Kognitivistisen mielenteorian kehitys ja nykytila

Historiatieteen ja filosofian laitos
Tampereen yliopisto
Pro gradu -tutkielma
Lokakuu 2010
Renne Pesonen

Tampereen yliopisto
Historiatieteen ja filosofian laitos

Renne Pesonen
Kognitivistisen mielenteorian kehitys ja nykytila

Pro gradu -tutkielma, 134 s.
Filosofia
Lokakuu 2010

Tutkielmassa käsitellään kognitivistista mielenteoriaa, joka samaistaa mielen toiminnan tietojenkäsittelyyn. Kognitivismi on tarjonnut 1900-luvun jälkipuoliskolla ehkä vaikutusvaltaisimman tieteellisen kuvan mentaalisten ilmiöiden luonteesta, ja teoria edustaa vielä tälläkin hetkellä jonkinlaista tieteellistä vakiokäsitystä mielestä. Kognitivismi on tarjonnut tieteenfilosofisen taustateorian niin kognitiiviselle psykologialle, kielitieteelle, tekoälytutkimukselle kuin muillekin kognitiotieteiden perheeseen kuuluville tutkimusohjelmille.

Kuitenkin viime vuosisadan kääntyessä loppuaan kohti on teoriaa kohtaan suuntautunut kritiiki ja yleinen tyytymättömyys selvästi lisääntynyt. Tämän työn tavoite on selvittää minkä luonteinen teoria kognitivismi oikeastaan on, miten se on syntynyt, mitkä sen keskeisimmät teoreettiset ideat ovat, mistä sen oletetut ongelmat johtuvat ja erityisesti onko se osoittautunut elinkelvottomaksi hankkeeksi samaan tapaan kuin behaviorismi viitisen vuosikymmentä sitten.

Työssä kuvataan kognitivistisen mielenteorian kahden pääkomponentin, komputationalismin ja funktionalismin, syntyä, hehitystä ja keskeisimpiä ideoita. Rinnan kognitivismin kanssa tarkastellaan myös muita viime vuosisadalla vaikuttaneita mielenteorioita, lähinnä behaviorismia ja psykoneuraalista identiteettiteoriaa. Tarkoitus on löytää ne teoreettiset ydinväitteet, jotka ovat ominaisia nimenomaisesti kognitivismille.

Lopuksi tarkastellaan minkälaisen muodon kognitivismi, ja erityisesti siihen sisältyvä komputationaalinen teoria ajattelusta, on tarkalleen ottaen saanut erityisesti filosofiassa ja tekoälytutkimuksessa. Tämän jälkeen katsastetaan teoriaan liittyviä ongelmia ja sen tulevaisuudennäkymiä. Johtopäätöksenä työssä esitetään, että kognitivistisen mielenteorian filosofiset perusajatukset ovat tietyissä piireissä saaneet ongelmallisia muotoiluja, mistä siihen liittyvät hankaluudet kumpuavat. Nämä tulkinnat eivät kuitenkaan ole millään muotoa välttämättömiä. Loppupäätelmänä on siis, että kognitivismin filosofinen ydin vaikuttaa yhä elinvoimaiselta, mutta monet sen ympärille kertyneet teoreettiset sitoumukset vaativat vakavaa uudelleenarviointia.

Avainsanat: kognitivismi, komputationaalinen mielenteoria, funktionalismi, naturalistinen mielenteoria, psykologian filosofia, kognitiotieteen filosofia

Kiitokset

Koen sekä kohtuulliseksi että tarpeelliseksi osoittaa kiitollisuudenvelkani niille ihmisille, jotka edesauttoivat tämän tekstin valmistumista ja hyödyllisillä kommenteillaan paransivat työn laatua. Asiaan liittyvien keskustelujen ja muiden tukimuotojen kautta heitä löytyisi runsaammin, mutta päätin kuitenkin rajoittua tässä mainitsemaan vain sen pienehkön joukon, joka oli tekstin kanssa suoraan tekemisissä jossain sen luontivaiheessa.

Kiitoksia siis professori Leila Haaparannalle ja parhaillaan toista professuuria laitoksellamme hoitavalle tohtori Petri Ylikoskelle sekä opiskelijakollegoilleni ja hyville ystävilleni Mikko Pelttarille, Risto Koskensillalle, Miika Haveriselle ja Antti Virnekselle.

Sisältö

1 Johdanto

1.1 Työn rakenne ja tavoitteet

Tämä työ käsittelee kognitivistisen mielenteorian koostumusta ja kehitystä filosofisine taustaoletuksineen. Teorian ydinajatus on samaistaa ajattelu ja muut psykologiset ilmiöt tietojenkäsittelyyn, tai vähintään pitää tietokoneiden toimintaa oleellisesti samankaltaisena mielen toiminnan kanssa. Tämä on ollut viimeisen viiden vuosikymmenen ajan ehkä merkittävin tieteellinen näkemys mielen olemuksesta sekä toiminut hyvin vaikutusvaltaisena käsitteen- ja teorianmuodostusperiaatteena niin psykologiassa kuin myös muissa mielen toimintaa koskettavissa tieteissä. Filosofisesti kognitivismissa erityisen mielenkiintoista on, että perinteinen mieli–ruumis-ongelma näyttäisi pitkälti ratkeavan, jos mielen ja ruumiin ongelmallista yhteyttä pidetään samanlaisena kuin tietokoneohjelman ja -laitteiston välistä käsitteellisesti ilmeisen ongelmatonta suhdetta.

Kognitivistinen mielenteorian voi katsoa syntyneen 50–60-lukujen aikana, joskin teorian tarkkaa synnyinsijaa tai -vuotta on varsin hankala paikantaa sen moninaisen taustan vuoksi. Ensinnäkin teorian taustalla olevat filosofiset näkemykset, erityisesti niin sanottu representationaalinen mielenteoria, ulottuvat 1600 ja 1700-luvuille uuden ajan filosofian niin rationalistiseen kuin empiristiseenkin perinteeseen. Kognitivismissa onkin havaittavissa kaikuja ainakin Descartesin, Leibnizin, Humen, Hobbesin ja Kantin filosofioista. Toiseksi kognitivistinen teoria kumpuaa pitkälti matematiikan perusteita koskevan tutkimuksen kehityskaaresta, joka alkaa 1800-luvun puolivälin paikkeilta modernin matemaattisen logiikan juurilta ja päätyy 1900-luvun puoliväliin mennessä ohjelmoitavien tietokoneiden syntyyn. Näihin älyllisen historian polkuihin sisältyvien ajatusten yhteenliittymä teki mahdolliseksi ymmärtää mieltä eräänlaisena koneena hyvin uudenlaisella ja hedelmällisellä tavalla. Kognitivismin teoreettinen perusta kumpusi filosofiasta ja matematiikasta, mutta itse vallankumous lähinnä kielitieteestä ja psykologiasta.

1900-luvun alun Pohjois-Amerikassa psykologinen tutkimus oli vahvasti behaviorismin pauloissa. Behaviorismin mukaan psykologian tulee kuulua osaksi empiiristen tieteiden perhettä ja ainoana tieteellisesti kunnioitettavana metodina pidettiin havaittavan käyttäytymisen tutkimista. Kaikki psykologia ei tuona aikana suinkaan ollut behavioristista, eikä sillä esimerkiksi manner-Euroopassa ole koskaan ollut erityisen vahvaa asemaa. Filosofisessa katsannossa behavioristinen suuntaus on kuitenkin erityisen mielenkiintoinen, koska se oli ensimmäinen vakavasti otettava yritys asettaa psykologia samaan jatkumoon luonnontieteiden kanssa. Säröjä tähän ohjelmaan syntyi kuitenkin jo toisen maailmansodan aikoihin myös Pohjois-Amerikkalaisen psykologian rintamalla, ja 50-luvun aikana uudenlaisen teorianmuodostuksen tarve tuli ilmeiseksi. Psykologian lisäksi muutospainetta aiheutti myös noina aikoina syntyt chmoskylainen kielitiede sekä muotoaan vielä hakeva tekoälytutkimus. Yhdessä nämä tieteenalat, tai tarkemmin ottaen monet niihin liittyvät tutkimusohjelmat, kerääntyivät kognitivistisen teorian ympärille sytyttäen niin sanotun kognitivistisen vallankumouksen. Kognitivismi ei siis ole vain psykologinen koulukunta eikä yksistään siitä syntynyt, vaan laajempi tieteen tekemisen taustateoria monille toisiinsa niveltyville tutkimusohjelmille.

Luku 2 käsittelee kognitivistisen teorian syntyä kahdesta edellä mainitusta perspektiivistä. Aluksi tutustutaan behaviorismiin sekä syihin, jotka johtivat sen korvaamiseen kognitivistisella teorialla. Jakson tarkoitus on luoda jonkinlaista taustaa, jota vasten tarkastella kognitivismin tieteenfilosofista luonnetta. Loppuluku keskittyy esittelemään, miten laskennan teoria, tietokoneet sekä käsitys tietojenkäsittelyn ja ajattelun yhteydestä syntyivät alkujaan matematiikan perustojen tutkimuksesta.

