomaan näitä ilmiöitä.

Kvanttimekaniikka saattaisi avata oven jonkinlaisella dualismille, jolle näiden ominaisuuksien selittäminen ei ole yhtä suuri ongelma. Henry Stappin (1996, 204) mukaan nimittäin kaikki yleisimmät tulkinnat kvanttimekaniikan luonteesta ovat pohjimmiltaan dualistisia. Varsinkin perinteisessä Kööpenhamina-tulkinnassa havaitsemisella on oleellinen rooli. Toisin sanoen, jonkinlainen tietoinen tapahtuma on sisäänrakennettuna kvanttimekaaniseen teoriaan, toisin kuin klassisessa fysiikassa. Tämä tietoinen tapahtuma liittyy juuri kvanttimekaaniseen mittaustapahtumaan, joka tarkoitti siis sitä, että fysikaalisen systeemin tila on aluksi usean tilan superpositio ja vakiintuu yhdeksi ominaistilaksi vasta kvanttimekaanisessa mittauksessa. Kuuluisin esimerkki tästä kvanttimekaniikan erityispiirteestä on Schrödingerin kissa -paradoksi. Siinä kissa on lukittu laatikkoon. Mittalaite mittaa tietyn elektronin spin-arvoa. Jos spin-arvo on "up", niin kissa tapetaan, jos se taas on "down", niin kissa säilyy hengissä. Tällöin seuraa, että jos spin-arvo on aluksi näiden kahden arvon superpositio, niin silloin myös kissa siirtyy superpositioon, joka on kuoleman ja elämän kombinaatio. Vasta kun laatikko avataan ja tietoinen olio havaitsee kissan, määräytyy eläkö kissa vai onko se kuollut. Perinteisen kvanttimekaniikan tulkinnan mukaan tällainen paradoksaalinen tilanne todellakin syntyy. On olemassa myös monia muita tulkintoja kvanttimekaniikan luonteesta, jotka saattavat välttyä vastaavanlaisilta paradokseilta, Stappin mukaan joka tapauksessa lähes kaikki muutkin kvanttimekaniikan tulkinnat sisältävät dualistisen elementin.

Jos tämä dualistisuus todella on läsnä kvanttimekaniikassa, antaa se dualismille mahdollisuuden uuteen tulemiseen. Kun tutkimme kappaleessa 3.2 dualismia,

päällimmäinen syy hylätä se, oli ongelmat mentaalisen ja fysikaalisen vuorovaikutuksessa. Kvanttimekaniikka saattaisi tarjota keinoja, joilla nämä ongelmat voitaisiin korjata. Nimenomaan kvanttikoherenssi ja superpositio voisivat olla ilmiöitä, jotka voisivat olla ratkaisu mentaalisen ja fysikaalisen vuorovaikutusongelmaan. Kaikki tämä on kuitenkin hyvin spekulatiivista keskustelua, eikä todennäköisesti ole olemassa empiirisiä keinoja testata näitä hypoteeseja. Halukas voi tutustua teorioihin kvanttimekaanisesta mielestä esimerkiksi Chalmersin (1996), Stappin (2006) tai Deutschin (1997) avulla.

Jos palaamme komputationalismiin ja mietimme sen suhdetta näihin teorioihin, niin voimme todeta, että se voi aivan hyvin olla yhteensopiva myös tällaisten kvanttimekaanisten mielen teorioiden kanssa, riippuen tietysti hyvin paljon teorian yksityiskohdista. Kuten muistamme, komputationalismi on täysin yhteensopiva myös esimerkiksi dualismin kanssa, jolloin tällaisetkin spekulatiiviset teoriat mahtuvat sen alle. Tästä on esimerkkinä Chalmersin kvanttimekaaninen mielen teoria (1996), jossa on selvästi piirteitä sekä dualismista että komputationalismista.

5.4 Searlen tekoälyn kritiikki

Dreyfusin tekoälyn kritiikki ei ole ehkä kovin helposti ymmärrettävää tai yksinkertaista. Sama pätee gödeliläisiin ja kvanttimekaanisiin argumentteihin. Tämän vuoksi ne ovat jääneet ehkä hieman vähemmälle huomiolle tekoälyn filosofiassa. Sen sijaan John Searlen vuonna 1980 artikkelissaan Minds, Brains and Programs esittelemä kiinalainen huone -argumentti oli hämmästyttävän yksinkertainen argumentti tekoälyä vastaan, ja sen myötä se sai valtavasti huomiota. Niinkin paljon, että moni mielenfilosofi ei suostu nykyään enää sanallakaan kommentoimaan tuota argumenttia. Argumentti oli yksi viime vuosisadan eniten keskustelua herättäneistä mielenfilosofian argumenteista. Se sai, ja saa edelleen valtavasti kritiikkiä sekä puolesta että vastaan.

Searle kohdisti argumenttinsa suoraan luvussa 2.3 esiteltyä Turingin testiä vastaan. Turingin testissä kone hyväksyttiin aidosti tietoiseksi, jos se kykenee keskustelemaan ihmisen kanssa siten, ettei ihminen huomaa keskustelevansa koneen kanssa. Searle pyrki kiinalainen huone -argumenttinsa avulla osoittamaan, että vaikka kone kykeneekin

tällaiseen keskusteluun, ei se silti riitä perusteeksi olettaa, että kone olisi aidosti tietoinen. Searle myönsi kyllä sen, että pelkkä komputaatio saattaa riittää siihen, että tietokone kykenee täydellisesti simuloimaan ihmisen käyttäytymistä. Hän siis hyväksyi Dreyfusin kritisoiman epistemologisen oletuksen. Searle oli kannassaan siis lievempi kuin ne, jotka uskovat gödeliläisiin tai kvanttimekaanisiin argumentteihin. Heidänhän mielestään koneet eivät koskaan kykyne edes simuloimaan ihmisen käyttäytymistä, koska ei-laskennalliset prosessit ohjaavat käyttäytymistämme. Vaikka aitoa tietoisuutta tietokoneelle ei Searlen mukaan pystytä luomaan, piti hän kuitenkin tietokonetta ja tekoälysovelluksia hyvinä välineinä tutkia mieltä ja kognitiota. Hän kutsui tätä näkemystään nimellä heikko tekoäly.

Kiinalainen huone -argumentin tarkoitus on osoittaa vääräksi teesi, että ihmisen mieli ja tietoisuus olisivat vain komputaatiota. Näkemystä, joka hyväksyy tämän teesin, Searle kutsui vahvaksi tekoälyksi, erotukseksi heikosta tekoälystä. Ja tämä teesihän vastaa Dreyfusin kritisoimaa psykologista oletusta. Argumentti pyrkii tähän päämäärään kumoamalla ensin ontologisen oletuksen. Eli se pyrkii osoittamaan, että maailmaa ei kyetä täysin analysoimaan kontekstittomien ja atomisten faktojen avulla. Jos argumentti kykenee tähän, seuraa tästä Searlen mukaan myös se, ettei tietoisuutta voida luoda pelkästään komputaatiolla, joka on nimenomaan vain tällaisten kontekstittomien faktojen käsittelyä.

Yksi syy siihen, miksi Searle muotoili argumenttinsa oli se, että 1970-luvun lopulla tietokoneiden kehittyessä tietyt tekoälypiirit väittivät kehittäneensä ohjelmia, jotka aidosti ymmärsivät englanninkielisiä lauseita käyttämällä hyväkseen tietokantaa englannin kieliopista, sanastosta sekä laajasta kokoelmasta taustatietoja siitä, millaisia käsitteellisiä suhteita sanoilla on toisiinsa. Nämä tiedot on annettu ohjelmille luonnollisesti täysin formaalissa muodossa. Tällaisten ohjelmien katsottiin tietyissä (helpotetuissa) tilanteissa pystyvän selviytymään Turingin testistä.

Tällainen oli esim. Roger Schankin (Schank & Abelson, 1977) kehittämä ohjelma, joka käytti skriptejä kuvaamaan käsitteellisiä suhteita. Annetaan tällaiselle ohjelmalle syötteeksi seuraavanlainen tarina: "Mies meni ravintolaan ja tilasi hampurilaisen. Kun hampurilainen tuotiin pöytään, se oli paistunut korpuksi, ja mies ryntäsi vihaisena ulos

ravintolasta jättäen maksamatta." Kun ohjelmalta nyt kysytään: "Söikö mies hampurilaisen?", ohjelma osasi vastata samoin kuin ihmisetkin luultavasti vastaisivat, eli "Ei syönyt".

Searle väitti, että Schankin ohjelman kaltaiset ohjelmat eivät oikeasti ymmärtäneet lauseita, vaan ne toimivat pelkästään formaalien sääntöjen mukaisesti. Searle halusi löytää analogian Schankin ohjelmalle, mutta joka kuitenkin olisi ilmeisellä tavalla kykenemätön tietoisuuteen. Tästä seuraisi hänen mukaansa, että ei pelkästään Schankin ohjelma, vaan kaikki mahdolliset komputaatioon perustuvat systeemit olisivat kykenemättömiä tietoisuuteen. Kiinalainen huone -ajatuskoe pyrkii olemaan juuri tällainen analogia Schankin ohjelmalle.

Searlen argumentin vetovoima perustuu sen yksinkertaisuuteen ja ymmärrettävyyteen, sekä siihen, että se on intuitiivisesti ja maalaisjärjellä ajateltuna hyvinkin uskottava. Siitä, ovatko Searlen itsensä tekemät oletukset oikeita vai vääriä, on kuitenkin paljon erilaisia näkemyksiä. Käytän loput luvusta 5 tämän argumentin käsittelyyn, koska se havainnollistaa hyvin tekoälyn ongelmia, ja siihen on helppo soveltaa useita erilaisia näkemyksiä tekoälyn vaatimuksista ja niiden ratkaisutavoista.

5.4.1 Kiinalainen huone -argumentti

Searle muodosti argumenttinsa seuraavasti: Oletetaan, että ihminen on lukittuna huoneeseen, joka on täynnä kiinankielistä kirjoitusta. Kyseinen ihminen ei kuitenkaan ymmärrä sanaakaan kiinankielisestä kirjoituksesta eikä puheesta. Oletetaan, että hän sen sijaan hallitsee englanninkielisen tekstin sekä puheen. Kiinankielisten kirjoitusten lisäksi huoneessa on valtavasti englanninkielisiä ohjeita siitä kuinka kiinankielisiä symboleita on mahdollista yhdistellä toisiinsa ja millaisia formaaleja suhteita niillä on toistensa kanssa. Symboleiden merkityksiä ei sen sijaan ole kerrottu missään vaiheessa. Nyt huoneeseen tuodaan erä kiinankielisiä kysymyksiä, joihin huoneessaolijan on tarkoitus vastata kiinankielisillä symboleilla, käyttämällä apunaan vain näitä kiinankielisiä kirjoituksia ja englanninkielisiä ohjeita. (Searle, 1980, 418.)

Nyt oletetaan, että nämä englanninkieliset ohjeet, joita Searlen mukaan voidaan kutsua ohjelmaksi, ovat niin hyviä, että huoneessaolija kykenee vastaamaan huoneeseen annettuihin kysymyksiin siten, etteivät ulkopuolella olevat kiinan kieltä osaavat kysymystenlaatijat pysty erottamaan henkilön antamia vastauksia äidinkielenään kiinaa puhuvan henkilön antamista vastauksista. Ainoa ero on siinä, että kiinaa puhuva henkilö ymmärtää käsittelemiensä symbolien merkityksen, kun taas huoneessaolija ei ymmärrä sanaakaan käydystä keskustelusta. Hän vain yhdistelee hänelle merkityksettömiä formaaleita symboleita toisiinsa ja palauttaa ne vastauksena huoneen ulkopuolelle. Huoneessaolija toimii tavallisen tietokoneen, eli sarjakäsittelynä toimivan Turingin koneen tavoin. Hän on tietokoneohjelman instanssi.

Tästä voidaan Searlen mukaan helposti huomata, että vaikka huoneessaolija toimii kuten kiinaa puhuva henkilö, toisin sanoen hänellä on samat syötteet ja tulosteet, hän ei silti ymmärrä keskustelusta mitään, eikä näin ollen ymmärrä vastaava Schankin tms. tietokonekaan, koska sillä ei ole mitään sellaista piirrettä jota huoneessaolijalla ei ole.

Tästä taas voidaan Searlen mukaan päätellä, etteivät tietokoneet kykene samankaltaiseen ymmärtämiseen kuin ihmiset. Ihmisen mieli ei toimi pelkästään sääntöjä noudattamalla, eli se ei ole laskennallinen prosessi. Laskennalliset operaatiot, jotka kohdistuvat formaaleihin symboleihin, eivät siis ole riittävä ehto ymmärryksen luomiseen. Eivätkä ne Searlen mukaan ole edes välttämätön tai merkittävä osa ymmärrystä. Hän pitää laskennallisuutta täysin merkityksettömänä asiana ihmisen mielen toiminnassa. (Searle, 1980, 418.)

Searlen kritiikki kohdistuu siihen, että huoneessaolijalla ei ole mitään keinoa liittää merkityksiä käsittelemiinsä symboleihin. Eikä siis vastaavalla tietokoneellakaan ole tätä kykyä. Kiinalainen huone -ajatuskoe on siis argumentti Dreyfusin kritisoimaa ontologista oletusta vastaan. Ei ole mitään keinoa miten tietokone voisi löytää kontekstittomista formaaleista symboleista niiden merkityksen. Ihmismieli sen sijaan on Searlen mukaan tietoinen näistä merkityksistä: tiedän mitä sana "kissa" merkitsee, tiedän mihin se viittaa. Kyseisen vajaavuuden takia tietokoneet eivät Searlen mukaan koskaan voi aidosti olla tietoisia. Hän on esittänyt päättelynsä myös seuraavanlaisena argumenttina (Searle, 1984):

1. Ohjelmat ovat puhtaasti formaalisia (syntaktisia).
2. Ihmisen mielellä on mentaalinen sisältö (semantiikka).
3. Syntaksi yksinään ei ole riittävä synnyttämään semantiikkaa.
4. Siksi ohjelman implementaatio ei yksinään ole riittävä synnyttämään mieltä.

Kiinalainen huone -ajatuskokeen on tarkoitus tukea kolmatta premissiä osoittamalla, että ihminen, joka käsittelee vain formaaleja symboleita, ei kykene ymmärtämään niiden semantiikkaa. Hänellä ei ole keinoa löytää niiden merkityksiä. Päättely saattaa päällepäin vaikuttaa varsin pätevältä, mutta itseasiassa sen jokaista premissiä ja myös johtopäätöstä täytyy tutkia hyvin tarkasti ennen kuin ne voidaan hyväksyä. Jokainen näistä lauseista nimittäin voidaan kyseenalaistaa hyvin perustein. Syitä tähän käsiteltiin jo hieman Dreyfusin kritiikkiä tutkiessa, mutta seuraavissa luvuissa tullaan käsittelemään näitä vielä huomattavasti lisää. Luku 6 käsitteleekin käytännössä pelkästään kolmatta premissiä.

Ennen kuin Searle julkaisi artikkelinsa, jossa hän julkaisi kiinalainen huone -argumentin, hän esitteli sen ensin kuitenkin useille eri yliopistoissa työskenteleville tekoälyn tutkijoille. Hän kohtasi paljon kritiikkiä ja sai vastaukseksi useita erityyppisiä vasta-argumentteja, jotka pyrkivät tukemaan vahvan tekoälyn mahdollisuutta. Nämä julkaistiin artikkelin yhteydessä, ja Searle pyrki myös vastamaan kaikkiin näihin reaktioihin. Lähden käsittelemään nyt näitä vasta-argumentteja.

Vasta-argumentit voidaan jakaa karkeasti kolmeen tyyppiin (Cole, 2004). Ensimmäiset hyväksyvät, että kiinalaisessa huoneessa oleva mies ei ymmärrä kiinaa, mutta jokin muu ymmärtää. Tällä tarkoitetaan yleensä, että jokin suurempi kokonaisuus, kuten koko huone, ymmärtää kiinaa, vaikka mies ei ymmärräkään. Tätä tyyppiä edustavat systeemivastaus ja virtuaalinen mieli -vastaus. Toisen tyypin vastaukset hyväksyvät, että kiinalainen huone sellaisenaan ei ymmärrä kiinaa, mutta jonkinlainen modifikaatio siitä voisi ymmärtää. Tällaisia vastauksia ovat muun muassa robottivastaus ja aivosimulaattorivastaus. Kolmannen tyypin vastaukset väittävät, että itseasiassa mies kiinalaisessa huoneessa ymmärtää kiinaa, toisin kuin Searle väittää. Nämä vastaukset perustelevat näkemyksensä esimerkiksi sillä, että tässä tapauksessa intuitiomme pettää. Kaikki riippuu siitä, miten määrittelemme ymmärtämisen.

5.4.2 Systeemivastaus

Ehkä useimmiten esitetty vastaus on systeemivastaus. Sen mukaan on totta, että kiinalaisessa huoneessa oleva ihminen ei ymmärrä kiinaa, mutta hän on vain osa koko systeemiä ja systeemi kyllä ymmärtää. Hän on siis vain keskusyksikkö laajemmassa kokonaisuudessa, joka sisältää myös kiinankieliset kirjoitukset eli tietokannan ja muistin sekä ohjeet eli ohjelmakoodin.

Tälle väitteelle Searlella on yksinkertainen vastaus (1980, 419): sehän olisi mieletöntä. Ja todellakin, arkijärjen mukaanhan näin on. Miten ihmeessä jonkinlainen ihmisen ja paperipinon konjunktio kykenisi olemaan tietoinen. Lisäksi Searle esittää seuraavanlaisen lisäargumentin: Ihminen voi teoriassa sisäistää koko systeemin painamalla mieleensä kaikki ohjeet ja kiinankieliset symbolit sekä suorittamalla kaikki operaatiot päässään. Nyt ihminen on sisäistänyt koko systeemin, eikä hän siltikään ymmärrä mitään kiinan kielestä, eikä ymmärrä myöskään systeemi, sillä systeemi on vain osa häntä. Hänellä ei ole edelleenkään keinoa liittää merkityksiä kiinankielisiin symboleihin.

Meidän täytyy nyt palata siihen mikä komputationalismin mukaan oli vaatimus tietoisuuden tai ymmärryksen syntyyn. Tietoisuus oli sen mukaan abstraktien symbolien formaalia manipulointia formaalien sääntöjen mukaisesti. Toisin sanoen tietty mentaalinen tila on Turingin koneen tai vastaavan formaalin systeemin kokonaistila, koostuen tilataulukoiden yms. yhteistilasta. Tästä voimme suoraan tehdä johtopäätelmän: Searlen argumentti ei ole pätevä, se ei sovellu komputationalismin kumoamiseen. Komputationalismi ei ole koskaan väittänytkään, että tietoisuus syntyisi jotenkin siihen Turingin koneen kirjoitus-/lukupäähän, joka operoi muistinauhaa, vaan tietoisuus syntyy Turingin koneen kokonaistilasta, johon kuuluvat kirjoitus- ja lukupää, muistinauha, sekä tilataulukko. Huoneessaolija, ja varsinkin huoneessaolijan tietoisuus, ovat siis täysin väärällä funktionaalisen organisaation tasolla, jotta niihin voisi syntyä ymmärrys.

Se, että ihminen sisäistäisi koko systeemin painamalla mieleensä kaikki ohjeet ja kiinankieliset symbolit, ja suorittaisi kaikki operaatiot päässään, ei auta asiaa ollenkaan.

Tämä suoritus tapahtuisi silti väärällä funktionaalisen organisaation tasolla. Kyseinen suoritus olisi tavallaan yhden kerroksen ihmisen tietoisuuden yläpuolella, eikä näin ollen ymmärrystä kiinankielestä voisi syntyä tälle alemmalle tasolle. Jotta kiinankielen ymmärtäminen voisi syntyä ihmisen mieleen, täytyisi tämän ymmärrykseen johtavan prosessin ujuttautua jollakin tavalla siihen prosessiin, joka tuottaa ihmisen tietoisuuden tällä alemmalla tasolla. Käytännössä tämä vaatii siis sen, että ihmisaivojen neuroneissa tapahtuisi muutoksia siten, että ne kykenisivät suorittamaan tämän prosessin. Ja näinhän juuri tapahtuu silloin kuin opettelemme normaalilla tavalla uuden kielen.

Tämä ei tietenkään estä sitä, etteikö kiinankielen ymmärrystä voisi syntyä tällä ylemmällä organisaation tasolla. Tällöin sillä ei vain olisi vaikutusta kyseisen ihmisen normaaliin ymmärrykseen. Tällaiselle ajatukselle on ainakin jotain perustetta, sillä suuri osa aivojemme prosesseistahan tapahtuu täysin tiedostamattamme. Esim. valtaosa ruuminosiemme liikkeistä tapahtuu tiedostamatta. Tämän ajatuksen voi myös muotoilla seuraavan virtuaalinen mieli -vastauksen muotoon, joka käyttää mallinaan tietokoneen ja virtuaalikoneen suhdetta. Tietokone voi normaalisti suorittaa (tai "ymmärtää") tiettyä käskykantaa, mutta se voi myös emuloida toista käskykantaa. Toisin sanoen sama tietokone voi toteuttaa yhden tai useampia virtuaalikoneita, joilla on täysin erilaiset ominaisuudet. Tällainen sisäistetty kiinalainen huone -systeemi on siis vain erillinen virtuaalikone tai mieli ihmisen aivoissa, eikä sillä ole vaikutusta ihmisen normaaliin ajatteluun. Ja kuten muistamme, komputationalismille tällainen ajatus on täysin normaali. Voimme myös kysyä, mitä psykologia ajattelee tällaisesta näkemyksestä, että ihmisaivoissa voisi olla monta erillistä mieltä. Psykologia ei välttämättä ole kovinkaan yllättynyt tällaisesta tuloksesta. Historia nimittäin tuntee paljon tapauksia sivupersoonahäiriöistä, joissa ihmisen persoona on jakautunut kahteen tai useampaan eri persoonallisuuteen, jotka voivat olla toisistaan tietämättömiä ja joista jokin aina ajoittain ottaa ihmisen valtaansa.

Itseasiassa Searlen suurin argumentti kiinalaisen huoneen ymmärrystä vastaan onkin se, että se tuntuu mielettömältä. Näin kieltämättä on. Ihmisen ja paperipinon kollektiivinen tietoisuus tuntuu jotenkin mielettömältä. Meidän täytyy kuitenkin muistaa, että myös se, miten ihmisaivojen harmaiden aivosolujen konjuktiosta voi syntyä tietoinen mieli, on aivan yhtä suuri mysteeri (ks. 2.2). Näin kuitenkin tapahtuu.

On selvää, että Searlen argumentin uskottavuus perustuu pitkälti intuitioon. On intuitiivista ajatella, että tietyt asiat eivät kykene ajattelemaan. Intuitiot pitäisi kuitenkin jättää syrjään, koska tieteen kehitys saattaa muuttaa myös intuitioitamme. Keskustelu kiinalainen huone -argumentin pätevyydestä päättyy umpikujaan jo siinä vaiheessa, kun tekoälyn kannattajat sanovat, että kiinalainen huone ymmärtää, mutta vastustajat pitävät ajatusta naurettavana.

Esimerkiksi Dennett (1987, 326) on kritisoinut Searlea tavasta, jolla hän johdattelee lukijaansa intuitioidensa perusteella. Hän kutsuukin Searlen ajatuskokeita intuitiopumpuiksi. Olemme tottuneet siihen, että tietoisuuden täytyy syntyä prosesseissa, jotka tapahtuvat hyvin pienissä ja monimutkaisissa neuroverkoissa, ja kaikenlisäksi äärimmäisen nopeasti. Kiinalaisessa huoneessa sen sijaan Turingin koneen virkaa hoitelevat ihminen, sekä kasa paperinpalasia. Kaiken lisäksi ihminen operoi äärimmäisen hitaasti. Jos huoneessa toimisi vain yksi ihminen, sopivien vastauksien tuottaminen kiinankielisiin kysymyksiin kestäisi lähes äärettömän kauan. Ei tunnu arkijärjellä ajatellen mahdolliselta, että tällainen systeemi voisi olla tietoinen. Dennett muistuttaa, että hitaat ajattelijat ovat tyhmiä eivätkä älykkäitä. Hänen mielestään nopeus on olennainen osa älykkyyttä, sillä jos jokin systeemi ei pysty tarpeeksi nopeasti toimimaan luonnon muuttuvissa olosuhteissa, ei se ole älykäs riippumatta siitä, kuinka monimutkainen sen rakenne on. Älykkyys on siis suhteellista verrattuna vallitseviin olosuhteisiin. Tosin tässä täytyy erottaa älykkyys tietoisuudesta, sillä olio voi toki olla tietoinen olematta älykäs. Perinteisessä komputationalismissa komputaation nopeus ei nimittäin vaikuta mitenkään tietoisuuden syntyyn.

Kiinalainen huone -argumentti ei siis vaikuta pätevältä. Ei kuitenkaan unohdeta sitä heti. Systeemivastauksen tarkoitus on lähinnä osoittaa, että Searlen argumentti ei ole muodoltaan pätevä kumoamaan vahva tekoäly. Useimmat systeemivastauksen kannattajat eivät välttämättä kuitenkaan hyväksy sitä, että kiinalainen huone sellaisenaan kykenisi aidosti ymmärtämään kiinankieltä, vaikka kiinalainen huone -argumentti ei kykenekään osoittamaan tätä. He siis ovat samaa mieltä Searlen kanssa siinä, että kiinalainen huone ei kykene aidosti ymmärtämään. Heidän syynsä tähän tulokseen ovat kuitenkin erilaiset. Robottivastaus on monien tutkijoiden antama

modifikaatioehdotus, jonka toteuttamalla kiinalainen huone voisi saavuttaa aidon tietoisuuden.

5.4.3 Robottivastaus

Searlen argumentti pohjautuu siis siihen, että hänen mukaansa kiinalainen huone ei voi ymmärtää käsittelemiensä symboleiden merkityksiä. Robottivastauksen kannattajat myöntävät, että kiinalainen huone sellaisenaan ei kykene löytämään merkityksiä sanoille, mutta vain siksi, että sillä ei ole oikeanlaista kausaalista yhteyttä reaalimaailmaan. Ajatellaan kuitenkin, että tällainen systeemi sijoitettaisiin robotin sisään, joka ei saisi syötteekseen pelkästään formaaleja symboleita vaan jolla olisi aistisensoreita, kuten videokamera ja mikrofoni, joilla se voisi aistia ympäristöään, sekä esimerkiksi kädet ja jalat, joilla se voisi vaikuttaa ympäristöönsä. Tällöin sillä olisi kausaalinen yhteys reaalimaailmaan ja se voisi oppia näkemästään ja tekemästään kuten lapsi ja liittää merkityksiä symboleihin. Tällainen robotti voisi robottivastauksen mukaan aidosti ymmärtää esimerkiksi kiinaa.

Searlen (1980, 420) mukaan aistimuksien ja motoriikan lisääminen tietokoneeseen ei tuo siihen yhtään sen enempää ymmärrystä. Hänen mukaansa sama kiinalainen huone -argumentti pätee myös tähän tapaukseen vain vähän muunneltuna. Kuvitellaan sen sijaan, että tietokone olisi robotin sisällä, jälleen kerran, että ihminen on huoneessa. Tällä kertaa symbolit, jotka tulevat huoneeseen, ovat vain peräisin videokameroista, jotka muokkaavat informaation formaaleiksi symboleiksi. Ihmisen vastaukset taas menevät huoneen motoriikasta huolehtivalle koneistolle. Taas voidaan Searlen mukaan huomata, että vaikka huoneella on kausaalinen yhteys ulkomaailmaan, eli se kykenee aistimaan ja myös vaikuttamaan ulkomaailmaan, ei sen sisällä oleva ihminen kuitenkaan ole lainkaan tietoinen näistä asioista vaan pelkästään manipuloi formaaleja symboleita.

Tämä on tietysti totta, kuten systeemivastauksen perusteella voimme huomata. Ihminen huoneessa ei edelleenkään ymmärrä mitään, mutta näin sen tulee vahvan tekoälyn mukaan ollakin. Ymmärtävä systeemi ei ole ihminen huoneessa, vaan koko systeemi, johon kuuluvat ihminen, paperipino, sekä nyt huoneeseen liitetyt motoriset ja sensoriset

laitteet. Näin ollen systeemivastaus kumoaa myös Searlen robottivastaukseen antaman kritiikin.

Kuten Stevan Harnad (1989) on huomauttanut: Jos ihminen huoneessa vain käyttää kyseisiä motorisia ja sensorisia laitteita, hän ei simuloi täysin niiden toimintaa, jolloin hän ei myöskään voi saada ymmärrystä niiden kautta. Jotta hän voisi käyttää näiden laitteiden kausaalisia voimia ja saada ymmärrys niiden kautta, hänen täytyisi kirjaimellisesti *olla* nämä laitteet.

Mikä siis oli se asia, minkä takia systeemivastauksen lisäksi on annettu myös robottivastaus? Robottivastauksen mukaan kiinalainen huone -systeemi saa ymmärryksen merkityksistä näiden aistisensorien ja motoristen laitteiden avulla. Miten tämä ymmärryksen synty sitten tapahtuu näiden kausaalisten linkkien avulla? Tulemme nyt merkittävään filosofiseen kysymykseen merkityksistä, eli semantiikasta. Tähän kysymykseen on periaatteessa kahdenlaisia vastauksia, sillä semantiikan luonteesta on kaksi erilaista koulukuntaa: eksternalismi ja internalismi.