Filosofien kosketus kognitivistiseen teoriaan välittyy pitkälti funktionalistisen teorian kautta. Funktionalismi ollut 1900-luvun lopun vaikuttavin teoria mielentiloista ainakin naturalistisesti suuntautuneen analyyttisen filosofian piirissä. Teoria polveutuu pitkälti Hilary Putnamin havainnosta, että mielen suhde ruumiiseen on hyvin samankaltainen kuin ohjelman suhde tietokoneeseen, joskin matemaatikko Alan Turing oli pohjustanut tätä näkemystä jo vuosikymmen aikaisemmin. Hän esitti varsin rajallista suosiota nauttineen hypoteesin, jonka mukaan psykologiset teoriat voitaisiin muotoilla Turing-koneina, mikä yhdistää laskettavuuden teorian psykologiaan hyvin tiiviisti. Idea havaittiin äkkiä varsin ongelmalliseksi, mutta funktionalismin perusajatus mielestä eräänlaisena abstraktina tietokoneohjelman kaltaisena systeeminä jäi elämään. Funktionalistisen teorian tieteenfilosofinen ydin on koettaa määritellä mielentilojen käsite ja perustella, miksi paras tapa ymmärtää mieltä on pitää sitä jonkinlaisena tietojenkäsittelyjärjestelmänä eikä esimerkiksi neurobiologisena ilmiönä. Siinä missä kognitiivisen psykologian vastinpari oli behaviorismi, funktionalismin suhteen samassa asemassa on niin sanottu psykoneuraalinen identiteettiteoria, jonka mukaan mielentilat ovat aivotiloja. Luku 3 käsittelee identiteettiteoriaan liittyviä ongelmia sekä esittelee funktionalismin kehitystä ja keskeisimpiä filosofisia ideoita.

Luku 4 sisältää tämän työn tieteenfilosofisten pyrkimysten kannalta oleellisimman osuuden. Luku alkaa jaksolla, jossa käydään läpi sekä kognitivistisessa filosofiassa että tekoälytutkimuksessa keskeisesti vaikuttaneita oletuksia mentaalisten representaatioiden ja prosessien luonteesta. Tätä seuraava jakso on omistettu osoittamaan, että nämä oletukset ovat hyvin ongelmallisia, sekä kartoittamaan kognitivismin heikkoja kohtia erityisesti tekoälytutkimuksen alueella. Tekoälytutkimus ei ole saavuttanut lähimainkaan niitä tavoitteita, jotka tutkimuksen alkuunpanijat asettivat 50-luvulla. Mikä oireellisinta, alan tavoitteita ja ennusteita ei juurikaan katsottu tarpeelliseksi korjata vuosikymmeniin, vaikka tutkimuksen näytöt jäivät alkuvuosien menestystä lukuun ottamatta melko vaatimattomiksi. Tämä ei suinkaan koske tekoälyn teknologisia sovelluksia vaan ainoastaan sen oletettua antia mielen tutkimukselle. Joka tapauksessa tämä on nostattanut kasvavaa kritiikkiä sekä tekoälytutkimusta että koko kognitivistista teoriaa kohtaan ja herättänyt perusteltuja syitä epäillä, että mielen ja laskennan samaistamisessa on jotain hyvin perustavanlaatuista vikaa. Luvun viimeinen kolmannes pyrkii selvittelemään tätä vyyhtiä tarkastelemalla, mistä tekoälytutkimuksen ongelmat pohjimmiltaan johtuvat, onko kognitivismin asema uskottavana tieteellisenä teoriana uhattuna ja mitä arvokasta teoriassa lopulta on riippumatta sen kohtalosta.

Viimeisen luvun ongelmanasettelu on oikeastaan syy tämän tutkielman olemassaololle. Työ lähti aikoinaan käyntiin ajatuksesta tarkastella reilun kahden viime vuosikymmenen aikana tapahtunutta kognitiotieteen kehitystä ja vertailla tänä aikana syntyneitä perintei-

selle kognitivismille vaihtoehtoisia käsitteen- ja teorianmuodostustapoja. Kognitiotieteen filosofien keskuudessa on esiintynyt kasvavaa tyytymättömyyttä tieteenalan taustateoriaa kohtaan ja tarvetta pyrkiä muotoilemaan tutkimuksen perustat ja keskeiset kysymykset uudella tavalla. Viime aikoina kognitiotieteissä onkin tapahtunut muutoksia. Lievimmillään tämä on tarkoittanut uusien mallinnustekniikoiden käyttöönottamista ja tiettyjen teoreettisten painotusten uudelleenarviointia, mutta myös koko kognitivismin loppua on povattu. Kognitiotieteen sisällä on sanottukin vallankumouksen jo kerran koittaneen. Tämä tapahtui 80-luvun aikana neurolaskentaan perustuvan mallintamisen noustessa nopeasti keskeiseksi kognitiotieteen metodologiaksi. Suuntaus kyllä tarjosi radikaalejakin uudistuksia moniin keskeisiin mielentiloja ja psykologisia prosesseja koskeviin käsityksiin, mutta lähemmin tarkasteltuna se jätti kognitivistisen teorian ytimen pitkälti ennalleen. Alunperin mielenkiintoni kohdistui pääasiassa siihen, onko kaikenlaisilla kognitiotieteen uusilla suuntauksilla joku selkeä yhteinen ydin, ja jos on, niin onko se itse asiassa vanha kognitivistinen teoria vai jotain muuta? Toiseksi mikäli kognitiotiede on muutoksen tarpeessa, onko kognitivismin hylkääminen kuitenkaan se mitä varsinaisesti halutaan? Näitä kysymyksiä varten tarvitsin perinpohjaisen analyysin kognitivismin rakenteesta ja teoreettisesta ytimestä.

Analyysi paisui ja työn sisältö lopulta typistyi käsittelemään ainoastaan kognitivistista teoriaa. Pohjimmiltaan keskeisin tutkimuskysymys kuitenkin jäi ennalleen: *onko kognitivistinen mielenteoria kriisissä, ja jos on, niin millaisessa?* Kysymyksen voisi muotoilla myös, onko kognitivistisen teorian puitteissa tehty tiede nyt samassa tilassa kuin behavioristinen psykologia viitisenkymmentä vuotta sitten? Jos teoria tulee hylätä, onko jotain sen elementtejä kuitenkin syytä säilyttää? Nämä kysymykset liittyvät ensisijaisesti tieteen tekemiseen, mutta syvemmälle porautuva filosofinen huoli tässä on, että mikäli kognitivismi menetetään, jääkö jäljelle mitään tieteellisesti uskottavaa teoriaa mielen olemuksesta. Laajemmassa filosofisessa katsannossa kognitivismin erityinen merkitys on tarjota naturalistinen teoria mielen ja materian perinteisesti ongelmalliseksi koetusta suhteesta, joten teorian hylkäämisen merkitys koskettaa myös kysymystä mielen ymmärtämisen mahdollisuudesta osana luonnonjärjestystä.

Suurin osa tästä työstä sisältää kognitivistisen teorian rakenteen ja keskeisimmän sisällön esiin kaivamista sekä sen vertailua muihin koeteltuihin ja hylättyihin mielenteorioihin. Esitystapa on erityisesti kahdessa seuraavassa luvussa luonteeltaan esittelevä ja historiallinen, mutta tavoitteeni on kuitenkin esittää lähinnä systemaattinen analyysi teorian rakenteesta ja luonteesta. Kognitivismin asiallinen ymmärtäminen lienee kuitenkin hankalaa ilman kohtuullisen hyvää käsitystä sen kehityksestä. Toisekseen teorian mielekkääksi kokemani esitystapa seurailee melko hyvin sen historiallista kehitystä, ja tarjoamalla jonkinlainen käsitys tästä kehityksestä on helpompaa lopuksi esittää, mistä kognitivismin ongelmalliset piirteet johtuvat. Tekstin tyylistä huolimatta en siis suosittele lukemaan tätä työtä ainakaan ensisijaisesti historiallisena vaan systemaattisena tutkielmana.

Tutkimuskysymykseen tarjoamani vastaus ei ole aivan yksiselitteinen. Jottei lukija kuitenkaan heti alkuun halkeaisi jännityksestä, sanottakoon, että teoria todella on vakavissa ongelmissa. Hyvin suuri osa sen keskeisistä teoreettisista ideoista on vakavan uudelleenarvioinnin tarpeessa. Kuitenkin nämä piirteet ovat enemmän tai vähemmän mielivaltaisia

hypoteeseja, jotka periytyvät teorian matemaattis-filosofisesta syntyhistoriasta. Erityisesti tekoälytutkimuksen ongelmat mitä ilmeisemmin ovat seurausta vääristä metodologisista valinnoista ja huonosta tutkimuskysymysten asettelusta, jotka ovat ikävä kyllä sementoituneet osaksi kognitivismia niin tiivisti, että niiden näkeminen erillisinä teorian ytimestä voi olla vaikeaa. Puhtaasta typeryydestä ei kuitenkaan ole kysymys, vaan huonosti tunnettua aluetta on lähdetty aikoinaan kartoittamaan tavoilla, jotka mitä ilmeisimmin ovat vaikuttaneet kaikkein hedelmällisimmiltä sen hetkisen tietämyksen ja teknologian valossa. Edellä sanotusta huolimatta kognitivistisen teorian filosofisesti merkittävimmät ydinideat lepäävät nähdäkseni ainakin toistaiseksi kohtuullisen vahvalla pohjalla.

Muutama terminologiaa koskeva kommentti lienee alkuun paikallaan. Käytän usein termejä ”intentionaalinen”, ”semanttinen” ja ”mentaalinen” pitkälti synonyymisinä. Termi ”mentaalinen” luonnehtii mitä tahansa mielen ilmiöitä mutta jatkossa erityisesti psykologisiin ilmiöihin viittaavia käsitteitä, jotka periytyvät arkisesta mieltä koskevasta kielenkäytöstämme. Tällaisia käsitteitä ovat *uskomus*, *halu* ja vastaavat, eli ”mentalistinen käsite” tarkoittaa usein samaa kuin ”propositionaalinen asenne” tai ”arkipsykologinen käsite”, joita käsitellään tarkemmin johdannon seuraavassa osassa. Psykologian käsitteet eivät välttämättä ole mentalistisia. Esimerkiksi behavioristit pyrkivät muotoilemaan psykologiset teoriat havaittavaan käyttäytymiseen, eivätkä mieleen, viittaavilla termeillä. ”Semanttinen” puolestaan tarkoittaa ilmiötä, jolla on jonkinlainen merkityssisältö, ja ”intentionaalinen” ilmiötä, joka viittaa johonkin tai koskee jotakin. Kognitivismissa mielentilojen sisältöjen yleensä ajatellaan palautuvan niiden viittauskohteisiin, joten jatkossa termit esiintyvät pääpiirteissään synonyymisinä. Aivan samaa asiaa nämä eivät kuitenkaan tarkoita. Esimerkiksi kivuissaan oleminen lienee kipua koskevassa intentionaalisessa mielentilassa olemista, mutta on hieman epäselvää, onko mielekästä ajatella, että kivulla on mitään varsinaista merkityssisältöä.