Eksternalismin mukaan symboleiden merkitykset määräytyvät ainakin osittain mielen ulkopuolisten tekijöiden perusteella. Tällaista näkemystä on kannattanut esimerkiksi Putnam (1975), jolta on peräisin kuuluisa lause: "Meanings just ain't in the head". Tässä näkemyksessä ihmismielen symbolit saavat merkityksensä jonkinlaista ulkomaailmaan ulottuvaa kausaalista linkkiä pitkin. Se mikä tämä linkki on käytännössä, on hyvin vaikea kysymys. Internalismissa sen sijaan symbolit eivät tarvitse tällaista linkkiä, vaan symbolit saavat merkityksensä suhteestaan muihin symboleihin. Internalismin mukaan siis ihmismielen symbolit ovat vain jonkinlaisia representaatioita ulkomaailmasta, ja ne saavat merkityksensä suhteestaan muihin representaatioihin. Internalismi ei välttämättä kiellä, etteikö näillä symboleilla voisi olla myös kausaalista linkkiä ulkomaailmaan, useimmiten näin onkin. Internalismin mukaan symboleiden merkitys ei kuitenkaan tule tästä linkistä, vaan pelkästään suhteesta muihin symboleihin. Esimerkiksi William Rapaport (1988) on kannattanut internalismia.

Voimme todeta, että Searle olettaa jonkinlaisen internalismin pätevän, ja hänen vastauksensa robottivastaukseen on itseasiassa argumentti tämän puolesta. Hän ei usko,

että kausaalinen linkki ulkomaailmaan voisi tuoda merkityksiä tietokoneelle. Se voi tuoda korkeintaan lisää formaaleita symboleita. Vaikka Searle ei siis usko, että kausaalinen linkki ulkomaailmaan tuottaa semantiikan, hän ei kuitenkaan kerro mikä ihmisellä semantiikan tuottaa. Hän vain väittää, että ihmisaivoissa on jokin biologinen ominaisuus, joka tämän mahdollistaa, mutta tämän tarkemmin hän ei sitä kerro. Tätä ominaisuutta ei voi kuitenkaan Searlen mukaan olla laskennallisilla systeemeillä, minkä kiinalainen huone -argumentti taas yrittää osoittaa. Tämän ominaisuuden täytyy kuitenkin olla hyvin mystinen. Se ei voi olla kausaalinen linkki ulkomaailmaan, koska Searle tämän kieltää, mutta se ei voi olla myöskään linkki mielen sisäisten representaatioiden välillä, koska tällöin ne olisivat simuloitavissa tietokoneella. Tulee mieleen, että Searle pitää merkityksiä biologisille systeemeille sisäisinä ominaisuuksina, itsessään merkityksellisinä. Mutta tämähän on juuri se väite, mistä Dreyfus syytti komputationalismia!

Joka tapauksessa, toimii semantiikka sitten eksternalistisesti tai internalistisesti, on uskottavampaa, että sen täytyy toimia näin sekä ihmisillä, että tietokoneilla. Ei ole mitään syytä olettaa, että semantiikka toimisi ihmisillä ja tietokoneilla eri tavoin. Joko kausaalinen linkki mielen sisältä ulkomaailmaan voidaan luoda sekä ihmisille, että tietokoneille, tai sitten semantiikka syntyy vain mielen sisäisten representaatioiden välisistä suhteista (Rapaport, 1988, 87). Luvussa 6 käsitellään miten semantiikka näistä suhteista voisi syntyä.

Komputationalismi on siis periaatteessa yhteensopiva molempien semantiikan teorioiden kanssa, vaikka perinteinen komputationalismi väittääkin, että semantiikka syntyy vain mielen sisäisten representaatioiden suhteesta. Palataan nyt kuitenkin kysymykseen, miksi komputationalistit ovat ylipäänsä esittäneet robottivastauksen systeemivastauksen lisäksi. Ne komputationalistit, jotka pitävät eksternalismia uskottavampana vaihtoehtona, tarvitsevat robottivastausta yksinkertaisesti siihen, että tietokoneen käyttämät symbolit voisivat saada merkityksensä ulkomaailmasta.

Mihin internalismia kannattavat komputationalistit sitten voisivat tarvita robottivastausta? Totesimme luvuissa 5.1.5 ja 5.1.6, että riittävän suuren kontekstin, toisin sanoen merkitysten viidakon, ohjelmoiminen käsin tietokoneelle voi olla

mahdoton tehtävä. Sen sijaan jos tietokoneelle annetaan aisti- ja motoriset elimet, se voi kasvaa ihmisten parissa ja oppia kuten pieni lapsi ympäristöstään. Tällöin sen semanttinen verkosto voisi hitaasti kasvaa, jolloin se oppisi hitaasti myös ymmärtämään paremmin symboleiden merkityksen. Näin ollen tässä tapauksessa robottivastaus on annettu vain käytännöllisistä, eikä periaatteellisista syistä.

Merkitysten ja semantiikan rooli on joka tapauksessa hieman epämääräinen tietoisuuden synnyssä. Esimerkiksi Putnam (1975) tuntuu sanovan, että merkitykset ovat jollakin tapaa määritelmällisesti riippuvaisia ulkomaailmasta. Jos käsite merkitys tai semantiikka halutaan määritellä tällä tavoin, ei ole kuitenkaan enää selvää haluaako esimerkiksi komputationalismi pitää tällaista käsitystä semantiikasta osana mielen teoriaansa. Tällöin komputationalismi voisi esimerkiksi väittää, että tällaiset eksternaaliset merkitykset eivät ole enää välttämättömiä tietoisuuden ja nimen omaan kvalioiden syntyyn, vaan välttämättömiä olisivat vain mielen sisäiset representaatiot, jotka eivät tarkoita samaa asiaa kuin näiden symbolien eksternaaliset merkitykset. Representaatiot ovat vain vastinkappaleita ulkomaailman objekteihin, eikä niillä ole välttämätöntä kausaalista suhdetta toisiinsa. Itse asiassa tämä tuntuukin olevan monen filosofin näkemys, vaikka sitä ei ääneen sanotakaan. Semantiikasta keskustellaan joka tapauksessa huomattavasti lisää luvussa 6.

5.4.4 Aivosimulaattorivastaus

Aivosimulaattorivastaus käskee meitä kuvittelemaan ohjelman, joka ei pyri kuvaamaan informaatiota maailmasta Schankin ohjelman tavoin (joka siis käytti skriptejä kuvaamaan käsitteellisiä suhteita), vaan ohjelman, joka yksinkertaisesti simuloi täydellisesti aitoja kiinaa puhuvan ihmisen aivojen hermoverkkoja ja synapsien lähettämiä signaaleita. Kone ottaa vastaan kiinankielisiä kysymyksiä ja simuloi aivojen formaaleja prosesseja tuottaen kiinankielisiä vastauksia. Voidaan ajatella myös, että tämä ohjelma suoritetaan rinnakkain siten kuin aivot toimivat eikä sarjakäsittelynä, kuten normaalisti tietokoneohjelmat suoritetaan perinteisissä yksiprosessorisissa koneissa. Nyt kysymys kuuluu: Eikö tällöin voida sanoa koneen ymmärtävän kirjoituksia, sillä jos me kiellämme sen, meidän täytyy väittää, etteivät tavalliset kiinaa puhuvat ihmisetkään ymmärrä kirjoituksiaan.

Searle esittää jälleen vasta-esimerkin (1980, 420): Sen sijaan, että ihminen kiinalaisessa huoneessa sekoittelisi kiinankielisiä symboleita, hän operoi monimutkaista vesijohtoverkkoa aukaisemalla ja sulkemalla siinä olevia venttiilejä. Kun hän vastaanottaa kysymyksen, hän katsoo ohjeista, mitkä venttiilit hänen täytyy aukaista ja mitkä sulkea. Vesijohtoverkko kuvaa aivojen synapseja ja vastaus kysymykseen putkahtaa ulos toisesta päästä verkkoa. Nyt Searle jälleen kysyy: Missä on tämän systeemin ymmärrys? Se tuottaa kyllä oikeat vastaukset kysymyksiin, mutta ainakaan ihminen ei taaskaan ymmärrä kiinaa, eivätkä myöskään vesiputket. Jos joku haluaa väittää, että ihmisen ja vesiputkien konjunktio jollain tapaa ymmärtää kiinaa, voimme periaatteessa taas painaa mieleemme systeemin formaalin rakenteen ja kuvitella synapsien signaalit mielessämme, Searle kommentoi viitaten systeemivastaukseen, emmekä silti ymmärrä mitään kiinan kielestä. Simuloimme jälleen vain aivojen formaaleja ominaisuuksia pääsemättä käsiksi aivojen kausaalisiin ominaisuuksiin, jotka tuottavat Searlen mukaan ajatuksemme.

Searlen vasta-esimerkki on läheistä sukua Ned Blockin jo aiemmin (1978) esittämälle funktionalismia kritisoivalle kiinalainen mieli -argumentille. Siinä kuvitellaan, että kiinan väestö kasvaisi yhtä suureksi ihmisen hermosolujen määrän kanssa. Nyt ajatellaan, että jokaiselle kiinalaiselle olisi jaettu lista puhelinnumeroita, joihin he soittavat aina saadessaan itse puhelun. Listat olisi laadittu siten, että soitettujen puhelujen kuviot (patterns) vastaisivat mentaalisessa tilassa olevien aivojen hermosolujen välisten signaalien kuvioita. Block pohti, voisiko tällainen kiinalainen mieli esimerkiksi tuntea kipua, ilman että kukaan yksittäisistä ihmisistä tuntee kipua.

Searlen vastaus on kuitenkin jälleen osoitettavissa vääräksi systeemivastauksen avulla, samalla tavoin kuin Searlen alkuperäinen kiinalainen huone -argumentti sekä vasta-argumentti robottivastaukselle osoitettiin vääräksi. Ihmisen mieli on jälleen väärällä funktionaalisen organisaation tasolla, jotta siihen voisi syntyä aito ymmärrys. Searlen vastaus ei edelleenkään kykene ottamaan kantaa siihen, syntyykö koko systeemiin ymmärrys.

Minkä takia aivosimulaattorivastaus siis on annettu? Aivosimulaattorivastaus on muodostettu sen takia, että monet tekoälyn tutkijat uskovat, että tietoisuuden syntyyn ei

kelpaa mikä tahansa laskennallinen systeemi. Se on siis robottivastauksen tavoin modifikaatioehdotus, jonka avulla kiinalainen huone voisi aidosti ymmärtää kiinankieltä, vaikka alkuperäinen versio siitä ei kykenekään tähän. Aivosimulaattorivastauksen kannattajat väittävät, ettemme voi tarkalleen tietää millainen ymmärrykseen kykenevän laskennallisen systeemin pitäisi olla, mutta koska ihmismieli ainakin varmuudella ymmärtää, niin varminta on kopioida ihmisaivojen funktionaalisuus täydellisesti.

Aivosimulaattorivastaus on reverse engineering -keino luoda tietoisuus (ks. 1.2). Koska on hyvin vaikea määritellä välttämättömät ominaisuudet, jotka tietoisuuteen vaaditaan, on parempi toteuttaa kaikki ominaisuudet, jotka ihmisaivoilla on. Evoluutio on miljoonien vuosien aikana löytänyt nämä ominaisuudet, joten on helpompaa kopioda suoraan nämä ominaisuudet, kuin keksiä nämä puhtaalta pöydältä.

Perinteinen komputationalismi ei tietenkään vaadi, että tietoisen systeemin tarvitsee olla juuri ihmisaivojen funktionaalisen organisaation kaltainen, vaikka monet konnektionistit näin kyllä väittävätkin. Kuten muistamme luvusta 4.2, yksinkertainen Turingin kone kykenee simuloimaan täydellisesti myös tällaisen systeemin, ja näin ollen moninainen realisoitavuus -periaatteen mukaisesti olemaan tietoinen. Aivosimulaattorivastaus siis periaatteessa pyrkii vain määrittelemään tietoisen ohjelman rakenteen, joka sitten voidaan suorittaa halutussa fysikaalisessa laitteistossa.

Erityisesti konnektionismi on siis puolustanut aivosimulaattorivastausta. Se on kritisoinut symbolista tekoälyä, jossa informaation prosessointi on peräkkäistä ja symbolista. Usein konnektionistien mukaan informaation prosessoinnin suorittavan fysikaalisen organisaation täytyy olla juuri konnektionistisen hermoverkon kaltainen (esim. Churchland, 1990, 35), vaikka monille konnektionisteille kyllä riittääkin, että vain ohjelman rakenteen täytyy olla konnektionistisen hermoverkon kaltainen, ja tämä ohjelma voidaan sitten suorittaa myös tavallisessa tietokoneessa. Aivot toimivat joka tapauksessa nimenomaan konnektionistisella tavalla: jokainen aivosolu toimii itsenäisenä yksikkönä, ja näin informaation prosessointi on rinnakkaista ja hajautettua. Konnektionististen ohjelmien on sanottu tuovan tietokoneisiin juuri sen luovan ja joustavan piirteen, mikä koneilta yleensä puuttuu. Tällaiset ohjelmat pärjäävätkin

käytännössä huomattavasti perinteisiä ohjelmia paremmin nimenomaan ihmisille helpoissa tehtävissä, kuten hahmontunnistuksessa.

Esimerkiksi Chalmers (1992) esittää myös, että semantiikan syntyyn vaaditaan konnektionistisen verkon kaltainen systeemi. Hän ei pidä robottivastauksessa esitettyä kausaalista linkkiä niinkään tärkeänä semantiikan syntyyn, vaan hän pitää uskottavampana sitä, että merkityksien synty tapahtuu niin sanotulla alisymbolisella tasolla toisin kuin symbolisessa tekoälyssä, jossa symbolit ovat itsessään merkityksellisiä. Konnektionismi mahdollistaa nimenomaan tällaisen alisymbolisen tason. Tästä asiasta lisää luvussa 6.

Searle kritisoi aivosimulaattorivastausta kuitenkin myös siitä, että vaikka simuloisimme aivoja täydellisesti, se olisi silti vain simulaatiota. Voimmehan simuloida esimerkiksi pyörremyrskyn toimintaa tietokoneella, mutta ei se silti ole sama asia kuin aito pyörremyrsky. Asia ei kuitenkaan ole aivan näin yksinkertainen, sillä tuntuu, että joskus X:n simulointi voi olla itse X. Voidaan kysyä, ovatko esimerkiksi tekosydämet tai tekonivelet vain aitojen simulaatiota. Simulaatio tuntuu jokseenkin vähättelevältä ilmaisulta tekoelimille, sillä ne ajavat täysin saman funktionaalisen asian kuin aidotkin. Voidaan myös löytää helposti systeemeitä X, joiden simulaatio varmasti on myös X. Esimerkiksi "systeemin, joka toimii kellona" täydellinen simulointi toimii taatusti myös kellona.

Chalmers ehdottaakin (1996, 327), että X:n simulointi on X tarkalleen silloin, kun ominaisuus X on organisatorinen invariantti, eli silloin, kun ominaisuus X riippuu pelkästään alla olevan systeemin funktionaalisesta organisaatiosta eikä mistään muusta. Tämä tarkoittaa käytännössä sitä, että X:n simulointi on X silloin, kun simuloitava ominaisuus riippuu pelkästään systeemin funktionaalisen organisaation sisältämien funktioiden syöte/tulos -arvoista funktionaalisesti merkittävälle tasolle asti. Yksinkertaistetusti, jos simulaatio toimii alkuperäisessä tehtävässään, ei se silloin ole pelkästään simulaatiota, vaan replikaatiota. Searlen ehdottama pyörremyrskynä oleminen ei ole organisatorinen invariantti, sillä se riippuu myös systeemin fyysisistä ominaisuuksista, kuten fysikaalisten kappaleiden nopeudesta ja muodosta. Pyörremyrskyn simulaatiolla ei ole näitä fysikaalisia ominaisuuksia. Sen sijaan kellona

toimiminen on organisatorinen invariantti. Kellon simulointi toimii myös kellona. Mielen fenomenaaliset ominaisuudet ovat Chalmersin mukaan myös organisatorisia invariantteja. Tietoisuus riippuu siis pelkästään realisoivan systeemin kuten aivojen funktionaalisesta organisaatioista. Ja tämähän on itseasiassa yksi funktionalismin pääteeseistä (ks. 3.5). Jos jokin systeemi kykenee toteuttamaan tietyn mentaalisen funktion, voidaan sille antaa tämän mentaalisen ominaisuuden status. Mitä Chalmers pitää sitten perusteena tälle funktionalismin teesille? Tulemme nyt ehkä tärkeimpään argumenttiin funktionalismin puolesta.

Mitä nimittäin tapahtuisi tietoisuudelle, jos ihmisen aivosolut korvattaisiin yksitellen jollain keinotekoisilla mutta funktionaalisesti vastaavilla komponenteilla, esim. aivosoluja mallintavilla elektronisilla piireillä? Ajatuksen on ilmeisesti ensimmäisenä esittänyt Clark Glymour 1970-luvun puolivälissä ja kirjallisesti sitä on käsitellyt Zenon Pylyshyn (1980). Oletetaan siis, että aivosolujen toiminta on täydellisesti mallinnettavissa formaalien sääntöjen avulla, ja kuten muistamme kappaleiden 5.1 ja 5.3 perusteella, meillä on hyviä syitä uskoa, että näin on. Kaikki fysikaaliset tapahtumat ovat periaatteessa formalisoitavissa, joten voimme hyvin perustein luoda funktionaalisesti aivosoluja vastaavat elektroniset piirit. Ja vaikka aivosoluissa olisi jopa kvanttitasolle asti funktionaalisesti merkitseviä ominaisuuksia, periaatteessa nämäkin voidaan mallintaa kvanttitietokoneilla. Mitä siis voisi tapahtua tietoisuudelle, jos kyseinen ajatuskoe aivojen asteittaisesti korvaamisesta elektronisilla piireillä toteutettaisiin?

Jos aikoo pitäytyä funktionalismin ja siis komputationalismin vastustajana, voi kysymykseen esittää vain kaksi vaihtoehtoista vastausta. 1. Tietoisuus häviää asteittain sitä mukaa, mitä enemmän aivosoluja korvataan ja lopulta tietoisuus häviää kokonaan. 2. Tietoisuus pysyy tiettyyn pisteeseen asti kirkkaana, kunnes se yhtäkkiä katoaa täysin. Jälkimmäinen vaihtoehto tuntuu äärimmäisen epäuskottavalta. Miten koko tietoisuus voisi kadota yhden ainoan hermosolun vaihdossa? Lisäksi tuon hetken löytäminen tuntuisi hankalalta. Olisiko se silloin, kun puolet aivosoluista on korvattu, vai kenties, kun neljännes on korvattu? Ensimmäinen vaihtoehto tuntuukin huomattavasti uskottavammalta. Arkielämässähän törmäämme useinkin tilanteisiin, jolloin

tietoisuutemme taso on jollain tapaa matalampi kuin yleensä. Näin on esimerkiksi silloin kun nukahdamme.

Onko vaihtoehto 1 siis uskottava vaihtoehto? On tärkeää nimittäin muistaa se, että kun aivosoluja korvataan elektronisilla vastineilla, niiden funktionaalinen toiminta pysyy muuttumattomana ja tällöin myös henkilön käyttäytyminen pysyy täysin muuttumattomana. Sen sijaan esim. nukahtaessamme myös funktionaalinen toimintamme muuttuu sitä mukaa, kun tajunnan taso laskee. Jos ajatellaan, että tietoisuutemme häviäisi vähitellen käyttäytymisemme pysyessä samana, tarkoittaisi se sitä, että esimerkiksi havaitessamme värejä, näkisimme ensin punaisen kirkkaana ja sitä mukaa, kun soluja korvataan, punaisen punaisuus jollakin ihmeellisellä tapaa vähitellen häviäisi, mutta itse asiassa emme silti edes huomaisi sitä, sillä funktionaalisen toimintamme täytyisi pysyä samana. Tällainen ajatus tuntuu jokseenkin kummalliselta.

Esimerkiksi Searle (1992) on kuitenkin yrittänyt puolustaa seuraavalla ajatuskokeella sitä, että tietoisuus voisi vähitellen kadota aivosoluja korvattaessa:

> ... kun piisiruja asteittain asetaan hupeneviin aivoihisi, tunnet kuinka tietoisen kokemuksesi laajuus pienenee, mutta tämä ei aiheuta muutosta ulkoiseen käyttäytymiseesi. Huomaat hämmästykseksesi, että olet todellakin kadottamassa kontrollisi ulkoiseen käyttäytymiseen. Huomaat esimerkiksi, että lääkäreiden testatessa näköäsi, he sanovat: "Pidämme punaista objektia edessäsi, kerro mitä näet." Haluat huutaa: "En näe mitään. Olen sokeutumassa täysin." Mutta kuuletkin sen sijaan äänesi sanovan ilman kontrolliasi: "Näen punaisen objektin edessäni."

Tarina ei ole kuitenkaan kovinkaan uskottava. Miten edellä kuvattu uskomus "En näe mitään. Olen sokeutumassa täysin." voisi edes syntyä aivoissa, jos siellä ei mikään muutu funktionaalisesti. Ei olisi mahdollista löytää mitään sellaista fysikaalista ilmiötä, joka aikaansaisi tämän uskomuksen. Tämä ei tietenkään nykyvalossa ole mahdollista. Jotta kyseinen uskomus voisi syntyä, vaadittaisiin tietoisuuden luonteesta vahvasti dualistinen teoria. Tällöin uskomusten ja fysikaalisten tilojen suhde olisi hyvin omituinen tai vaihtoehtoisesti ihmisen täytyisi olla suuresti erehtynyt omista mentaalisista tiloistaan siitä huolimatta, että käyttäytyy täysin rationaalisesti (Chalmers, 1996, 258). Ja nämä ovat hyvin epäuskottavia vaihtoehtoja.

Lisäksi Chalmers (1996, 270) ehdottaa, että voimme periaatteessa tehdä funktionaalisesti isomorfisen elektronisen laitteen aivoista ja kytkeä sen ns. varmuuspiiriksi aivoillemme. Nyt voimme liittää oikeiden aivojen ja "vara-aivojen" välille kytkimen, jolla voimme valita kumpi niistä "ajattelee" vuorollaan. Funktionalismin vastustajien mukaan aina, kun kytkin käännetään vara-aivojen puolelle, tietoisuus yhtäkkiä katoaa ja takaisin käännettäessä tietoisuus taas yhtä nopeasti palaa. Siltikään kyseisen henkilön ei tulisi huomata mitään eroa. Ajatus on omituinen. Miten voisi olla mahdollista, että elämääni ja tietoisuuteeni jäisi esimerkiksi monen vuoden mittainen aukko (jos kytkintä pidettäisiin niin kauan toisessa asennossa), ja kun kytkin jälleen käännettäisiin takaisin, tietoisuuteni palaisi, enkä silti edes huomaisi kyseistä ammottavaa aukkoa tietoisuudessani? Paljon uskottavampi vaihtoehto on, ettei tietoisuus katoa minnekään, vaan pysyy samankaltaisena koko ajan.

Kyseisille argumenteille funktionalismin puolesta ei ole tullut toistaiseksi vielä yhtäkään uskottavaa vasta-argumenttia. Argumentti siis tukee hyvin vahvasti sitä, että tietoisuuden syntyyn riittää oikeanlaiset funktionaaliset ominaisuudet, eikä esimerkiksi realisoivalla materialla ole merkitystä. Loogisesti saattaa olla mahdollista, että tietoisuus voisi kadota edellä kuvatuissa ajatuskokeissa, mutta todellisuudessa se kuulostaa äärimmäisen epäuskottavalta. Se asettaisi kaiken lisäksi hyvin kiistanalaiseksi tiedon omista mentaalisista tiloistani ylipäänsä. En tällöin voisi olla varma edes siitä, olenko juuri nyt tietoinen, ja kuten muistamme, tämä oli juuri se ainoa asia, jota Descartes piti varmana tietona.

5.4.5 Muita vasta-argumentteja

Edellä mainitut kolme vasta-argumentia ovat yleisimmin kiinalainen huone -ajatuskoetta vastaan esitetyt argumentit. Näiden lisäksi on esitetty monia variaatioita näistä. Searlen perusargumentti joka tapauksessa kumoutuu jo pelkästään loogisen muotonsa perusteella, jonka systeemivastaus osoittaa.

Usein sitä, että kiinalainen huone voisi aidosti ymmärtää, on puolustettu yksinkertaisesti sillä perusteella, että se kykenee selviytymään Turingin testistä. Emme voi kieltää siltä ymmärrystä sillä perusteella, että se tuntuu meistä mahdottomalta. Ainoa syy minkä

perusteella pidämme muita ihmisiä ymmärtävinä, on heidän käyttäytymisensä, joten jos aiomme pitää muita ihmisiä ymmärtävinä, täytyy meidän pitää myös testin läpäisevää tietokonetta ymmärtävänä.

Searlen mukaan kognitiotieteissä kuitenkin oletetaan ihmisten mentaaliset tilat ja niistä tietäminen samalla tavoin kuin luonnontieteissä oletetaan fysikaalisten objektien olemassaolo ja niistä tietäminen. Hän myöntää, että jos robotti käyttäytyisi täysin kuten ihminen ja näyttäisikin vielä samalta, meidän täytyisi pitää sitä tietoisena niin kauan, kun meillä ei olisi näyttöä toiseen suuntaan. Mutta heti jos tietäisimme, että sen käyttäytyminen perustuu laskennallisiin operaatioihin, niin emme enää voisi Searlen mukaan pitää sitä tietoisena. Kuitenkin edellisten kappaleiden perusteella voimme todeta, ettei tämä väite ole mitenkään perusteltu.

Searle väittää myös, etteivät mentaaliset tilat voi olla vain laskennallisia prosesseja ja niiden tulosteita, koska laskennallisia prosesseja ja niiden tulosteita voi olla ilman, että niillä on mentaalisia tiloja. Tämä on sinänsä kiinnostava väite, vaikka ei pidäkään paikkaansa. Onhan loogisesti täysin mahdollista, että mentaaliset tilat voisivat olla laskennallisia prosesseja, vaikka kaikilla laskennallisilla prosesseilla ei olisikaan mentaalisia tiloja. Mutta voisiko olla niin, että kaikilla laskennallisilla prosesseilla olisikin myös mentaalinen tila. Uskaliain esimerkki tästä tulee *Artificial Intelligence* termin keksijältä John McCarthyltä, jonka mukaan jopa termostaatilla on mentaalisia tiloja. Termostaatilla voi olla hänen mukaansa kolme mentaalista tilaa: nyt on liian kuuma, nyt on liian kylmä, tai nyt on juuri sopivaa. Ehkä termostaatti voidaan kuitenkin arkijärjen perusteella tuomita tiedottomaksi, vaikka arkijärki voikin pettää. Asiasta on kuitenkin paljon vakaviakin teorioita. Esimerkiksi Chalmers (1996, 298) ehdottaa, että kaikkiin laskennallisiin tiloihin eli informaatiotiloihin liittyy ns. protofenomenaalinen tila, jota voidaan pitää tietoisuuden tai kognition rakennuspalana. Tällainen tila ei tietenkään vastaa mitenkään ihmismielen monimutkaista ja rikasta kokemuksellista tilaa, mutta voisi olla yksittäinen rakennuspala, josta kokonaisuus koostuu. Tällainen panpsykistinen teoria on tietenkin hyvin arveluttava, mutta nykyään moni mielenfilosofi pitää panpsykistisiä mielenteorioita jopa sangen varteenotettavina vaihtoehtoina. En nyt paneudu tähän hypoteesiin kuitenkaan enempää.

Jotkut ovat kritisoineet kiinalainen huone -argumenttia myös siten, että vaikka analogiset tai digitaaliset tietokoneet eivät kykenisikään tuottamaan ymmärrykseen tarvittavia kausaalisia prosesseja, mitä ikinä nuo prosessit sitten ovatkaan, saattaa meillä kuitenkin tulevaisuudessa olla teknologioita, jotka kykenevät tällaisia kausaalisia prosesseja tuottamaan. Ehkä siis jonakin päivänä pystymme tällaisella tekniikalla luomaan aitoa ymmärrystä tietokoneisiin.

Tällainen puolustus komputationalismin puolesta ei voi kuitenkaan toimia. Komputationalismi määriteltiin nimenomaan siten, että mentaalisuus on komputaatiota. Jos luonnosta löydettäisiin jokin ei-laskennallinen prosessi, ja voitaisiin osoittaa, että tietoisuus syntyy tämän prosessin myötä, täytyisi komputationalismin tällöin tunnustaa virheellisyytensä. Mitään varteenotettavaa vaihtoehtoa tällaiseksi prosessiksi ei kuitenkaan ole olemassa. Ainoa tunnettu ehdotus tällaiseksi prosessiksi on luvussa 5.3 käsitellyt kvanttimekaaniset prosessit. Ja näissäkin prosesseissa ainoa ei-laskennallinen ilmiö on aaltofunktion romahdus ja senkin ei-laskennallisuus johtuu sen satunnaisuus-luonteesta. Tuntuu epäilyttävältä, että satunnaisuus olisi se tietoisuuden kaikkein olennaisin ominaisuus. Joku voisi kyllä väittää, että tämä satunnaisuus voisi olla ratkaisu esimerkiksi ihmisen vapaan tahdon ongelmaan. Tiedämme kuitenkin filosofiasta, että indeterminismi ei ratkaise tätä ongelmaa. Jos ihmisen ratkaisut ovat satunnaisia eli indeterministisiä, eivät ne silloin ole yhtään sen vapaampia, kuin jos ne olisivat deterministisiä.