Huomautettakoon myös, että tässä työssä ”mentaalinen” ja ”psykologinen” tarkoittavat yleensä samaa, mutta niillä on hienovarainen ero. Mentaalisella tarkoitan mitä tahansa mieleen liittyvää ilmiötä, siinä missä psykologiset ilmiöt liittyvät nimenomaisesti psykologian tutkimuskohteeseen. Esimerkiksi havaittava käyttäytyminen voi kuulua psykologisen tutkimuksen alaan, mutta en kutsuisi sitä varsinaisesti mentaaliseksi ilmiöksi. Lisäksi ”psykologia” tarkoittaa tässä työssä teoriaa, jonka tarkoitus on tuottaa kuvaus mielentilojen välisestä dynamiikasta sekä käyttäytymisen etiologiasta. Esimerkiksi niin sanotut kvaliat, eli tietoisen kokemuksen laadulliset piirteet, varmaankin ovat mentaalisia ilmiöitä, mutta on epäselvää, onko niillä varsinaista kausaalista merkitystä mielen toiminnassa. Näin ollen ne eivät välttämättä ole varsinaisesti psykologisia ilmiöitä edellä tarkoitetussa mielessä.

Toinen käsiteryväs, joka vaatinee täsmennystä, sisältää ilmaukset ”materialistinen”, ”fysikalistinen” ja ”naturalistinen”. Yleensä kahdella ensiksi mainitulla ei jatkossa ole juurikaan eroa, mutta viittaan termillä ”materialismi” ensisijaisesti metafyysiseen käsitykseen, jonka mukaan maailma koostuu pohjimmiltaan materiaalisista kappaleista ja niiden vuorovaikutuksista, mitä ne sitten lienevätkään. Termi ”fysikalistinen” puolestaan esiintyy yhteyksissä, joissa fysiikkatieteellä on erityistä merkitystä. Esimerkiksi ”materialistinen mielenteoria” viittaa näkemykseen, jonka mukaan mieli on pohjimmiltaan materiaalinen

olio tai prosessi, siinä missä ”fysikalistinen mielenteoria” taas, että mentaalisia ilmiöitä koskevat lainalaisuudet ovat johdettavissa fysiikan teorioista. Nämä ovat ainakin käsitteellisesti kaksi eri asiaa, mutta materialismin ja fysiikan suhdetta tässä työssä ei liene tarpeellista tämän enempää pohtia. Käytän termiä ”naturalistinen” luonnehtimaan mielenteorioita, joiden tarkoitus on auttaa ymmärtämään mieltä luonnollisena ilmiönä ja asettamaan psykologisen tutkimuksen jonkinlaiseen jatkumoon luonnontieteiden tutkimuskohteiden kanssa. Jatkossa ”tieteellinen mielenteoria” tarkoittaa pitkälti samaa asiaa sisältäen kuitenkin sivumerkityksen, että teoria ei ole pelkästään filosofinen vaan ottaa kantaa myös esimerkiksi psykologian käsitteen- ja teorianmuodostukseen. ”Mielenteoria” taas ei varsinaisesti viittaa psykologiaan, vaan filosofisiin tai muuten käsitteellisiin teorioihin mentaalisten ilmiöiden luonteesta. Jokaiseen empiiriseen tutkimusohjelmaan liittyy jokin tällainen taustateoria tietoisesti tai tiedostamatta. Huomautettakoon vielä erityisesti, että teorian naturalistisuus tai tieteellisyys ei edellytä fysikalistista reduktionismia. Esimerkiksi kognitivismi ja filosofinen behaviorismi ovat molemmat tieteellisiä mielenteorioita, joista ensiksi mainittu on nimenomaisesti mielentilojen suhteen antifysikalistinen, mistä lisää kolmannessa luvussa. Myöskään tarkoitukseni ei ole viljellä käsitystä, että naturalismi olisi edellytys teorian tieteellisyydelle. En näe mitään syytä miksi esimerkiksi ihmis- ja yhteiskuntatieteet yleensä pyrkisivät naturalisoimaan tutkimuskohteitaan, ja sama koskee merkittävää osaa psykologiasta. Tässä työssä luonnehdinnan ”tieteellinen” on siis tarkoitus olla enemmänkin teorian luonnetta kuvaava kuin normatiivinen ilmaus.

Seuraava alaluku on ehkä epätyypillisen laaja ja yksityiskohtainen johdantoon sisällytettäväksi. Kuitenkin työssä vahvasti esillä olevien arkipsykologian, representationaalisen mielenteorian ja reduktiomallin käsitteet edellytävät yllä oleviin selvennyksiin verrattuna hieman pikällisempää esittelyä. Seuraavaksi käsiteltävät asiat ovat työn varsinaisesta sisällöstä jonkin verran irrallisia ja liittyvät lähinnä yleisempiin tieteenfilosofisiin kysymyksiin. Osio ei sisällä mittään erityisen oleellista heti seuraavan luvun ymmärtämisen kannalta, joten lukija voi halutessaan hypätä alla olevan jakson yli ja tarvittaessa palata siihen myöhemmin, mikäli esitys ilman sen sisältämiä taustatietoja vaikuttaa hämmentävältä.

1.2 Tieteenfilosofisia taustoja

Aloitetaan tarkastelemalla arkipsykologiaa. Termillä viitataan käsitekokoelmaan ja selitysmalliin, joiden avulla ihmiset normaalisti selittävät sekä ymmärtävät omaa ja toistensa käyttäytymistä. Arkipsykologiset käsitteet ovat tuttuja mielentiloja: haluja, uskomuksia, havaintoja, tunteita ja niin edelleen. Yleisesti ottaen kyseinen käsitteistö koostuu niin sanotuista *propositionaalisista asenteista*. Esimerkiksi lause ”Stalin toivoi Trotskin kuolevan” ilmaisee Stalinille kuuluneen asenteen *toivoa* propositiota tai asiaintilaa *Trotski on kuollut* kohtaan. Arkipsykologinen selittämismalli taas koostuu muun muassa havainnointia, mielentiloja ja käyttäytymistä koskevista väittämistä sekä niiden välisiä yhteyksiä koskevista päätelmistä. Tämänlaisesta selittämisestä selkein esimerkki on niin sanottu *praktinen syllogismi*, jonka yleinen muoto on tapana esittää seuraavasti:

1. *S haluaa*, että *H*. (arvoarvostelma)
2. *S uskoo*, että tekemällä teon *T* halu *H* toteutuu. (keino tai uskomusarvostelma)

3. Joten *S suorittaa* teon *Q*. (toiminnallinen seuraus)

Siis esimerkiksi: 1. Stalin toivoi, että Trotski olisi kuollut, 2. Stalin uskoi, että lähettämällä salamurhaajan Trotskin perään, Trotski tulee kuolemaan, siispä 3. Stalin lähetti murhaajan Trotskin perään. Koska praktinen syllogismi on paraatiesimerkki arkipsykologisesta selittämisestä, kutsutaan arkipsykologiaa usein myös nimellä *uskomus–halu-psykologia*.

Lienee ilmeistä, että syllogismin sovellettavuus on usein varsin rajallista, koska toiminnalle saattaa olla sisäisiä – siis psykologisia – esteitä, kuten laiskuus, ristiriitaiset halut sekä uskomukset. Lisäksi mielitekojen toteuttamiselle on yleensä ulkoisia esteitä, kuten rahan ja ajan puute. Tässä mielessä arkipsykologia ei aina tarjoa kovin tehokkaita välineitä käyttäytymisen ennustamiseen, mutta käyttäytymisen ymmärtämiseen ja selittämiseen malli on perin soveltuva, mahdollisesti jopa välttämätön. Toisaalta se, että arkipsykologisten yleistysten ennustusvoima saattaa yksittäistapauksissa olla heikonlainen, voi kertoa enemmänkin olosuhteiden ja tarkasteltavan toimijan psykologisten vaikuttimien monimutkaisuudesta, ei niinkään selitysmallin periaatteellisesta heikkoudesta. Muodostavatko arkipsykologiset selitykset varsinaista systemaattista teoriaa, jonka totuudellisuutta voidaan tieteellisesti tutkia, on herättänyt jonkin verran keskustelua, eikä vastaus tähän kysymykseen ole aivan yksiselitteinen. Jotkut katsovat arkipsykologisen selittämisen olevan tapa käsitteellistää sosiaalista ymmärrystämme, joka ei ole pohjimmiltaan teoreettista vaan perustuu kykyymme kuvitella itsemme toisten asemaan. Tätä näkemystä kutsutaan *simulaatioteoriaksi*. Toinen, *teoriateoriaksi* kutsuttu, valtavirtanäkemys pitää arkipsykologiaa suurpiirteisenä teoriana mielen tiloista ja prosesseista. Simulaatioteoria on kohtuullisen uusi oivallus, joka tässä työssä sivuutetaan. Jatkossa keskitytään jälkimmäiseen näkemykseen, joka on ollut huomattavan oleellinen kognitivismin kehityksen, tavoitteiden ja teorianmuodostuksen kannalta.[1]