Palataan nyt takaisin kysymykseen syntaksista ja semantiikasta sekä Searlen argumentin loogisesta rakenteesta. Kiinalainen huone -ajatuskoe oli osa seuraavaa päättelyä: Ohjelmat ovat syntaktisia, mielellä on semantiikka, syntaksi ei ole riittävä semantiikkaan, siksi ohjelman implementaatio ei ole riittävä luomaan mieltä. Kiinalainen huone -ajatuskokeen oli tarkoitus tukea premissiä, jonka mukaan syntaksi ei ole riittävä luomaan semantiikkaa. Olemme todenneet, ettei kiinalainen huone ajatuskoe kykene tukemaan tuota premissiä. Se ei siis osoita sitä, etteikö pelkkä syntaksi voisi tuottaa semantiikkaa. Mutta onko niin, että syntaksi todella on riittävä keino luoda semantiikkaa? Entä ovatko Searlen päättelyn muut premissit perusteltuja? Tämän tutkielman seuraava eli viimeinen luku ennen yhteenvetoa pyrkii valottamaan näitä kysymyksiä.

Katsotaan vielä ennen kuitenkin seuraavaa Chalmersin analogiaa Searlen päättelylle. Se antaa viitteitä siitä, mikä Searlen päättelyssä saattaa olla vikana. Chalmers (1996, 327) parodioi Searlen päättelyä seuraavalla tavalla:

1. Reseptit ovat syntaktisia.
2. Kakut ovat mureita.
3. Syntaksi ei ole riittävä luomaan mureutta.
4. Siksi reseptin implementointi ei ole riittävä luomaan kakkua.

Tämä on vahvasti analoginen esimerkki Searlen päättelylle. Voimme huomata, että jossain kohtaa on nyt virhe. Argumentti ei nimittäin erota reseptejä, jotka ovat syntaktisia objekteja, reseptin implementaatioista, jotka ovat fysikaalisia objekteja. Tietokoneohjelmat ovat abstrakteja laskennallisia objekteja, jotka ovat puhtaasti syntaktisia. Ilmiselvästikään pelkkä ohjelma ei kykene tietoisuuteen. Sen sijaan ohjelman implementaatiot ovat konkreettisia systeemejä, joilla on kausaalinen dynamiikka, eivätkä ne ole pelkästään syntaktisia. Ohjelman implentaatiolla on kausaalinen suhde maailmaan, ja tämä suhde on se, joka saattaa mahdollistaa tietoisuuden syntymisen myös laskennallisiin systeemeihin.

6 SEMANTIIKKA

Kirjoitetaan vielä kerran ylös Searlen edellisessä luvussa esittämä argumentti tietoisten koneitten mahdottomuudesta:

1. Ohjelmat ovat puhtaasti formaalisia (syntaktisia).
2. Ihmisen mielellä on mentaalinen sisältö (semantiikka).
3. Syntaksi yksinään ei ole riittävä synnyttämään semantiikkaa.
4. Siksi ohjelman implementaatio ei yksinään ole riittävä synnyttämään mieltä.

Semantiikka ajatellaan tässä siis hyvin tärkeäksi osaksi tietoisuutta, ja kieltämättä tämä tuntuu varsin uskottavalta. Ihmisen mielellä selvästikin on jonkinlainen sisältö, ja tätä sisältöä on usein kutsuttu semantiikaksi. Mutta mitä semantiikka tarkemmin ottaen tarkoittaa? Sillä on eri tieteenaloilla hieman eri merkityksiä. Semantiikka käsitetään kuitenkin yleensä tarkoittavan oppia merkityksistä, tai tarkemmin sanottuna oppia ihmisten kommunikaatiossa käyttämien symbolien merkityksistä. Syntaksi-semantiikka kysymyksen yhteydessä on parempi puhua kuitenkin metafyysisestä semantiikasta. Se tutkii sitä, mikä on merkityksen fundamentaalinen luonne, erityisesti, mikä on se tekijä, joka antaa ajatuksille niiden merkityksen (Block, 1998).

Mutta mitä sitten käsite merkitys tarkoittaa tarkasti ottaen? Tämä on kiistelty kysymys filosofiassa. Merkitys on usein jaettu kahteen osaan. Jonkin symbolin merkitys jaetaan nimittäin usein symbolin referenttiin tai denotaatioon, eli siihen objektiin mihin tuo symboli aktuaalisessa maailmassa viittaa, ja tämän lisäksi merkitys jaetaan jonkinlaiseen ajatukseen, jonka tuo symboli ilmaisee. Esimerkiksi sanan aurinko merkitys viittaa siihen fysikaaliseen objektiin, joka paistaa taivaalla, mutta tämän lisäksi sanalla aurinko on jonkinlainen abstrakti, kenties mielen sisäinen merkitys. Tätä jaottelua on kuitenkin usein kritisoitu. Voidaanko jaottelua ylipäänsä tehdä ja riippuvatko nämä merkityksen eri puolet toisistaan?

Nyt Searlen toisen premissin mukaan ihmisen mielellä on mentaalinen sisältö eli semantiikka. Mitä tämä sitten tarkoittaa? Missä merkityksessä semantiikka pitäisi ymmärtää tässä? Mieli käsittelee selvästikin merkityksiä, mutta käsitteleekö mieli vain niitä merkityksiä, jotka ovat täysin mielen sisäisiä ja riippuvaisia vain mielen sisäisistä

ominaisuuksista, vai ovatko nämä merkitykset riippuvaisia myös mielen ulkopuolisista tekijöistä, toisin sanoen symbolien referenteistä. Ovatko siis mentaaliset tilat riippuvaisia ulkomaailmasta?

Käsittelin luvussa 5.4.3 hieman internalismia ja eksternalismiä. Internalismin mukaan mentaalisten tilojen sisältö riippuu vain mielen sisäisistä ominaisuuksista, kun taas eksternalismin mukaan mentaaliset tilat riippuvat myös ainakin osittain mielen ulkopuolisista ominaisuuksista. Tämän yhteydessä puhutaan myös mielen sisällön laajasta ja kapeasta sisällöstä (wide/narrow content). Laaja sisältö ottaa huomioon myös mielen ulkopuoliset tekijät semantiikassa, kun taas kapea sisältö määritellään siten, että siihen vaikuttavat ainoastaan mielen sisäiset tekijät. Nyt internalismi voidaan määritellä siten, että sen mukaan psykologinen teoria ei tarvitse selityksessään laajaa sisältöä vaan kapea sisältö riittää, kun taas eksternalismin mukaan myös laaja sisältö vaaditaan.

Totesin samaisessa luvussa 5.4.3, että komputationalismi ei sinänsä ota kantaa siihen kumpi näistä teorioista on oikeassa. Se on käytännössä yhteensopiva molempien teorioiden kanssa. Jos ihmismielen käsittelemien symbolien merkitykset riippuvat ulkomaailmasta, on tämän riippuvuuden tultava jonkinlaista kausaalista linkkiä pitkin. Ja jos ihmisellä on tämä kausaalinen linkki, niin tämä linkki on luotavissa myös tietokoneelle. Periaatteessa komputationalismi kannattaa kuitenkin internalismia. Tällöin siis merkitykset syntyvät yksinään tietokoneen formaaleista symboleista, toisin sanoen puhtaasti syntaksista.

Tämä näkemys ei ehkä intuitiivisesti ole kovin uskottava. Kuitenkin internalismin puolesta on helppo löytää useita argumentteja. Totesin luvussa 2, että tärkein tietoisuuteen liittyvä piirre ovat kvaliat. Vaikka kvaliat usein kuvataan aistimusten perusteella, esimerkiksi sininen on se tuntemus, jonka koen katsoessani pilvetöntä taivasta, kvaliat eivät kuitenkaan tunnu riippuvan juurikaan ulkomaailmasta. Kykenen kokemaan kvalioita esimerkiksi unessa, jolloin aistini eivät ole lainkaan yhteydessä ulkomaailmaan. Kvalioitten lisäksi olen unessa ollessani yleensä tietoinen myös siinä olevien elementtien merkityksistä, jolloin nämä merkitykset eivät myöskään tunnu olevan riippuvaisia ulkomaailmasta. Korkeintaan ne ovat riippuvaisia jostain aikaisemmin kokemastani tai aistimastani ulkomaailman objektista, mutta kyseisellä

hetkellä suoraa kausaalista linkkiä näihin objekteihin ei ole. Jos kausaalinen linkki vaadittaisiin, kuten eksternalismi väittää, kumoaisi se tällöin myös sen hyvin vahvan periaatteen, että ihmisellä on suora ja välityn pääsy omiin mentaalisiin tiloihinsa.

Searlen kolmas premissi oli siis se, ettei pelkästä syntaksista voi syntyä semantiikkaa. Searlen näkemys on tästä huolimatta internalistinen. Hänen mukaansa merkitykset eivät tule myöskään kausaalisia linkkejä pitkin, mikä tuntuu uskottavalta edellisen kappaleen perusteella. Ja myös Searlen oma vastaus robottivastaukseen tuki tätä. Mistä semantiikka syntyy siis Searlen mielestä? Hänen mukaansa ihmisaivoissa on jokin biologinen ominaisuus, joka synnyttää merkitykset. Hän väittää kuitenkin, että laskennallisilla systeemeillä tätä ominaisuutta ei voi olla. Searle ei silti kerro meille mikä tämä ominaisuus ihmisaivoissa on. Tämä selitys tuntuu siis varsin omituiselta. Käsittelen nyt seuraavaksi joitain teorioita, jotka yrittävät selittää paremmin miten semantiikka voisi syntyä. Nämä ovat yhteensopivia myös sen kanssa, että koneille voisi syntyä semantiikka.

6.1 Fysikaalinen symbolisysteemi -hypoteesi

Newell ja Simon esittivät vuonna 1976 kuuluisan fysikaalinen symbolisysteemi -hypoteesin. Se kuuluu yksinkertaisesti näin: "Fysikaalisella symbolisysteemillä on välttämättömät ja riittävät ominaisuudet yleiseen älykkääseen toimintaan". Tämä tarkoittaa sitä, että systeemi, joka toimii älykkäästi, on välttämättä fysikaalinen symbolisysteemi, ja tällainen systeemi on myös riittävä kaikkeen älykkääseen toimintaan. Tämä tarkoittaa käytännössä myös sitä, että tietoisen systeemin täytyy olla fysikaalinen symbolisysteemi. Mikä sitten on fysikaalinen symbolisysteemi? Se on yksinkertaisesti fysikaalinen systeemi, joka manipuloi symboleita. Suurin osa nykyisin käytössä olevista tietokone/ohjelma -pareista ajatellaan tällaisiksi systeemeiksi. Ne käsittelevät symboleita jollakin tasolla.

Mutta mitä ovat symbolit? Newellille ja Simonille tärkein asia symboleissa on se, että ne viittaavat. Ne viittaavat johonkin kohteeseen, joka on yleensä ihmisen määrittelemä. Symboli ei ole siis symboli, ellei se symboloi jotain asiaa. Symbolit ovat tällöin representaatioita kyseisestä asiasta. Tämän lisäksi symbolit ovat atomisia. Symboleita

on mahdollista yhdistellä muodostaen laajempia ilmaisuja, mutta yksittäistä symbolia ei ole mahdollista jakaa osiin. Symbolisessa tekoälyssä komputaatio tapahtuu siis tällaisten atomisten symbolien manipulaationa. Tällaisella systeemillä on siis fysikaalisen symbolisysteemi hypoteesin mukaan riittävät ja välttämättömät ominaisuudet kaikkeen älykkääseen ja myös tietoiseen toimintaan.

Oleellista tälle hypoteesille on semantiikan kannalta se, että symbolit, joita käsitellään atomisina entiteetteinä, ovat yhtä aikaa sekä manipuloinnin kohteita, että ne kantavat representaation jostain ulkoisesta asiasta. Nyt tätä asiaa on helppo kritisoida siitä, että miten nämä itsenäiset atomiset symbolit voivat itsessään kantaa merkityksiä, ja miten kone voisi olla tietoinen näistä merkityksistä. Kone voi manipuloida esimerkiksi seuraavankaltaisia merkkijonoja, jotka koostuvat atomisista symboleista: "KISSA ON ELÄIN". Mutta koska nämä kolme symbolia ovat atomisia, eivätkä ne kanna mitään sisäistä rakennetta, on ongelmallista, miten symboli KISSA voisi kantaa jonkinlaisen merkityksen koneelle. Kyseinen symboli olisi mahdollista vaihtaa esimerkiksi symboliin KIVI, ilman että tietokoneen toiminta muuttuisi lainkaan, koska molemmat ovat atomisia. Tämä ei tunnu kuitenkaan enää järkevältä. Symboli KISSA tuntuu saavan merkityksensä ainoastaan siitä, että ihminen joka ohjelmoi kyseisen ohjelman, asettaa omassa mielessään symbolille kyseisen merkityksen. Tämä on siis juuri se, mitä Searle kritisoi argumentissaan. Tätä ongelmaa on kutsuttu myös nimellä symbolin pohjustus (symbol grounding) -ongelma (Harnad, 1990). Miten manipulaation kohteena oleva symboli voi siis saada merkityksensä?

Fysikaalinen symbolisysteemi -hypoteesin kannattajalla on kaksi mahdollisuutta vastata tähän syytteeseen. Ensimmäinen vaihtoehto on se, että symbolit eivät saa merkitystään siitä, miten niitä manipuloidaan, vaan nimenomaan siitä, miten ne ovat kausaalisesti sidoksissa ulkomaailmaan. Tämä on siis eksternalistinen vastaus. Symbolit saavat merkityksensä siitä, että esimerkiksi symboli KISSA on manipulaation kohteena silloin kuin on olemassa jonkinlainen kausaalinen linkki ulkomaailmassa olevaan kissa-olioon. Tässä on läsnä kuitenkin sama eksternalismin ongelma. Miten on mahdollista, että koneen käsittelemillä symboleilla voisi olla merkityksiä silloin, kun koneella ei ole kausaalista yhteyttä ulkomaailmaan. Ainakin ihmismieli on tietoinen merkityksistä myös esimerkiksi unessa.