Teoriateorian juuret löytyvät Wilfrid Sellarsin artikkelista ”Empiricism and the Philosophy of Mind” (1956), jossa hän esitti, että arkipsykologia voisi olla kulttuuriperäinen teoria eikä esimerkiksi introspektioon perustuvaa mielentilojen kuvailua. Taustalla tässä on huomio, että ajatukset ovat samalla tavalla intentionaalisia kuin lauseet. Molempien ominaispiirteenä on viitata johonkin tai koskea jotakin, joten kuvaukset ajatuksista ovat väitelauseiden kanssa rakenteellisesti samanlaisia. Sellars esitti, että tämä yhtäläisyys ei itse asiassa johdu siitä, että mentaaliset tilat olisivat välttämättä olemukseltaan intentionaalisia, vaan siitä, että ajatukset ovat olemukseltaan kielellisiä. Intentionaalisuus on siis ensisijaisesti kielen ominaisuus, joka mahdollistaa intentionaaliset mielentilat. Kaikki psykologiset ilmiöt, kuten havainnot, eivät välttämättä ole kielellisiä, mutta voidaksemme pitää toisiamme uskovina ja haluavina olioina, tulee meidän ymmärtää esimerkiksi halun ja sen kohteen välinen ero. Näin ollen Sellarsin mukaan käsitys intentionaalisuudesta loogisesti edeltää käsitystä intentionaalisista mielentiloista. Tämä kuitenkin on helppoa yleistää käsitykseksi, jonka mukaan kaikki mentaaliset tapahtumat ovat kielellisiä, ja edelleen teoreettiseksi malliksi, jonka mukaan älykäs toiminta saa alkunsa eräänlaisesta sisäisestä puheesta. Sellarsin mukaan arkipsykologia saattaisi ensisijaisesti perustua kielisiin käy-

[1]Ks. (Stich & Nichols, 1992) teorioiden vertailusta ja suhteesta perinteiseen kognitiotieteeseen.

täntöihimme, jotka liittyvät muiden käyttäytymistä koskevaan puheeseen. Tällöin teoria siis olisi pohjimmiltaan behavioristinen ja arkipsykologinen ymmärtäminen mahdollista vasta kun intersubjektiivinen kielenkäyttö on olemassa. Sisäinen puhe on puolestaan tämänlaisesta kielenkäytöstä periytyvää kielellistä käyttäytymistä ilman havaittavaa ulkoista puhumista. Lopulta tällainen behavioristinen kielenkäyttö voitaisiin ulottaa myös oman toiminnan kuvaamiseen. Tällöin ihmisten väitteet myös heidän omista mielentiloistaan olisivat eräänlaista teoreettista kielenkäyttöä, eikä introspektioon perustuvaa mentaalisten tilojen kuvaamista. Sellars esitti tämän käsityksen fiktiivisenä ajatuskokeena kielen kehittymisestä. Mallin tarkoitus oli tarjota vaihtoehto perinteiselle intentionaalisten termien semantiikalle, jonka mukaan uskomus–halu-käsitteet viittaavat meille introspektiossa annettuihin mutta muilta salattuihin sisäisiin ilmiöihin. (Sellars, 1956, s.309–321.)

Arkipsykologian teorialuonteen ympärillä käyty keskustelu liittyy läheisesti *eliminatiiviseen materialismiin*. Eliminativismi tässä yhteydessä tarkoittaa kantaa, jonka mukaan arkipsykologia on virheellinen mielenteoria ja sen ontologia illuusio, joten uskomus–halukäsitteistö tulisi hävittää tieteellisen psykologian kielestä. Paul Churchland, eräs eliminativismin tunnetuimmista kannattajista, katsoo arkipsykologian täyttävän litanian tyypillisiä huonon teorian tuntomerkkejä. Hänen mukaansa teoria kumpuaa jo antiikin ajoista, eikä ole sittemmin juurikaan kehittynyt.[2] Lisäksi hän huomauttaa, ettei uskomus–halukäsitteistöllä voi edes mielekkäästi puhua tietyistä mielen ilmiöistä, kuten joistain mielisairauksista, luovasta mielikuvituksesta, yksilöiden välisistä älykkyyseroista, unen puutteen vaikutuksesta mielen toimintaan ja niin edelleen. (Churchland, 1981, s.6–7) Tyypillisesti eliminativistit pyrkivät korvaamaan arkipsykologian mentalistiset käsitteet neurotieteiden käsitteistöllä, jolloin kyseessä on nimenomaan eliminativistinen materialismi. Eliminativistien argumentaatio riippuu oleellisesti arkipsykologian teorialuonteesta, koska osoittaakseen uskomus–halu-selitysten muodostavan kelvottoman teorian, on niiden osoitettava muodostavan ylipäätään jonkinlaisen teorian (Stich, 1996, s.3–4).[3]

Teoriateorian kannattaminen ei tietenkään sinänsä tarkoita eliminativismin kannattamista. Päinvastoin esimerkiksi Sellars itse piti arkipsykologiaa hyvin toimivana teoriana ja katsoi olevan täysin uskottavaa, että sen termit viittaavat todellisiin mentaalisiin ilmiöihin (Sellars, 1956, s.307). Kognitivistisissa piireissä vaikuttaneista filosofeista erityisesti Jerry Fodor on moneen otteeseen puolustanut arkipsykologian pätevyyttä ja sen keskeistä asemaa tieteellisen mielenteorian prototyyppinä. On syytä huomata, että mitään kanonisoitua arkipsykologian teoriaa ei ole olemassa, vaan keskustelu koskee sitä, onko syytä olettaa, että jokin tieteellisesti uskottava, naturalistinen ja kohtuullisissa määrin arkipsykologiaa muistuttava uskomus–halu-käsitteistöön perustuva mielenteoria olisi ylipäätään muodostettavissa (Fodor, 1987, s.10.). Fodorin mukaan arkipsykologiaa ei yleensä sovelleta tietoisesti, vaan piilevästi ja automaattisesti luontevana osana normaalia inhimillistä toimintaa. Tätä voisi verrata siihen, ettemme käytä arkista gravitaatioteoriaa, esimerkiksi toteamusta ”esineet ilman tukea tippuvat maahan”, tietoisesti normaalissa toiminnassamme, mutta tosiasiallisesti toimimme ikään kuin jatkuvasti ottaisimme huomioon tämän lain. Vastaavasti Fodorin mukaan kaikessa ihmisten välisessä toiminnassa me jatkuvasti

[2]Tästä kyllä voi olla helposti toistakin mieltä. Esimerkiksi psykodynaamista teoriaa voitaneen pitää jonkinlaisena sofistikoituneena arkipsykologiana.

[3]Churchland 1981 ja 1988 ovat varsin edustavia esimerkkejä eliminativistisesta ajattelusta.

käytännön toimin ilmituomme vakaumustamme arkipsykologian pätevyyteen (*ibid.*, s.2–3). Siinä missä Sellars esitti arkipsykologian mahdollisesti olevan pohjimmiltaan behavioristinen teoria, jonka kulttuurimme jäsenet käytännössä poikkeuksetta omaksuvat, Fodor on esittänyt kyvyn ja taipumuksen arkipsykologisointiin olevan sisäsyntyistä (Fodor, 1992). Mikäli näin on, voidaan sanoa, ettemme vain toiminnassamme jatkuvasti ilmennä arkipsykologista mielenteoriaamme, vaan melko kirjaimellisesti ruumiillistamme sen.

Keskustelu arkipsykologisten selitysten luonteesta liikkuu siis seuraavien kysymysten äärellä: Muodostavatko nämä selitykset minkäänlaista koherenttia ja systemaattista kokoelmaa, ja onko kyseessä varsinainen teoria, jonka pätevyyttä voidaan mielekkäästi arvioida? Mikäli on, niin onko se hedelmällinen ja edes jokseenkin totuudenmyötäinen kuvaus mielen toiminnasta ja käyttäytymisen etiologiasta vai ei? Näiden kysymysten mielekkyys saattaa tuntua jokseenkin arveluttavalta, ellei kiinnitä huomiota keskustelun taustalla vaikuttaviin tieteenfilosofisiin motiiveihin. Kysymys on loppujen lopuksi siitä, mitä tieteellisen mielenteorian oikeastaan tulisi selittää. Selittääksemme mitä mielentilat ja mentaaliset prosessit oikeastaan ovat, tulee meillä olla jonkinlainen käsitys siitä, minkälaisia mentaalisia ilmiöitä ylipäätään on olemassa. Muussa tapauksessa meillä kai ei ole mitään käsitystä, mitä olemme selittämässä. Tietenkään mielentilojen inventaarion ei tutkimuksen alkaessa tarvitse olla täydellinen, mutta mitä eliminativistit koittavat sanoa on, että mielenteoreetikon ei kannata tuhlata aikaansa arkipsykologian parissa, kuten kognitivismiin läheisesti liittyvässä funktionalismissa on pitkälti tehty.

Arkipsykologian pätevyyden arviointi edellyttää jonkinlaista systemaattista analyysiä siitä, mitä mielentilojen luonnetta koskevia oletuksia uskomus–halu-käsitteisiin oikeastaan sisältyy. Esimerkiksi lause ”Kari haluaa jäätelöä” kertoo jotain melko itsestään selvää siitä, miten Kari käyttäytyy kun jäätelöä on saatavilla, mutta minkälaista käsitystä mielentiloista tällaiset väitteet edustavat? Mitä Karin mielessä oikeastaan on? Tuskin sentään jäätelöä. Onko haluaminen jonkinlaista sisäistä puhetta? Arkipsykologisia väitteitä on usein tapana tulkita yleisemmän teorian näkökulmasta, joka kulkee nimellä *representationaalinen mielenteoria*. Kyse ei ole juurikaan sisällöllisestä teoriasta, vaan mielenteorian muodosta, joka voidaan tiivistää kahteen teesiin seuraavasti (Fodor, 1987, s.16–21):

1. *Propositionaalisten asenteiden luonne*: Organismilla on asenne A propositioon p, jos ja vain jos sillä on tälle asenteelle ominainen kausaalinen suhde A_M mentaaliseen representaatioon p_M, joka tarkoittaa propositiota p. Mentaalinen tila siis koostuu tietynlaisesta kausaalisesta relaatiosta organismin ja tiettyä propositiota vastaavan mentaalisen representaation välillä.

2. *Mentaalisten prosessien luonne*: Mentaaliset prosessit ovat mentaalisten representaatioiden esiintymien kausaalisia ketjuja.

Kohdan 1. mukaan esimerkiksi lause ”henkilö x uskoo, että alkaa sataa” tarkoittaa, että henkilöllä x on mielessään mentaalinen representaatio p, joka viittaa asiaintilaan, että alkaa sataa, ja että x käyttäytyy p:n suhteen joillain uskomiselle ominaisilla tavoilla. Miten representaatio viittaa kohteeseensa, on asia erikseen. Voidaan ajatella, että representaatiot ovat esimerkiksi jonkinlaisia mielessä olevia kuvia, jotka muistuttavat kohteitaan, mut-

ta kognitivistisen mielenteorian puitteissa ne yleensä mielletään jonkinlaisiksi lauseiden kaltaisiksi symbolirakenteiksi.[4] Oleellista kuitenkin on, että representaation ja sen kohteen välinen suhde on sama asia, kuin mielensisältöjen ja maailman välinen suhde, joten selonteko mentaalisten representaatioiden semantiikasta on samalla selonteko mielentilojen intentionaalisuudesta. Representaation kausaalinen rooli käyttäytymisen etiologiassa puolestaan määrittelee henkilön x asenteen sitä kohtaan. Esimerkiksi uskomus, että alkaa sataa lienee vaikutuksiltaan erilainen kuin pelko, että alkaa sataa, joten uskominen ja pelkääminen vastaavat organismin erilaisia kausaalisia suhteita mahdollisesti samaan mentaaliseen representaatioon. Kohta 2. taas sanoo, että ajattelu tai muu mentaalinen toiminta on kausaalinen prosessi, jossa mielentilat saavat aikaan muita mielentiloja. Esimerkiksi pelko, että alkaa sataa voi aiheuttaa halun perua iltapäivälle suunniteltu puistokaljoittelu ja halun ottaa kauppareissulle sateenvarjo mukaan ja niin edelleen. Empiirisen psykologian ja kognitiotieteen tehtävä on sitten antaa representationaaliselle mielenteorialle varsinaista sisältöä, eli selventää minkälaisia kausaalisia suhteita havainnoinnin, propositionaalisten asenteiden ja käyttäytymisen välillä tosiasiassa vallitsee.

Arkipsykologiaa myötäilevä käsitys mielestä on historian saatossa kokenut sekä ylä- että alamäkiä niin mielenfilosofiassa kuin tieteellisessä psykologiassakin. Sillä on kuitenkin ollut keskeinen asema erityisesti 1900-luvun loppupuoliskoa hallinneessa kognitiivisessa mielenteoriassa. Periaatteeltaan tämä saattaa kuulostaa epäilyttävältä, jos kyse on kerran antiikkisesta arkijärkisestä käsitekokoelmasta, mutta pienellä tarkastelulla saattaa päätyä johtopäätökseen, että mielekkäät vaihtoehdot ovat melko vähissä. Asiaa lienee parasta valaista esimerkin avulla:[5] Jalankulkija kävelee jalkakäytävällä ja lähtee yllättäen ylittämään tietä. Samaan aikaan auto lähestyy jalankulkijaa nopeasti. Kuski suorittaa äkkijarrutuksen ja menettää autonsa hallinnan. Auto ohjautuu ulos tieltä ja osuu lyhtypylvääseen. Jalankulkija säikähtää, epäröi hetken, juoksee auton vierelle ja katsoo kuskin puoleisesta ikkunasta sisään. Välittömästi tämän jälkeen hän kaivaa puhelimensa esiin ja näppäilee numerot 1 ja 1 ... Mitä tilanteessa oikein tapahtuu? Miksi jalankulkija otti puhelimensa esiin, ja minkä numeron hän näppäilee seuraavaksi, jos minkään?

Fysiikan avulla jalankulkijan käyttäytymistä ei edellä olevien tietojen avulla voida selittää eikä ennustaa, koska tilanteen kuvauksessa ei anneta jalankulkijasta mitään yksityiskohtaista fysikaalista tietoa, jonka avulla hänen – tai jos suhtaudumme jalankulkijaan fysikaalisena kappaleena, niin ennemminkin *sen* – liikeratojaan voisi laskea. Fysiikan perspektiivistä tarvitsisimme jonkinlaista tietoa esimerkiksi niistä voimista, jotka liikuttavat hänen kättään. Vastaavasti neurofysiologian kannalta on mahdotonta esimerkiksi ennustaa, minkä numeron hän näppäilee seuraavaksi. Mikäli jalankulkijan toiminnan selittäminen ja ennustaminen edellyttäisi neurologisiin mekanismeihin viittaamista, olisi tietomme toiminnan vaikuttimista ja syistä tässä tapauksessa täysin olematonta. Jalankulkijan aivojen toiminnasta voimme tietysti päätellä yhtä ja toista. Esimerkiksi, että hänen motorinen aivokuorensa aktivoitui tavalla, joka sai aikaan kuvatunlaisia kädenliikkeitä ja niin

[4] Ks. esim. (Pylyshyn, 1984, s.193–196). Kuvan- ja lauseenkaltaisten mentaalisten representaatioiden eroista ja asemasta kognitivismissa tarkemmin ks. mainitun teoksen luku 8: "Mental Imagery and Functional Architecture".

[5] Asian yksityiskohtaisempaa käsittelyä löytyy esimerkiksi Zenon Pylyshynin teoksesta *Computation and Cognition* (1984), s.1–12, josta esimerkki ja sen johtopäätökset ovat lainattu pienin muutoksin.

edelleen. Mutta yllä olevassa ei ole tarpeeksi tietoa siitä, miten jalankulkijan hermosto aktivoitui ikkunasta katsomisen jälkeen, jotta voisimme neurofysiologian avulla ennustaa, mitä seuraavaksi tapahtuu.

Tapahtumaketju on kuitenkin helposti selitettävissä, kun ajattelemme kuskin *säikähtäneen* ja *yrittäneen* väistää jalankulkijaa, joka *päätti tarkistaa*, onko kuski loukkaantunut, *päätteli havainnostaan*, että on, *halusi* auttaa soittamalla hätänumeroon ja numeroiden 1 ja 1 jälkeen hän näppäilee numeron 2, koska hän *tietää*, että hätänumero on 112. Tarvitsemme siis intentionaalisia käsitteitä tapahtuman kuvaamiseen ja ymmärtämiseen, ja voimme ennustaa jalankulkijan käyttäytymistä käytännössä parhaiten niiden avulla. Tämä ei tietenkään tarkoita, etteikö tapahtuma olisi monimutkainen fysikaalinen prosessi ja täysin fysiikan avulla ennustettavissa, mikäli siitä olisi tarpeeksi tietoa käytettävissä. Emme kuitenkaan tarvitse tällaista tietoa selittämään tapahtumaa ja ennustamaan jalankulkijan tai kuskin toimintaa, vaan intentionaalinen kuvaus riittää. Toisekseen intentionaalinen kuvaus lienee myös välttämätön, jotta voisimme oikeastaan ymmärtää, mitä tilanteessa oikeastaan tapahtuu. Voisimme ehkä rakentaa neurologisen oheistarinan, jonka mukaan jalankulkijan motorinen aivokuori aktivoitui ohjaten kättä näppäilemään numerosarjan 112, koska hänen aivonsa olivat tilassa, joka vastaa halua auttaa kuskia, ja hänen aivoihinsa oli tavalla tai toisella koodattuna tieto hätänumerosta. Joka tapauksessa pystyäksemme kuvaamaan jalankulkijan käyttäytymistä mielekkäänä toimintana, tarvitsemme intentionaalisia termejä ”tietää”, ”haluta” ja niin edelleen. Siis kaikesta huolimatta tarvitsemme kuvausta, joka vastaa asiaintilaa *jalankulkija tietää, että hätänumero on 112*, olipa väite annettu suoraan tällaisena jalankulkijaa koskevana intentionaalisena kuvauksena tai sitten neurofysiologisin tai fysikalistisin termein.

Yllä olevan kaltaiset tarkastelut ovat vaikuttaneet kognitivistiseen mielenfilosofiaan luomalla uskoa, että psykologiset yleistykset tulee muotoilla intentionaalisin termein, ja mielen olemuksen sekä toiminnan selittäminen on intentionaalisten mielentilojen ja intentionaalisesti kuvatun käyttäytymisen selittämistä. Yleensä tämä tarkoittaa sitoutumista suurin piirtein arkipsykologisen mallin mukaiseen käsitykseen mielestä. Kyse tässä siis on tutkimuskohteen määrittelystä sekä siitä, mikä katsotaan mielen selitykseksi ja mitä ei. Tämän teoreettisen näkökannan seurauksia kognitivismin kehitykseen tarkastellaan laajemmin kolmannessa luvussa. Toinen kysymys sitten on, voidaanko mieli ja sen toiminta täydellisesti kuvata mainitulla tavalla, kuten representationaalisen mielenteorian puitteissa oletetaan, vai poimivatko arkipsykologiset kuvaukset vain jonkin palasen mielenilmiöiden kokonaisuudesta. Eli vaikka intentionaalisten tilojen selittäminen olisi välttämätön osa mielenteoriaa, onko se kuitenkaan kattava perspektiivi mielen ymmärtämiseksi? Tätä kysymystä puolestaan kommentoidaan työn viimeisessä luvussa.