Harnadin (1990, 344) ehdotus tähän ongelmaan on lähellä kausaalista selitystä, mutta silti internalistinen. Hänen mielestään yksi vaihtoehto on, että nämä atomiset symbolit ovat linkitettyinä aisti-ikoneihin, jotka antavat merkityksen symboleille. Aisti-ikonit ovat jonkinlaisia aistien välittämiä ja tallentamia kuvia ulkomaailman objekteista aivoihin. Harnadin mukaan aivot eivät kuitenkaan manipuloi suoraan näitä aisti-ikoneita vaan atomisia symboleita, jotka ovat linkitetty näihin ikoneihin. Tällöin suoraa linkkiä ulkomaailmaan ei tarvita, vaan riittää että linkki on ollut joskus olemassa ja tätä kautta aivoihin on tallentunut representaatio kyseisestä asiasta. Vastaavanlainen systeemi olisi helppo implementoida myös tietokoneelle.

Toinen mahdollisuus fysikaalinen symbolisysteemi -hypoteesin kannattajalle on esittää, että symbolit saavat merkityksensä suhteestaan muihin symboleihin. Tämä on siis täysin internalistinen käsitys merkitysten ja mentaalisen sisällön synnystä. Käsittelen luvussa 6.3 teoriaa, jonka alle voidaan lukea myös tällainen näkemys, mutta sitä ennen käsittelen kuitenkin teoriaa, joka usein on käsitetty hyvin vastakohtaiseksi teoriaksi verrattuna fysikaalinen symbolisysteemi -hypoteesiin. Se on nimeltään konnektionismi.

6.2 Semantiikka konnektionismissa

Konnektionismi on siis tekoälyohjelma, jossa älykkyyttä pyritään luomaan ottamalla mallia siitä, miten ihmisaivot käsittelevät informaatiota. Konnektionismi käyttää mallinaan neuroverkon kaltaisia systeemeitä. Mikä on semantiikan ja tietoisuuden kannalta kiinnostavaa konnektionistisessa semantiikassa ja mikä erottaa sen fysikaalinen symbolisysteemi -hypoteesista on se, että konnektionismissa systeemi ei manipuloi atomisia symboleita, jotka fysikaalinen symbolisysteemi -hypoteesissa vastasivat mentaalisia representaatioita, vaan konnektionismissa mentaaliset representaatiot ovat hajautuneet ympäri neuroverkon solmuja ja niiden välisiä yhteyksiä. Konnektionismissa symboleita ei siis manipuloida suoraan, vaan siinä manipulaatio tapahtuu niin sanotulla alisymbolisella tasolla. Manipulaatio tapahtuu yksittäisissä neuroverkon solmuissa, jotka yksinään eivät symboloi mitään. Mentaaliset representaatiot sijaitsevat siis korkeammalla tasolla kuin se taso, missä komputaatio tapahtuu. Fysikaalinen symbolisysteemi -hypoteesissahan mentaaliset representaatiot ja komputaatio sijaitsevat samalla tasolla. (Chalmers, 1992)

Konnektionismin usein ajatellaan erottuvan symbolisesta tekoälystä siinä, että neuroverkot ovat arkkitehtuuriltaan erilaisia kuin Turingin koneet, joihin symbolinen tekoäly luottaa. Tämä ei ole kuitenkaan merkittävä asia semantiikan kannalta, kuten ei komputationalismin kannalta ylipäänsä, mikä on todettu aikaisemmissa luvuissa. Molemmat arkkitehtuurit ovat universaaleja: kumpikin pystyy tekemään kaiken, minkä toinenkin pystyy (Franklin & Garzon, 1990). Ja usein konnektionistiset mallit luodaankin korkean tason ohjelmointikielillä, jotka sitten implementoidaan Turingin koneelle. Merkittävä asia tietoisuuden synnyn suhteen näiden mallien välillä on nimenomaan siinä, miten ne suhtautuvat semantiikkaan.

Nyt siis ero fysikaalinen symbolisysteemi -hypoteesin ja konnektionismin välillä voidaan nähdä seuraavassa esimerkissä. Jos koneella on representaatio esimerkiksi kissasta, tällöin symbolisen tekoälyn mukaisella koneella tämä representaatio on vain atominen symboli kuten "KISSA", jolla ei ole minkäänlaista sisäistä rakennetta. Sillä ei tunnu olevan keinoa saada merkitystään mitään kautta. Ainoastaan ihminen voi asettaa tuolle symbolille merkityksen. Voisi kyllä ehdottaa, että Harnadin ehdottama aisti-ikoni voisi olla antamassa tuolle symbolille merkityksen. Tällä aisti-ikonillahan saattaa olla myös rikas sisäinen rakenne. Tärkeää on kuitenkin se, että myös Harnadin aisti-ikoni on atominen, se on mahdollista korvata toisella aisti-ikonilla siten, että koneen toiminta ei muutu lainkaan. Sen sijaan konnektionismissa reprentaatioilla on aina rikas sisäinen rakenne, koska tuo representaatio on jakautunut ympäri neuroverkkoa. Tämä representaatio ei myöskään ole atominen. Jos yritämme vaihtaa representaation toiseen, toisin sanoen vaihdamme tiettyjä neuroverkon synapsien aktivaatiovoimakkuuksia, vaikutamme tällöin välttämättä myös muuhun neuroverkon toimintaan, jolloin koneen käyttäytyminen ei myöskään pysy samana. (Chalmers, 1992)

Palataan nyt takaisin Searlen väitteeseen, ettei pelkästä syntaksista voi syntyä semantiikkaa. Searle tuskin on kovinkaan vakuuttunut kummastakaan näistä ratkaisuyrityksestä selittää kuinka ne voisivat luoda semantiikan koneelle, ei fysikaalinen symbolisysteemi -hypoteesista eikä edes konnektionismista. Vaikka symboleilla olisi rikas sisäinen rakenne esimerkiksi konnektionismin tapaan, tämäkin rakenne olisi silti pelkästään syntaktista, eikä syntaksista voi edelleenkään syntyä semantiikkaa.

Chalmers (1992) vastaa Searlelle kuitenkin seuraavasti: Ensinnäkin, mitä tarkoittaa, että systeemi on syntaktinen? Se kaiketi tarkoittaa jotain sen kaltaista, että systeemi toimii sääntöjä seuraamalla. Semantiikka taas tarkoittaa jonkinlaista mentaalista sisältöä. Tällöin Searlen väite voidaan ilmaista: Systeemillä, joka koostuu vain sääntöjä noudattavasta toiminnasta, ei voi olla mentaalista sisältöä. Tämä väite on kuitenkin väärä. Esimerkiksi ihmisaivoilla varmuudella on mentaalinen sisältö. Silti ihmisaivot voidaan periaatteessa kuvata systeeminä, joka noudattaa vain tiukkoja sääntöjä, nimittäin luonnonlakeja. Tällöin Searlen väite yksinkertaisesti ei pidä paikkaansa.

Ehkä ei ole täysin selvää, voidaanko luonnonlait rinnastaa niihin sääntöihin, joita syntaktinen systeemi noudattaa. Chalmersin perusidea on kuitenkin oikea. Luultavasti myös ihmisaivojen toiminta on kuvattavissa syntaktisten sääntöjen avulla, olivat ne sitten luonnonlakeja tai korkeamman tason sääntöjä.

Chalmersin mukaan motivaatio sille, että syntaksista ei voisi syntyä semantiikkaa, tulee lingvistiikasta. On totta, että syntaktiset säännöt, joita kieli noudattaa, eivät ole riittäviä synnyttämään semantiikkaa sanoille tai lauseille. Mutta tämä johtuu siitä, että tällöin syntaksi, joka määrä sanojen ja lauseiden muotoa, sekä semantiikka, eli sanojen ja lauseiden merkitykset sijaitsevat samalla tasolla. Tällöin näillä ilmaisuilla voi olla vain ihmisen antama merkitys, mutta ei sisäistä merkitystä. Sen sijaan aivoissa syntaksi sijaitsee äärimmäisen matalalla tasolla. Atomien, molekyylien tai edes neuronien syntaktiset ominaisuudet eivät tunnu merkityksellisiltä kun puhutaan käsitteellisestä tasosta. Chalmers esittääkin nyt väitteen: Syntaksi tietyllä tasolla ei ole koskaan riittävä synnyttämään semantiikkaa samalla tasolla. Tästä ei kuitenkaan seuraa, etteikö syntaksi voisi synnyttää semantiikkaa ylemmällä tasolla.

Jos Chalmersin väitteet ovat tosia, seuraa tästä, että symboliset tekoälysysteemit ovat huomattavasti haavoittuvaisempia Searlen antamalle kritiikille, koska näissä symbolien manipulaatio ja symbolien representationaalinen sisältö sijaitsevat samalla tasolla. Sen sijaan konnektionismi on huomattavasti paremmin turvassa kritiikiltä, koska syntaktinen manipulaatio tapahtuu alisymbolisella tasolla neuroverkon solmuissa, kun taas representationaalinen sisältö sijaitsee korkeammalla tasolla, hajautettuna ympäri neuroverkkoa.

Fysikaalinen symbolisysteemi -hypoteesille jää kuitenkin keino kiertää tämä ongelma väittämällä, että yksittäiset symbolit saavat merkityksensä suhteestaan muihin systeemin symboleihin, kuten edellisessä luvussa vihjasin. Tällöin merkitys sijaitsisi yksittäisten symbolien ulkopuolella, mutta silti koko systeemin sisällä. Tämä onkin jo varsin lähellä konnektionismia. Tällöin mentaalinen representaatio olisi jakautunut ympäri systeemiä, eikä enää yksittäisessä symbolissa. Tämä ei kuitenkaan olisi ehkä enää täysin puhtaasti fysikaalinen symbolisysteemi -hypoteesin mukainen ajatus, vaan todellakin lähempänä konnektionismia. Seuraava luku käsittelee tämän kaltaista näkemystä.

Näissä näkemyksissä tuntuu olevan kuitenkin yksi tärkeä piirre, joka täytyy huomata. Vaikka nämä teoriat onnistuisivatkin kumoamaan Searlen väitteen, etteikö syntaksista voisi syntyä semantiikkaa, eivät nämä kuitenkaan kerro juurikaan sitä, miten tuo synty tarkalleen tapahtuu. Ne pohjautuvat yksinkertaisesti jonkinlaiseen oletukseen, että semantiikka syntyy ylemmälle tasolle jollakin tapaa emergentisti alemman tason syntaksista. Syntaksi-semantiikka -ongelma vastaa hyvin pitkälle yleistä mieli-ruumis -ongelmaa. Viimeisessä luvussa ennen yhteenvetoa käsittelen vielä William Rapaportin näkemystä siitä miten tämä ongelma voitaisiin ratkaista.

6.3 Rapaportin käsitys semantiikasta

Rapaport (1995) kysyy mitä tarkoittaa kielen ymmärtäminen tai ymmärtäminen ylipäänsä. Hänen mukaansa semanttinen ymmärtäminen on korrespondenssi eli vastaavuussuhde kahden määrittelyjoukon välillä. Toinen näistä ymmärretään toisen avulla. Esimerkiksi vieras kieli ymmärretään oman äidinkielen avulla. Mutta miten tällöin tuo toinen aikaisemmin ymmärretty ymmärretään? Rekursiivisesti vielä kolmannen avulla. Rekursio tarvitsee kuitenkin lähtöpisteen, jota ei ymmärretä enää minkään muun avulla. Tällöin tuo täytyy ymmärtää itsensä avulla. Mutta miten tämä on mahdollista? Rapaport vastaa: syntaktisesti.

Rapaport esittää, että on olemassa korrespondenssipareja, joissa toisella pareista on syntaktinen rooli ja toisella semanttinen rooli. Tällaisia voisivat olla esimerkiksi seuraavat: Elokuva voi olla semanttinen tulkinta käsikirjoituksesta. Talo voi olla semanttinen tulkinta rakennuspiirustuksista. Ihminen, joka lukee kirjaa, muodostaa

mentaalisen mallin kirjan tapahtumista, ja tämä mentaalinen malli voidaan käsittää semanttiseksi tulkinnaksi tekstistä. Tietokoneohjelman suoritus voi olla semanttinen tulkinta ohjelmasta, joka on syntaktinen, jne.

Tällaisia pareja on lukematon määrä. Jos on olemassa korrespondenssi kahden asian välillä, tällöin toinen näistä voidaan ymmärtää toisen avulla. Syntaktisen määrittelyjoukon ei tarvitse olla "kieli". Riittää, että se voidaan analysoida osiin (tai symboleihin), joita voidaan yhdistellä ja manipuloida sääntöjen mukaan. Syntaktinen ja semanttinen määrittelyjoukko täytyy myös ajatella pariksi. Se kumpi on syntaktinen ja kumpi semanttinen määrittelyjoukko riippuu näkökulmasta. Semanttinen osapuoli on se, joka on ennalta ymmärretty. (Rapaport, 1995)

Nyt kuuluu kysymys: miten ihminen voi ymmärtää esimerkiksi kieltä tai ylipäätään mielen ulkopuolisia asioita? Tähän vaaditaan se, että ihmisellä on jonkinlainen mielen sisäinen malli näistä asioista. Malli on aina jonkinlainen abstraktio tai yksinkertaistus mallinnettavasta asiasta. Mielessäkin täytyy olla siis yksinkertaistettu malli rikkaasta ja monimuotoisesta ulkoisesta maailmasta. Tällöin ihminen ymmärtää maailmaa sisäisen mallin avulla, se on semanttinen tulkinta siitä.

Rapaport käyttää mielen mallinaan symbolista systeemiä. Hän ei kuitenkaan pidä tätä kovin merkityksellisenä. Konnektionistinen systeemi kelpaisi aivan yhtä hyvin, sillä se on aivan yhtä laskennallinen ja syntaktinen kuin symbolinen systeemikin. Systeemin täytyy kuitenkin kyetä representoimaan ja käsittelemään intesionaalisia objekteja, toisin sanoen objekteja, jotka eivät ole vaihdettavissa toiseen intensionaalisessa kontekstissa, sekä näiden lisäksi epätäydellisiä, ei-olemassaolevia, mahdottomia jne. objekteja. Tällainen on systeemi, joka mahdollistaa toimimisen mallina ulkoisesta maailmasta.

Merkityksiä voidaan oppia monella tapaa, kuten ostensiivisesti, eli aistien kautta, esimerkiksi näkemällä objektin. Toinen tapa on lingvistinen. Perinteinen sanakirjasta oppiminen voisi tarkoittaa tätä. Kolmas tapa on oppia siitä, miten ja missä yhteydessä kyseistä termiä käytetään. Kun sama termi tulee vastaan monissa eri konteksteissa, sen merkitys hitaasti vakiintuu. Merkitykseksi muodostuu se asia, joka parhaiten tuntuu sopivan kyseiseen kontekstiin.

Merkitys on Rapaportille riippuvainen muista mallin representaatioista. Oppii systeemi merkityksen mitä kautta tahansa, se oppii sen kuitenkin kontekstinsa eli ennestään ymmärtämiensä representaatioiden avulla. Ja tuo konteksti on systeemin sisäinen. Vaikka systeemi oppii esimerkiksi käsitteen kissa merkityksen ostensiivisesti näkemällä kissan, tämäkin synnyttää vain linkin sisäisen KISSA symbolin ja representaation, joka kertoo mitä näen edessäni, välillä. Vaikka on olemassa kausaalisia linkkejä mielen ja maailman välillä, semantiikassa ne voivat toimia vain näin: ulkoisten objektien aiheuttamat mielen sisäiset representaatiot näistä voivat toimia referentteinä muille sisäisille symboleille, mutta koska ne ovat kaikki sisäisiä, mieli ja merkitykset ovat loppujen lopuksi täysin sisäisiä ja syntaktisia.

Kysymys kuuluu silti edelleen: jos uudet representaatiot opitaan jo tunnettujen representaatioden avulla rekursiivisesti, niin miten tuo ketju alun perin voi saada alkunsa. Rapaport vastaa seuraavasti: on olemassa myös toinen tapa ymmärtää jokin asia, kuin pelkästään ymmärtää se toisen avulla. Yksinkertaisesti, me vain totumme kyseiseen asiaan. On esimerkiksi sanottu, että ei ole mahdollista ymmärtää sellaista monimutkaista teoriaa kuten kvanttimekaniikka, siihen on vain mahdollista tottua. Tällöin asia ymmärretään tavallaan itsensä avulla. Jollei siis mielessä ole ennestään mallia, joka soveltuisi kuvaamaan uutta asiaa, tällöin ei synny semanttista linkkiä mihinkään ennestään tuttuun representaatioon, vaan syntyy vain identiteettikuvaus representaatioon itseensä. Myöhemmin myös tämän uuden käsitteen ympärille voi muodostua semanttista verkkoa ja tämä voi myös yhtyä muuhun verkkoon, jos sopiva kuvaus löytyy näiden välille.

Tärkeää Rapaportin teoriassa on joka tapauksessa se, että merkitys syntyy ympäröivän kontekstin perusteella. Voidaan ottaa analogia esimerkiksi algebrasta: kun ratkaistaan tuntematonta yhtälöstä, tämän tuntemattoman ratkaisu seuraa yksinkertaisesti ratkaisemalla yhtälön muu osa. Merkitys selviää siis ratkaisemalla konteksti tuntemattomalle.

Tämä teoria on holistinen teoria, merkitykset rikastuvat ja selventyvät sitä mukaa, mitä useammin kyseinen termi on yhteydessä muiden representaatioiden kanssa. Merkitys riippuu myös koko semanttisesta verkosta. Jos jokin osa verkosta muuttuu, muuttuu

myös kaikkien muiden käsitteiden merkitykset. Holismilla on kyllä paljon vastustajia. Sitä on helppo syyttää esimerkiksi kehämäisyydestä. Kehämäisyys voidaan kuitenkin kiertää kronologisella teorialla verkon syntymisestä. Uudet merkitykset pohjautuvat aina ennestään tunnettuihin merkityksiin. Ensimmäinen merkitys voidaan ymmärtää itse itsensä avulla, me vain totumme siihen. Tällöin sillä ei tavallaan ole edes mitään mielekästä eksternaalista merkitystä. Voidaan ajatella, että olisi esimerkiksi ihminen, jolla olisi vain yksi ajatus. Mutta mitä tämä ajatus merkitsisi? Tuntuu, ettei sillä voi olla mitään merkitystä, koska se ei ole yhteydessä mihinkään muihin ajatuksiin.

Holismilla on myös monia seurauksia. Tärkeimpänä se, että kaksi ihmistä eivät voi koskaan tarkoittaa täsmälleen samaa asiaa sillä mitä he sanovat. Tämä tekee täydellisestä kääntämisestä mahdottoman. Tosin tämän ei pitäisi ylipäänsä olla mikään yllätys. Esimerkiksi Russell (1918, 195) on sanonut, että jos tarkoittaisimme täsmälleen samoja asioita sillä mitä sanomme toisillemme, ei olisi tarvetta kommunikoida. Tarkan synonyymisyyden puute voi olla ehto kommunikaation mahdollisuudelle (Rapaport, 1996, 142). Vaikka täydellinen kääntäminen onkin mahdotonta, ei se estä sitä, etteivätkö osapuolet voisi ymmärtää toisiaan riittävän hyvin. Jos he ovat kasvaneet riittävän samankaltaisessa ympäristössä, heidän käsitteistönsä on luultavasti riittävän samankaltainen, jotta kommunikaatio onnistuu. Kontekstin täytyy olla joka tapauksessa tarpeeksi suuri, jotta tarkoitettu merkitys on mahdollista selvittää.

Rapaportin teoriaa kutsutaan käsitteellisen tai funktionaalisen roolin semantiikaksi. Tämä tarkoittaa näkemystä, että tietyn sanan merkitys on se funktionaalinen rooli, joka sillä on kontekstissa. Merkityksen ontologinen perusta on siis funktionaalinen ominaisuus. Tämä on semantiikan teoria, joka on täysin seurausta funktionalismin vaikutuksesta. Funktionalismissahan mentaalinen ominaisuus on funktionaalinen ominaisuus, ja nyt käsitteellisen roolin semantiikassa semanttinen ominaisuus on myös funktionaalinen ominaisuus.

Jos käsitteellisen roolin semantiikka on tosi, kumoaa se Searlen kolmannen premissin, jonka mukaan syntaksista ei voisi syntyä semantiikkaa. Syntaksista voi syntyä semantiikkaa, koska symbolin merkitys on se funktionaalinen rooli, joka sillä on syntaktisessa prosessissa. Käsitteellisen roolin semantiikka ei myöskään ole enää kovin

kiinnostunut siitä, toteuttaako semanttisen verkon symbolinen, konnektionistinen tai mikä tahansa muu systeemi. Riittää, että systeemi kykenee esittämään ja manipuloimaan ylempänä kuvatun kaltaisia entiteettejä.

Käsitteellisen roolin semantiikan avulla käy turhaksi myös edellisessä luvussa esitetty oletus siitä, että semantiikka syntyy emergentisti syntaksista. Tällöin semantiikka identifioidaan funktionaaliseen rooliin, joten ei ole mitään ylimääräistä ilmiötä, joka tarvitsee selityksen. Saattaa silti tuntua, että merkityksissä on jotain, joka kaipaa lisäselitystä. Esimerkiksi käsitteen kissa merkitys tuntuu olevan myös jotain muuta kuin sen funktionaalinen rooli. Tässä kuitenkin sekoitamme merkityksen kvalioihin. Ajatellessamme käsitteen kissa merkitystä, tämä ajattelu tuottaa kvalian, jonka synty kaipaa kyllä lisäselvitystä, mutta itse käsitteen kissa merkitys ei tarvitse lisäselvitystä.

Palaamme siis jälleen kvalian käsitteeseen. Vaikka muut mentaaliset ominaisuudet ovat ehkä selitettävissä reduktiivisesti, saattaa olla, että kvalioiden synty tarvitsee selityksekseen jonkinlaisen emergentistisen teorian.

Semantiikka on hyvin hankala ilmiö, josta keskustellaan hyvin paljon nykyäänkin. Se ei välttämättä ole kuitenkaan mieli-materia -ongelman hankalin aspekti. On löydettävissä teorioita, jotka osoittavat miten semantiikka voisi syntyä sekä ihmisiin että koneisiin. Semantiikka ei näytä olevan erottava tekijä ihmisten ja koneiden välillä, kuten ei mikään muukaan tekijä näytä olevan. Pahin ongelma on edelleen selittää kvaliat, mutta tämän ongelman selittäminen on vaikeaa riippumatta siitä, onko kyseessä ihmiset vai koneet. Yhteenvetoluvussa yritän vielä selventää tätä asiaa, sekä kokoan yhteen tärkeimmät asiat siitä, mitä olen konetietoisuudesta tämän tutkielman aikana todennut.

7 YHTEENVETO

Olen käsitellyt tässä tutkielmassa varsin laajasti tietoisen koneen ongelmaa. Olen esitellyt tärkeimmät argumentit sekä tietoisen koneen mahdollisuuden puolesta että sitä vastaan. En ole juurikaan ottanut kantaa siihen, mitkä ovat ne tarkat vaatimukset, joita koneen täytyy toteuttaa, jotta se voisi aidosti olla tietoinen. Jätän tämän asian selvittämisen kognitiotieteilijöiden ja insinöörien harteille. Olen keskittynyt siihen filosofisesti merkittävään kysymykseen, että onko konetietoisuus ylipäätään mahdollista. Ennen kuin vastaan tähän kysymykseen, palautetaan ensin mieleen tärkeimpiä asioita tutkielman sisällöstä.

Komputationalismi, siis näkemys jonka mukaan tietoisuus on komputaatiota, on ollut 1960-luvulta lähtien yksi suosituimmista näkemyksistä mieli-materia -ongelman ratkaisuksi. Se käyttää Turingin konetta mallinaan selittämään mentaaliset ominaisuudet. Turingin kone on abstrakti kone, joka kykenee suorittamaan minkä tahansa laskennallisen funktion. Komputationalismi uskoo, että prosessi, joka synnyttää mentaalisuuden ihmisaivoissa, on laskennallinen. Tästä seuraa, että systeemi, joka toteuttaa Turingin koneen, voi simuloida myös tuon mentaalisuuden synnyttävän prosessin. Ja koska komputationalismin mukaan mentaalisuus on funktionaalinen ominaisuus, niin tällöin mielen simulointi ei ole enää pelkästään simulaatiota, vaan tällöin se toteuttaa myös täysin vastaavat mentaaliset ominaisuudet, mitkä alkuperäisessä simuloinnin kohteessakin olivat.

Komputationalismia vastaan on kuitenkin kehitetty paljon vasta-argumentteja. Tärkeimpinä niistä olen käsitellyt Dreyfusin ja Searlen, sekä matematiikasta ja kvanttimekaniikasta tulleita vasta-argumentteja. Ne pyrkivät kumoamaan esimerkiksi yllämainitut komputationalismin teesit: ne voivat yrittää osoittaa, että se prosessi, joka synnyttää mentaalisuuden ihmisaivoissa ei ole laskennallinen, tai sitten ne voivat yrittää osoittaa, että mieli ei olisi funktionaalinen ominaisuus. Searle on erityisesti kritisoinut komputationalismia myös siitä, että laskennalliset systeemi eivät voi saada tietoa merkityksistä eli semantiikasta siitä syystä, että laskennalliset systeemit ovat perimmäiseltä luonteeltaan syntaktisia, eikä pelkästään syntaksista voi syntyä semantiikkaa.

Nämä argumentit osoittautuvat kuitenkin varsin hyvin perustein vääriksi. Ne saattavat ehkä kyetä horjuttamaan tiettyjä komputationalismin muotoja, mutta eivät kaikkia. Omasta mielestäni mielenkiintoisimmat vasta-argumentit tulevat kuitenkin kvanttimekaniikan piiristä. Eivät ehkä siinä mielessä, että ne pystyisivät osoittamaan vääräksi komputationalismin, vaan siinä, että modernin fysiikan teorioiden kehitys saattaa auttaa meitä selventämään mentaalisen suhdetta fysikaaliseen. Jo nyt on olemassa mielenkiintoisia teorioita siitä, miten kvanttimekaniikan mysteerit saattaisivat olla avuksi tietoisuuden mysteerien selvittämisessä.

Vasta-argumenttien lisäksi käsittelin myös ehkä tärkeintä argumenttia komputationalismin ja ylipäänsä funktionalismin puolesta. Tämä oli argumentti aivosolujen asteittaisesta korvaamisesta funktionaalisesti vastaavilla elektronisilla piireillä. Tätä on kutsuttu myös häipyvä kvalia (fading qualia) -argumentiksi. Jos tietoisuus olisi jokin ei-funktionaalinen ominaisuus, tulisi tietoisuuden jollakin tapaa kadota sitä mukaan kun aivoja korvattaisiin näillä elektronisilla piireillä. Tämä ei tuntunut kuitenkaan mahdolliselta. Jos siis tämä argumentti on tosi, seuraisi tästä, että myös funktionalismin täytyy olla tosi. Eikä tätä argumenttia vastaan ole tiedossa hyviä vasta-argumentteja.

Mutta palataan nyt kuitenkin kysymykseen kvalioista. Lupasin jo tutkielman alussa keskittyä siihen kysymykseen, miten nimenomaan tietoisuus ja sen tärkein piirre kvaliat voivat syntyä tekoälysysteemeissä. Onko tämä selventynyt tutkielman aikana? Ehkä hieman, mutta tuskin riittävästi. Onko niin, että komputationalismi ei kykenekään vastaamaan tähän kysymykseen yhtään sen paremmin kuin mikään muukaan mielen teoria? Yritän nyt hieman vielä selventää tätä asiaa.

Tein tutkielman alussa jonkinlaisen oletuksen, että fysikalismi on tosi. Fysikalismin voi nähdä koostuvan kahdesta teesistä: 1. Ontologisesta fysikalismista, eli siitä, että on olemassa vain materiaa, sekä 2. mieli-materia supervenienssi -teesistä, eli siitä, että materia determinoi mentaaliset ominaisuudet. Miten siis kvaliat suhtautuvat tähän näkökantaan? Mitä ovat kvaliat tai mentaaliset ominaisuudet? Ovatko ne redusoitavissa tai selitettävissä fysikaalisten ominaisuuksien avulla?

Lähdetään liikkeelle reduktiosta. Mitä on reduktio? Reduktio voidaan yksinkertaisesti kuvata siten, että jos jokin asia on redusoitavissa toiseen, niin ei ole pohjimmiltaan olemassa mitään muuta kuin tuo jälkimmäinen asia. Reduktiota voidaan nähdä olevan kuitenkin monen tyyppistä. Kim (2005, 275) jakaa reduktion mieli-materia -ongelman kannalta katsoen kolmeen tärkeään eri tyyppiin: silta-laki (bridge-law) -reduktioon, identiteettireduktioon, sekä funktionaaliseen reduktioon.