Siirrytään sitten tarkastelemaan erityisesti funktionalismia käsittelevän luvun taustalla olevaa reduktiomallia. Kun kysytään, mitä mieli on ja miten se toimii, mielekäs vastaus voi näyttää monenlaiselta. Esimerkiksi representationalistinen mielenteoria vastaa tähän omalla tavallaan antamalla määritelmän mielentiloille ja mentaaliselle kausaatiolle. Tämä määritelmä on intentionaalinen, koska mielen sisältöjen sanotaan olevan representaatioita, ja representaation määrittelevä ominaisuus on sen viittaussuhde representoituun. Toisekseen mentaalisesta kausaatiosta sanotaan, että kyseessä on mentaalisten representaa-

tioiden esiintymien ketju, joka määräytyy sen perusteella, minkälainen kausaalinen suhde organismilla on mihinkin representaatioon. Naturalistisesti suuntautuneen teoreetikon kysymys mielen olemuksesta kuitenkin koskee sitä, mikä nämä ilmiöt oikeastaan saa aikaan ja miten ne kuuluvat osaksi luonnonjärjestystä. Kysymys tällöin on, miten intentionaalisuus voidaan määritellä ei-intentionaalisesti, ja jos mentaaliset oliot ovat fysiikan lain alaisia olioita, niin miten mielensisältöjen määräämä kausaatio on ylipäätään mahdollista. Katalogi mentaalisista ilmiöitä ei siis ole vastaus mielenteoreetikon kysymykseen siitä, mitä mieli on. Jonkinlaisen mentaalisten ilmiöiden luettelon hän tarvitsee tietääkseen, mitä hän on selittämässä, mutta tämä on tutkimuksen lähtökohta, ja päämäärä on selittää mentaaliset ilmiöt ei-mentaalisesti. Tällaista selittämistä kutsutaan ilmiön *redusoimiseksi* tai *palauttamiseksi* toiseen ilmiöluokkaan. Esimerkiksi neurobiologisen reduktionismin mukaan mielentilat olevat aivotiloja ja mentaalinen kausaatio palautuu aivojen toimintaan. Jos tämä pitää paikkansa, niin tiedämme miten mieli toimii, jos tiedämme miten aivot toimivat.

Reduktion käsite ei ole aivan yksiselitteinen, ja eri tieteenfilosofit tarkoittavat reduktiolla hieman eri asioita. Kun tässä työssä puhutaan mielentilojen palauttamisesta käyttäytymiseen, laskentaan tai aivotiloihin, taustalla on eräs versio niin sanotusta *klassisesta reduktiomallista*. Sen perusidea on, että redusoitava teoria määritellään jonkun toisen teorian kielellä, joka tyypillisesti kattaa redusoitavaa teoriaa laajemman ilmiöluokan. Tämänlaisesta reduktiosta paraatiesimerkki on lämpöopin johtaminen statistisesta mekaniikasta, eli fysiikan haarasta, joka yhdistelee klassista mekaniikkaa ja tilastollisia menetelmiä. Klassinen esimerkki tästä on ideaalikaasulain johtaminen kaasun kineettisestä teoriasta. Ideaalikaasulain mukaan $paine \times tilavuus = kaasun\ ainemäärä \times kaasuvakio \times lämpötila$. Reduktio alkaa määrittelemällä yhtälössä esiintyvät käsitteet klassisen mekaniikan termeillä: ”kaasu” ymmärretään kokoelmana ainehiukkasia, joista tehdään sellainen idealisoitu oletus, etteivät ne vaikuta toisiinsa muuten kuin suorien törmäysten kautta. ”Kaasun ainemäärä” tarkoittaa kaasussa olevien hiukkasten lukumäärää. ”Lämpötila” on hiukkasten keskimääräinen liike-energia ja ”paine” puolestaan ilmiö, joka aiheutuu hiukkasten törmäyksistä kaasun sisältävän säiliön seinämiin. ”Kaasuvakio” on empiirinen perussuure. Nyt kun termit ovat määritelty mekaniikan kielellä viittaamalla hiukkasiin, liike-energiaan ja niin edelleen, varsinainen yhtälö voidaan johtaa klassisen mekaniikan laeista. Mikäli kaikki lämmön käyttäytymistä koskevat lainanalisuudet pystytään samaan tapaan johtamaan klassisesta mekaniikasta, voidaan sanoa, että lämpöilmiöt on palautettu mekaniikkaan, ja kysymys lämmön olemuksesta – siis onko lämpö jonkinlaista ainetta, energian muoto vai mitä – on ratkaistu.

Reduktion seurauksena ei siis tule selitetyksi vain redusoitavan teorian termit vaan myös sen lainalaisuudet. Ideaalikaasun tapauksessa esimerkiksi lämpötilan ja paineenkasvun yhteys tulee ymmärrettäväksi: paine kasvaa lämpötilan kasvaessa, koska lämpötilan nouseminen tarkoittaa kaasuhiukkasten kineettisen energian lisääntymistä, joka puolestaan tarkoittaa hiukkasten nopeuksien lisääntymistä, jolloin hiukkaset törmäävät säiliön seinämiin suuremmalla voimalla ja tiheämpään tahtiin.[6]

[6]Ks. (Nagel, 1961) luku 11: ”The Reduction of Theories”, erityisesti sivut 352–354, missä tämä reduktiomalli esitettiin ensimmäisen kerran.

Usein redusoitavaa teoriaa kutsutaan *ylemmän tason* teoriaksi, koska reduktion yleensä ajatellaan tapahtuvan jostain erityisemmästä, ylemmän abstraktiotason teoriasta *alemman tason* teoriaan, joka taas koskee yleisempää ja perustavanlaatuisempaa ilmiöluokkaa. Reduktiossa ylemmän tason teorian käsitteet muotoillaan alemman, redusoivan teorian kielellä niin sanottujen *siltalakien* avulla. Nämä ovat lainomaisia yleistyksiä, jotka kiteyttävät reduktion teoreettisen idean ja osoittavat, miten ja miksi teoria on palautettavissa toiseen. Esimerkiksi ideaalikaasulain tapauksessa siltalait kertovat, että lämpötila on sama asia kuin hiukkasten keskimääräinen liike-energia sekä miten laki johdetaan kineettisestä kaasuteoriasta ja niin edelleen. Siltalakien lainomaisesta luonteesta seuraa, että redusoitavan teorian ilmiöt osoittautuvat sulkeutuvan alemman tason teorian käsittelemien ilmiöiden joukkoon. "Perustavammanlaatuisempi" tarkoittaa, että esimerkiksi kaikki termodynaamiset ilmiöt voidaan periaatteessa kuvata klassisen mekaniikan tilastollisina ilmiöinä, mutta ei toisin päin, joten termodynamiikan perusta on mekaanisissa ilmiöissä. (McCauley, 1996, s.432–433)

Huomautettakoon, että reduktion tavoite ei millään muotoa ole osoittaa ylemmän tason ilmiöitä näennäisiksi, vaan oikeastaan päinvastoin. Siltalait osoittavat, että korkeamman tason ilmiöt seuraavat lainomaisesti alemman tason teorian perustavanlaatuisemmista ilmiöistä. Tässä mielessä onnistuneen reduktion nimen omaan pitäisi vahvistaa uskoa redusoitavan teorian pätevyyteen ja lisätä sitä koskevaa ymmärrystä, ainakin mikäli redusoivaa teoriaa pidetään ymmärrettävänä ja jokseenkin ongelmattomana. Jos hyvin käy, reduktion lopputuloksena kyetään ennustamaan uusia redusoitavan teorian ilmiöitä ja selittämään pois siihen liittyviä mahdollisia ongelmia.

Jerry Fodorin kehittelemä klassisen reduktiomallin laajennus tiivistää hyvin kognitivismiin, ja erityisesti kolmannessa luvussa käsiteltävään funktionalismiin, liittyvän käsityksen reduktiivisen selittämisen edellytyksistä. Malli voidaan tiivistää kolmeen ehtoon: Teoria T_Y redusoituu teoriaan T_A, jos ja vain jos 1° jokainen redusoitavan teorian laki on johdettavissa redusoivasta teoriasta, 2° jokainen redusoitavan teorian ilmiö kuuluu nomologisesti välttämättä redusoivan teorian kattamaan ilmiöluokkaan ja 3° siltalait yhdistävät jokaisen redusoitavan teorian predikaatin johonkin redusoivan teorian luonnolliseen luokkaan. (Fodor, 1974, s.98–103) Ehtoa 1° on edellä jo käsitelty, mutta ehdot 2° ja erityisesti 3° vaatinevat selventämistä.

Ehto 1° yksinkertaisesti siis edellyttää, että redusoitavan teorian lait ovat johdettavissa redusoivasta teoriasta. Reduktio on kuitenkin kaksisuuntainen tie. Sen lisäksi, että redusoivan teorian ilmiöstä A tulee siltalakien nojalla olla johdettavissa redusoitavan teorian ilmiö Y, ehdon 2° mukaan on myös välttämätöntä, että jokainen ilmiön Y esiintymä todella on redusoivan teorian ilmiö. Ei esimerkiksi olisi mielekästä sanoa valon oleva sähkömagneettista säteilyä ja optisten ilmiöiden palautuvan sähkömagneettisen säteilyn ominaisuuksiin, jos valo olisi sähkömagneettista säteilyä *tai* jotain muuta. Ehto 2° siis tarkoittaa, että redusoitava ylemmän tason teoria tulee olla johdettavissa vain redusoivasta alemman tason teoriasta, jolloin redusoivan teorian tulee olla riittävän laaja kattaakseen kaikki redusoitavan ilmiöluokan tapaukset. Ehdossa mainittu "välttämättömyys" tulee ymmärtää nomologisena, eli välttämättömyytenä luonnonlain mielessä. Esimerkiksi ideaalikaasulaki ja monet muut lämpöön liittyvät ilmiöt voidaan kyllä johtaa myös kalorik-

kiteoriasta (Brush, 1976, s.555), mutta koska kalorikkia ei ole olemassa, ei ole mielekästä väittää, etteivät kaasun lämpöilmiöt redusoidu kineettiseen kaasuteoriaan koska loogisessa mielessä lämpö voi olla molekyylien liikettä tai kalorikkia tai jotain aivan muuta. Itse asiassa mikä tahansa teoria voidaan loogisesti johtaa äärettömästä määrästä mielikuvituksellisia teorioita, joten mikään ei redusoituisi mihinkään, jos ”johdettavuus vain tietystä teoriasta” ajatellaan loogisena, eikä empiirisenä kriteerinä. Reduktiivisen selittämisen tavoite on siis selventää eri tieteenalojen käsittelemien ilmiöiden välisiä empiirisiä eikä loogisia suhteita.

Ehto 3° on näistä kolmesta filosofisesti mielenkiintoisin, eli kaikkein hämärin. *Luonnollinen luokka* tai *laatu* on melko epämääräinen käsite. Mitä luonnolliset luokat sitten ovatkaan, mielekkäintä lienee määritellä ne suhteessa johonkin teoriaan. Fodor määrittelee luonnollisen luokan luonnollisten ominaisuuksien avulla siten, että predikaatti P poimii luonnollisen luokan teoriassa T, jos ja vain jos T:ssä on ainakin yksi laki, jossa predikaatti P esiintyy (Fodor, 1974, s.102). Käytän jatkossa luonnollisen luokan määritelmää siinä merkityksessä, että predikaatti P joko on teorian T peruskäsite tai määriteltävissä T:n peruskäsitteiden avulla. Luonnollisten luokkien ja ominaisuuksien perusajatus on helppo hahmottaa, jos tieteellisten teorioiden luonteesta tehdään muutamia idealisoivia oletuksia. Oletetaanpa, että hiukkasfysiikka on teoria, jonka peruskäsitteitä ovat muiden muassa ”hiukkanen”, ”massa” ja ”sähkövaraus”, joiden avulla hiukkasteorian lait ja predikaatit muotoillaan. Nyt esimerkiksi termi ”elektroni” poimii erään luonnollisen luokan, kun predikaatti ”x on elektroni” määritellään siten, että x on hiukkanen, jolla on tietty massa, sähköinen varaus ja niin edelleen. Vastaavasti ”x on negatiivisesti varautunut hiukkanen” poimii luonnollisen luokan, johon kuuluvat elektronit, myonit ja muutama muu hiukkanen. ”Negatiivisesti varautunut hiukkanen” on luonnollinen ominaisuus, koska on olemassa yleisiä lakeja, jotka koskevat nimen omaisesti negatiivista sähkövarausta kantavia hiukkasia.

Tarkalleen ottaen Fodor pyrkii reduktiomallinsa avulla osoittamaan, etteivät erityistieteiden lait yleensä palaudu fysiikkaan, muttei tämä kuitenkaan ole ongelma metafyysiselle fysikalismille eikä erityistieteille. Pääpiirteissään argumentti etenee seuraavasti: Erityistieteet, eli oleellisesti muut tieteenalat kuin fysiikka, eivät käsittele suoranaisesti fysiikan alaan kuuluvia ilmiöitä – juuri tämän takia ne ovat erityistieteitä – vaan pyrkivät muodostamaan lainomaisia yleistyksiä ilmiöistä, joiden esiintymillä ei ole olemassa mitään ainakaan ilmeistä yhteistä fysikaalista nimittäjää. Tämä tarkoittaa, että ilmiöiden esiintymiä ei voida luokitella mielekkäästi fysikaalisten predikaattien avulla. Tästä Fodor käyttää esimerkkinä valuutan kiertoa koskevia taloustieteen lainalaisuuksia (Fodor, 1974, s.103). Tietynlainen rahanvaihtotapahtuma voi toteutua fysikaalisesti lukemattomilla tavoilla: tilisiirrolla, kädestä käteen oravannahoilla tai kolikoilla ja niin edelleen. Fysikaalisesti tarkasteltuna eri valuutanvaihtotapahtumilla ei tarvitse olla mitään tekemistä keskenään, vaikka taloustieteellisesti tarkasteltuna ne olisivat identtisiä. Käsittääkseni eri tapahtumat voivat olla taloustieteellisestä kannalta identtisiä jopa siinä tapauksessa, että ne toteutuisivat eri maailmoissa, joissa on voimassa eri fysiikan lait. Näin ollen taloustieteellisiä lainalaisuuksia tuskin voi määritellä fysiikan lakien avulla, olkoonkin niin, että jokainen taloudellinen tapahtuma on jonkinlainen fysikaalinen tapahtuma. Ehto 3° on tässä oleellinen, koska tuskin on olemassa siltalakeja, jotka palauttaisivat taloustieteelli-

set termit, kuten ”vaihtokurssi” mihinkään fysikaaliseen luonnolliseen luokkaan. Yleisesti ottaen sama pätee useiden muidenkin erityistieteiden tapauksessa, poikkeuksena luonnollisesti fysiikkaan palautuvat teoriat, kuten vaikkapa kemia. Ilmiöitä, jotka voivat toteutua fysikaalisesti periaatteessa rajattoman monenlaisilla tavoilla, joita ei yhdistä mikään fysikaalinen luonnollinen luokka, kutsutaan *monitoteutuviksi*.

Tarkoittaako monitoteutuvuus sitten samaa, kuin redusoitumattomuus? Ajatellaan, että siltalaki on muotoa ”Y jos ja vain jos A_1 *tai* A_2 *tai*... *tai* A_n”, eli ilmiö Y voidaan palauttaa disjunktiiviseen luonnollisista ominaisuuksista $A_1, A_2, \ldots, A_n$ koostuvaan luokkaan. Jos nyt tarkastellaan miten ehto 3° on edellä määritelty, kysymys siitä, palautuuko Y kyseiseen disjunktiiviseen luokkaan vai ei on sama kuin onko tuo luokka luonnollinen teoriassa T_A. Tämä taas on sama kysymys kuin, että onko teoriassa T_A lakeja, jotka koskevat mainittua luokkaa vai ei. Mainitsin käyttäväni luonnollisen luokan käsitettä siinä merkityksessä, että luokka on määriteltävissä teorian peruskäsitteiden avulla ja käsittääkseni loogisessa mielessä disjunktiivinen määritelmä on aivan kelvollinen määritelmä siinä missä mikä tahansa muukin. Varsinaisen ongelman muodostavat avoimet disjunktiiviset määritelmät muotoa ”Y jos ja vain jos A_1 *tai* A_2 *tai*...”, missä predikaatit A_i muodostavat äärettömän joukon, jota ei voi äärellisesti määritellä, siis tavallaan sulkea, teorian T_A lakien tai predikaattien avulla. Jatkossa tarkoitan monitoteutuvalla ilmiöllä juuri ilmiötä, jonka mahdollisia toteutumia vastaa avoin disjunktiivinen luokka.[7] Esimerkiksi ”hiukkasjoukon x keskimääräinen kineettinen energia” viittaa äärettömän monenlaiseen fysikaaliseen tilaan, jossa hiukkasilla on erilaiset nopeudet ja suunnat. Tilojen joukko on kuitenkin suljettu, koska hiukkasten keskimääräinen energia on mekaniikan luonnollinen ominaisuus, joka määrittelee täsmällisesti, mitä yhteistä noilla kaikilla tiloilla on. Tässä mielessä lämpötila ei ole monitoteutuva ilmiö, vaikka ”kaasun lämpötila” viittaa äärettömään määrään kaasun mahdollisia tiloja. Kuitenkin esimerkiksi *valuutta* lienee fysiikan suhteen monitoteutuva ilmiö, koska käypää valuuttaa ja sen arvoa ei määritä fysiikka, vaan lainsäädäntö ja eräät yhteiskunnalliset instituutiot. Tietysti on loogisesti mahdollista, että laki ja yhteiskunnan instituutiot, ja sitä myötä valuutta, jollain tavalla palautuvat fysiikkaan, mutta tämä mahdollisuus tuntuu lievästi sanottuna äärimmäisen etäiseltä. Jos monitoteutuvuus tarkoittaa, että ilmiön mahdollisia toteutumia vastaa avoin disjunktiivinen predikaatti, niin monitoteutuvuus on ainakin riittävä ehto redusoitumattomuudelle, koska ”Y jos ja vain jos A_1 *tai* A_2 *tai*...” ei ole minkäänlainen laki. Predikaatti ”A_1 *tai* A_2 *tai*...” ei ole edes määritelty, joten on mahdotonta sanoa mihin tämä ilmaus viittaa. Näin ollen sen sijoittaminen siltalakiin ei tuota minkäänlaista ymmärrettävää totta tai epätotta yleistystä.