Silta-laki -reduktio on suunniteltu selittämään korkean tason teoria matalamman tason teorian avulla. Siinä asetetaan korkeamman ja matalamman teorian välille laki, joka yhdistää nämä toisiinsa. Esimerkiksi optiikka voidaan redusoida elektromagneettiseen teoriaan. Kysytään nyt, onko mentaalisuus silta-laki -redusoitavissa fysikaaliseen ja kykeneekö tämä reduktio myös selittämään miten mentaalisuus syntyy fysikaaliseen? Oletetaan, että reduktio voitaisiin todellakin tehdä. Oletetaan, että olisi osoitettu, että seuraava silta-laki pätee: Kaikille x, x tuntee kipua aikana t jos ja vain jos x on fysikaalisessa tilassa N aikana t. Tämä voidaan siis nähdä tietynlaisena korrelaatiolakina. Mutta selittääkö tämä miksi kipu syntyy fysikaalisesta tilasta N? Ei. Itseasiassa tämä silta-laki on nimenomaan se asia, joka kaipaa selitystä. Tämän asian totesimme jo tutkielman alussa puhuessamme korrelaatiosta. Toisekseen, silta-laki reduktio on täysin yhteensopiva kaikkien dualististen mielen teorioiden kanssa. Jos mentaalisuus on ainoastaan silta-laki redusoitavissa fysikaaliseen, seuraa tästä, että mentaalisuus on emergentti ilmiö. Toisin sanoen, on vain raaka fakta että mentaalisuus syntyy fysikaalisesta. Mentaalisuus ei ole selitettävissä silta-laki -reduktion avulla.

Entä identiteettireduktio? Onnistuuko se paremmin selittämään mentaalisen? Identiteettireduktio oli se tapa, jolla identiteettiteoria pyrki selittämään mieli-materia suhteen. Se siis identifioi nämä asiat. Mutta onko tämä edelleenkään mikään selitys? Identiteettireduktio poistaa tämän ongelman väittämällä, ettei ole mitään selitettävää. Sen mukaan on siis yksinkertaisesti vain fysikaalisia ominaisuuksia, eikä mitään muuta mikä kaipaisi selitystä. Tavallaan identiteettireduktio kyllä ratkaisisi selitys -ongelman, mutta voidaanko mentaalinen todella identifioida fysikaaliseen. Identiteoriaa tutkiessamme huomasimme, että tämä sisältää paljon ongelmia. Emmekä voi siis hyväksyä identiteettireduktiota mielen redusoimiseksi pelkästään sillä perusteella, että se onnistuisi poistamaan selitys -ongelman.

Jäljellä on funktionaalinen reduktio. Millainen se on ja onnistuisiko se selittämään mentaalisen? Kim (2005, 280) esittää funktionaalisen reduktion seuraavalla tavalla:

> 1. Ominaisuudelle F, joka on tarkoitus redusoida, annetaan funktionaalinen määritelmä, joka on seuraavaa muotoa: Omata F $=_{def}$ omata jokin ominaisuus tai mekanismi P, siten että C(P), missä C määrittelee kausaalisen tehtävän, jonka P suorittaa.
> 2. Löydä ominaisuus tai mekanismi, joka suorittaa kausaalisen tehtävän, jonka C määrittelee. Toisin sanoen, identifioi F:n "realisoija" (tai "realisoijat") tutkittavasta systeemistä tai systeemeistä.
> 3. Kehitä teoria, joka selittää kuinka F:n realisoija(t) suorittaa kausaalisen tehtävän C annetussa systeemissä tai systeemeissä.

Voitaisiinko mentaalisuus redusoida ja selittää tällä tavalla? Jos otetaan esimerkiksi käsite "kipu", voitaisiin sille antaa seuraavankaltainen määritelmä: Tuntea kipua $=_{def}$ olla sellaisessa tilassa S, että S:n on aiheuttanut kudosvaurio ja S aiheuttaa vaikerointia ja käyttäytymistä, joka pyrkii välttämään kudosvaurion aiheuttanutta tekijää (Kim, 2005, 280).

Nyt voitaisiin kuvitella, että todella löytyisi jokin tila aivoista, joka olisi aiheutunut kudosvaurioista, ja joka aiheuttaisi vaikerointia sekä kudosvaurion aiheuttaneen tekijän välttämiseen pyrkivää käyttäytymistä. Tämä on itseasiassa varsin uskottava väite. Tämän perusteella kipu todellakin olisi funktionaalisesti redusoitavissa fysikaalisiin ominaisuuksiin. Mutta onnistuuko funktionaalinen reduktio antamaan mentaalisuudelle selityksen?

Tarkastellaan lausetta 1. Siinä annetaan mentaaliselle ominaisuudelle määritelmä. Toisin kuin esimerkiksi silta-laki, määritelmä ei tarvitse kuitenkaan selitystä. Määritelmiä ei lasketa ylimääräisiksi premisseiksi todistuksissa ja päättelyissä. Jäljelle jää siis löytää se fysikaalinen tila, joka toteuttaa kyseisen funktionaalisen määritelmän, sekä luoda teoria siitä, miten tuo tila toteuttaa kausaalisen tehtävänsä. Nämä ovat kuitenkin selitettävissä kausaalisesti. Ne eivät tarvitse silta-laki -tyyppisiä ratkaisuja, jotka eivät ole selitettävissä.

Näyttäisi siis olevan niin, että funktionaalinen reduktio todellakin kykenee myös selittämään redusoimansa asian synnyn. Asiaan liittyy kuitenkin suuri mutta. Kaikki

tämä nimittäin riippuu siitä, onko mentaalisille ominaisuuksille annettavissa funktionaalinen määritelmä. Erityisesti, ovatko kvaliat määriteltävissä funktionaalisesti? Mentaalisen funktionaalinen reduktio ei ole mahdollista, jos mentaalinen ei ole funktionalisoitavissa (Kim, 2005, 290).

Kysymys siis kuuluu: onko tämä mahdollista? Kun käsittelimme tietoisuutta ja kvalioita, totesimme, että on loogisesti täysin mahdollista, että se miltä sininen väri näyttää minusta, näyttää sinusta samalta kuin se miltä minusta näyttää punainen väri. Jos tämä kvalia-inversio on mahdollinen, seuraa tästä suoraan, että kvaliat eivät ole määriteltävissä sen tehtävän mukaan, minkä ne suorittavat. Esimerkiksi kivulla saattaa kyllä olla myös funktionaalinen rooli yllä kuvatulla tavalla. Kivulla on silti myös fenomenaalinen aspekti, joka ei ole kuvattavissa funktionaalisesti. Jos nämä huomiot pitävät paikkaansa, mentaalisuus ei ole redusoitavissa funktionaalisesti.

Mitä tästä seuraa? Onko funktionalismi tuhoon tuomittu, koska kvaliat eivät ole määriteltävissä funktionaalisesti? Ei ollenkaan. Siitä, että mentaalisuus ei ole redusoitavissa tai selitettävissä funktionaalisesti, ei seuraa, etteikö funktionalismi olisi tosi. Meillä on nimittäin erittäin hyviä perusteita olettaa, että funktionalismi todellakin on tosi. Tästä kertovat useat tämän tutkielman varrella esitetyt argumentit funktionalismin puolesta, sekä vastakkaisten argumenttien osoittaminen vääriksi. Jos nämä perusteet pitävät paikkaansa, syntyy mentaalisuus todellakin pelkästään funktionaalisen organisaation perusteella. Tämä synty ei vain ole selitettävissä. Funktionalismi on siis ei-reduktiivista fysikalismia. Saattaakin olla, että loppujen lopuksi joudumme turvautumaan jonkinlaisiin fundamentaalisiin luonnonlakeihin tietoisuuden selityksessä, jos haluamme linkittää mentaalisen fysikaaliseen. On vain raaka fakta tai luonnonlaki, että tiettyjä mentaalisia ominaisuuksia syntyy tietyistä funktionaalisista tapahtumista.

LÄHTEET

Anderson, David (1989): Artificial Intelligence and Intelligent Systems: The Implications. Ellis Horwood Limited.

Block, Ned (1978): "Troubles with Functionalism". *Perception and Cognition: Issues in the Foundation of Psychology*, ed. C.W. Savage, University of Minnesota Press.

Block, Ned (1998): "Semantics, Conceptual Role". *Routledge Encyclopedia of Philosophy*, ed. E. Craig, Routledge.

Boden, Margaret (1988): *Computer Models of Mind*. Cambridge University Press.

Chalmers, David (1992): "Subsymbolic Computation and the Chinese Room". *The Symbolic and Connectionist Paradigms: Closing the Gap*, ed. J. Dinsmore, Lawrence Erlbaum.

Chalmers, David (1994): "A Computational Foundation for the Study of Cognition". *Philosophy-Neuroscience-Psychology Technical Report 94-03*, Washington University.

Chalmers, David (1995): "Minds, Machines, and Mathematics". PSYCHE, 2(9), June.

Chalmers, David (1996): *The Conscious Mind*. Oxford University Press.

Church, Alonzo (1936): "An Unsolvable Problem of Elementary Number Theory". *American Journal of Mathematics*, 58:345-363.

Churchland, Paul & Patricia (1990): "Could a Machine Think?". *Scientific American*, 262(1):32-37.

Cole, David (2004): "The Chinese Room Argument". *Stanford Encyclopedia of Philosophy*, ed. Edward Zalta.

Cotogno, Paolo (2003): "Hypercomputation and the Physical Church-Turing Thesis". *The British Journal for the Philosophy of Science*, 54:181-223.

Crick F. & Koch C. (1995): "Are We Aware of Neural Activity in Primary Visual Cortex?". *Nature*, 375:121-3.

Dennett, Daniel (1987): "Fast Thinking". *The Intentional Stance*. Cambridge: MIT Press.

Dennett, Daniel (1991): *Consciousness Explained*. Boston, MA: Little, Brown & Co.

Descartes, René (1641): *Meditationes De Prima Philosophia*.

Deutsch, David (1997): *The Fabric of Reality*. Penguin Books, London.

Dreyfus, Hubert (1965): *Alchemy and Artificial Intelligence*. The RAND Corporation.

Dreyfus, Hubert (1972): *What Computers Still Can't Do*. The MIT Press, 1992.

Franklin S. & Garzon M. (1990): "Neural Computability". *Progress in Neural Networks*, ed. O. Omidvar, 1:127-145. Norwood, NJ: Ablex.

Glymour, Clark (1988): "Artificial Intelligence Is Philosophy". *Aspects of Artificial Intelligence*, Kluwer Academic Publishers, 195-207.

Gödel, Kurt (1934): "On Undecidable Propositions of Formal Mathematical Systems". *The Undecidable*, ed. Martin Davis, 39-74.

Hameroff, Stuart (1994): "Quantum Coherence in Microtubules: A Neural Basis for Emergent Consciousness". *Journal of Consciousness Studies*, 1:98-118.

Harnad, Stevan (1989): "Minds, Machines and Searle". *Journal of Theoretical and Experimental Artificial Intelligence*, 1:5-25.

Harnad, Stevan (1990): "The Symbol Grounding Problem". *Physica D*, 42:335-346.

Harnad, Stevan (2003): "Can a Machine Be Conscious? How?". *Machine Consciousness*, Imprint Academic, 67-75.

Hauser, Larry (1993): *Searle's Chinese Box: The Chinese Room Argument and Artificial Intelligence*. Michigan State University.

Kim, Jaegwon (2005): *Philosophy of Mind*. Westview Press, 2nd Edition.

Kleene, Stephen (1936): "General Recursive Functions of Natural Numbers". *Math Annual*, 112:727-742.

Knill, Emanuel (1996): *Quantum Randomness and Nondeterminism*. LANL report LAUL-96-2186.

Koch C. & Hepp K. (2006): "Quantum Mechanics in the Brain". *Nature*, 440:611-2.

Kurzweil, Ray (2005): *The Singularity Is Near: When Humans Transcend Biology*. Viking Penguin.

Leibniz, Gottfried (1714): *Monadologia*. Gaudeamus, Helsinki 1995.

Lucas, John (1963): "Minds, Machines, and Gödel". *Philosophy*, 36:112-127.

Miller, Galanter & Pribram (1960): *Plans and the Structure of Behavior*. New York: Holt, Rinehart and Winston.

Nagel, Thomas (1974): "What Is It Like to Be a Bat?". *Philosophical Review*, 4:435-50.

Newell A. & Simon H. (1976): "Computer Science as Empirical Inquiry: Symbols and Search". *Communications of the Association for Computing Machinery*, 19(3):113-126.

Penrose, Roger (1989): *The Emperor's New Mind*. Oxford University Press.

Penrose, Roger (1994): *Shadows of the Mind*. Oxford University Press.

Putnam, Hilary (1967): "Psychological Predicates". *Art, Mind, and Religion*, ed. W. H. Capitan and D. D. Merrill, University of Pittsburgh Press.

Putnam, Hilary (1975): 'The Meaning of "Meaning"'. *Language, Mind, and Knowledge*, ed. K. Gunderson, University of Minnesota Press.

Putnam, Hilary (1988): *Representation and Reality*. MIT Press.

Pylyshyn, Zenon (1980): 'The "Causal Power" of Machines'. *Behavioral and Brain Sciences*, 3:442-44.

Pylyshyn, Zenon (1985): *Computation and Cognition*. MIT Press, Cambridge.

Rapaport, William (1988): "Syntactic Semantics". *Aspects of Artificial Intelligence*, ed. J. Fetzer, Kluwer Academic Publishers, 81-131.

Rapaport, William (1995): "Understanding Understanding: Syntactic Semantics and Computational Cognition". *AI, Connectionism and Philosophical Psychology*, ed. J. Tomberlin, Philosophical Perspectives, 9:49-88.

Rapaport, William (1996): *Understanding Understanding: Semantics, Computation, and Cognition*. Tech. Rep. 96-26, SUNY Buffalo Computer Science Department.

Russell, Bertrand (1918): "Logical Atomism?" Reprinted in *Logic and Knowledge*, ed. R. Marsh, New York: Capricorn, 1956.

Russell, Bertrand (1948): "Analogy". *From Human Knowledge: Its Scope and Limits*, George Allen and Unwin, 482-86.

Russell S. & Norvig P. (1995): *Artificial Intelligence: Modern Approach*. Prentice Hall.

Schank R. & Abelson R. (1977): *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Press.

Searle, John (1980): "Minds, Brains and Programs". *Behavioral and Brain Sciences*, 3:417-24.

Searle, John (1984): *Minds, Brains and Science*. Cambridge: Harvard University Press.

Searle, John (1990): "Is the Brain a Digital Computer?". *Proceedings and Addresses of the American Philosophical Association*, 64:21-37.

Searle, John (1992): *The Rediscovery of the Mind*. MIT Press.

Sloman Aaron (1986): "What Sorts of Machines Can Understand the Symbols They Use?". *Proceedings of the Aristotelian Society*, Supp. 60:61-80.

Sloman A. & Chrisley R. (2003): "Virtual Machines and Consciousness". *Machine Consciousness*, Imprint Academic, 133-72.

Stapp, Henry (1996): "The Hard Problem: A Quantum Approach". *Explaining Consciousness - The 'Hard Problem'*, ed. Jonathan Shear, MIT Press.

Steane, Andrew (1998): "Quantum Computation". *Reports on Progress in Physics*, 61:117-173.

Turing, Alan (1936): "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*, 2:42:230-265.

Turing, Alan (1950): "Computing Machinery and Intelligence". *Mind*, LIX:433-460.

Turing, Alan (1969): "Intelligent Machinery". *Machine Intelligence*, 5:3-23.