Yllä kuvatun reduktiomallin kannattajan ei millään tavalla tarvitse kiistää metafyysisen materialismin pätevyyttä. Mallista seuraa vain, että on olemassa lainomaisia erityistieteiden yleistyksiä, jotka voidaan – ja usein täytyy – määritellä sellaisten predikaattien

[7] Huomautettakoon, että teoria T_Y voi olla monitoteutuva myös muidenkin teorioiden kuin fysiikan suhteen, jolloin oleellisesti samat tarkastelut pätevät. Asiaa käsittelevässä artikkelissaan Fodor antoi melko hyviä perusteluja epäillä, etteivät myöskään äärelliset disjunktiiviset predikaatit yleensä vastaa luonnollisia luokkia (Fodor, 1974, s.108–109). En nyt puutu tähän sen syvällisemmin, koska varsinaisesti monitoteutuvien ominaisuuksien tarkastelu riittää jatkossa eikä ole kovin selvää, että argumentti pätee äärellisten disjunktioiden tapauksessa. Kritiikistä ks. (Kim, 1992) ja Fodorin myöhempi vastaus (Fodor, 1997).

avulla, jotka eivät redusoidu fysiikan luonnollisiin luokkiin. Fysikalisti voi toisaalta hyväksyä tällaisen monitoteutuvuuden, mutta vedota siihen, että jos ilmiön Y nomologisesti mahdolliset fysikaaliset toteutumat muodostavat avoimen joukon $F_1, F_2, \ldots$, niin tällöin itse asiassa on totta, että Y on F_1 *tai* F_2 *tai*..., ja predikaatti "F_1 *tai* F_2 *tai*..." on hyvin määritelty, vaikka emme pysty muotoilemaan tai tietämään kyseistä määritelmää. Joka tapauksessa fysikaalisia ilmiöitä $F_1, F_2, \ldots$ yhdistää ainakin se, että ne toteuttavat tarkasteltavan ylemmän tason ilmiön, joten juurikin ilmiö Y määrittelee kyseisen luokan. Käsittääkseni tämä pitää paikkansa. Tällaisessa luokan $F_1, F_2, \ldots$ määrittelmässä on kuitenkin ongelmallista, että se on muotoiltu redusoitavan eikä redusoivan teorian kielellä. Näin ollen mainitun yhtäpitävyyden oletettu pätevyys ei millään tavalla auta ymmärtämään ilmiötä Y fysikaalisesti, ja edelleen koska predikaatin "F_1 *tai* F_2 *tai*..." määritelmää ei voi muodostaa, ei mainittu tosiasia mahdollista minkään varsinaisen reduktiivisen teorian muotoilemista.

Perinteisesti reduktion analyysissä on painotettu reduktioiden formaaleja ehtoja, mutta toisaalta taas teoriareduktiot ovat vahvasti kytköksissä tieteelliseen selittämiseen, joka loppujen lopuksi on epistemologista tai kognitiivista toimintaa. Tässä mielessä siltalakien muodostaminen ei paljolti poikkea mistä tahansa tieteellistä teorianmuodostuksesta. Vaikka tieteellisten teorioiden tarkoitus on kuvata todellisuutta oikein, niin ensisijaisen tärkeää on, että teoriat kuvaavat todellisuutta meille. Ehdon 3° muodollisen määritelmän varsinainen ydin on hylätä sellaiset reduktiot, jotka eivät millään tavalla auta ymmärtämään redusoitavan teorian ilmiöitä redusoivan teorian avulla.

Tarjottu reduktiomalli pitää sisällään muutamia epäilyttäviä oletuksia, esimerkiksi että tieteelliset teoriat ovat hyvin määritellyistä peruskäsitteistä rakentuvia lausejoukkoja. Myös Fodor suhtautui hieman varauksella reduktiomallin sisältämiin käsityksiin teorioiden luonteesta ja luonnollisista luokista (Fodor, 1974, s.102). Argumentista kuitenkin tekee mielenkiintoisen nimenomaan se, että vaikka perinteiset positivistiset käsitykset tieteellisistä teorioista, selittämisestä ja reduktiosta hyväksyttäisiinkin metafyysisen materialismin ohella, niin reduktiivinen fysikalismi on silti tarpeettoman vahva ja epäuskottava sitoumus. Joka tapauksessa edellä esitetyn kaltaisilla tarkasteluilla on ollut voimakas vaikutus mielenfilosofian kehitykseen suurinpiirtein 60-luvulta eteenpäin.[8] Seuraavissa luvuissa tarkastellaan lähemmin kehitystä, joka johti mentaalisuuden monitoteutuvuuden melko yleiseen hyväksymiseen ja psykofysikaalisen reduktionismin hylkäämiseen. Tämä ei kuitenkaan johtanut metafyysisisen materialismin kiistämiseen eikä mielen reduktiivisen selittämisen tavoitteesta luopumiseen, vaan uudenlaiseen naturalismiin, jota usein kutsutaan antireduktionistiseksi materialismiksi. Hieman tarkemmin sanottuna tämä tarkoitti kognitivismin nousua ja ryhtymistä tarkastelemaan mieltä tietojenkäsittelyjärjestelmänä, minkä katsotiin tarjoavan mentaalisuudelle metafyysisen materialismin kanssa yhteensopivan, mutta fysiikkaan palautumattoman teorian.

[8] Ks. esim. (Kim, 1992, s.1–4). Myös (Kim, 1989) sisältää hyvän yhteenvedon ja kriittisen analyysin mielenfilosofiassa yleiseen antireduktionismin hyväksymiseen johtaneista argumenteista.

2 Komputationaalisen mielenteorian juuret

Kognitivistinen mielenteoria koostuu oleellisesti kahdesta osasta. Toinen näistä on niin sanottu komputationaalinen mielenteoria, jonka perusidea on, että psykologiset prosessit voidaan kuvata mentaalisia representaatioita käsittelevinä algoritmeina. Tähän liittyy hyvin tiivisti käsitys, että mentaaliset representaatiot ovat lauseiden kaltaisia symbolirakenteita, jotka muodostavat eräänlaisen kielen. Tässä yhteydessä ”kieli” tarkoittaa oleellisesti formaalikielen kaltaista symbolijärjestelmää, jota usein kutsutaan *ajattelun kieleksi*. Koputationalismi nojaa siis representationaaliseen mielenteoriaan, ja täsmentää sitä toteamuksella, että mentaalisten representaatioiden esiintymien kausaaliset ketjut ovat komputationaalisia prosesseja. Esimerkiksi päättely on teorian mukaan ajattelun kielen lauseiden algoritmista tuottamista ja manipuloimista.

Ajatus, että järjenkäyttö olisi toteutettavissa säännönmukaisena symbolirakenteiden manipuloimisena, kumpuaa pääasiassa 1800- ja 1900-lukujen taitteessa syntyneestä aksiomaattisesta logiikasta. Samasta tutkimusohjelmasta syntyneet tietokoneet puolestaan ovat konkreettinen todiste, että looginen päättely tai mikä tahansa muu algoritmisoitavissa oleva formaali toiminta voidaan suorittaa mekaanisesti fysikaalisella laitteella. Mikäli representationalistisen mallin mukaisesi kaikki mentaalinen toiminta järkeilyn lisäksi on mentaalisten representaatioiden manipulointia, voidaan periaatteessa kaikkia psykologisia prosesseja vastaavat algoritmit toteuttaa tietokoneilla, ainakin mikäli lisäksi oletetaan, että mieli on jonkinlainen mekanismi, kuten naturalistisessa filosofiassa usein tapana on.

Filosofisesti erityisen mielenkiintoista komputationalismissa on, että jos psykologiset prosessit todella voidaan algoritmisoida, mielen mysteereistä on ratkaistu ainakin seuraavat: mitä ajatukset ovat, mitä ajattelu on, ja ennen kaikkea miten fyysinen olio ylipäätään voi ajatella, eli miten mentaalinen toiminta voi nousta materiaalisista mekanismeista? Viimeinen mysteeri lienee näistä keskeisin. Mikäli psykologian palautus algoritmeihin ja laskentaan onnistuu, on psykofyysinen ongelma pitkälti ratkaistu, koska tietokoneiden toiminnassa ei liene mitään erityisen mystistä. Ennen kaikkea tietokoneiden toiminta ei vaadi minkäänlaista järjellistä, ajattelevaa tai muuten intentionaalista komponenttia, joten koneet tarjoavat houkuttelevan mallin ajattelun naturalistiseksi ymmärtämiseksi.

Tämä luku pitkälti keskittyy siihen tieteellisen kehitykseen, jota vasten ajatus ajattelun algoritmisoinnista vaikuttaa itse asiassa yllättävänkin luontevalta. Tarina alkaa 1800-luvun loppupuolelta Fregestä ja päättyy tietokoneiden syntyyn sekä 1950 julkaistuun Alan Turingin artikkeliin ”Computing Machinery and Intelligence”. Luvun tarkoitus ei ole luoda tarkkaa historiallista kuvausta 1900-luvun alun matematiikan tutkimuksesta, vaan lähinnä esitellä komputationalismin asianmukaiseen ymmärtämiseen tarvittavien teorioiden kehitystä, mikä kattaa lähinnä modernin matemaattisen logiikan sekä formaalikielten ja laskennan teorian. Luvun lopuksi käsitellään näiden teorioiden sekä tietokoneiden mielenfilosofista merkitystä. Tehdään ennen tätä kuitenkin pikainen yhteenveto behaviorismin aikakaudesta sekä sen syntymään ja kuolemaan johtaneista syistä. Tarkoitus on luoda kuvaa siitä tieteellisestä viitekehyksestä, johon kognitivismi syntyi. Kuten johdannossa todettiin, komputationaalinen teoria kehittyi pitkälti filosofiasta ja matematiikasta, mutta teorian tarve kumpusi lähinnä psykologiasta ja sille läheisistä tutkimusohjelmista.

2.1 Käyttäytymistieteistä kognitivismiin

Arkipsykologia on ollut markkinoilla hyvän aikaa muodossa tai toisessa, ja siihen törmätään jo ainakin Aristoteleen kirjoituksissa. Esimerkiksi *Nikomakhoksen etiikkassa* kirjan VII kolmannessa luvussa käsitellään praktista syllogismia ja muun muassa todetaan, että "... halu johtaa maistamiseen, sillä se voi liikuttaa kaikkia jäseniämme" (Aristoteles, 1989, s.127). Nimitys "arkipsykologia" ei pyri olemaan väheksyvä samaan tapaan kuin esimerkiksi "kyökkipsykologia", vaan mallin arkisuus liittyy juurikin sen pitkään perinteeseen, ainakin länsimaisen ajattelun historiassa, jonka myötä uskomus–halu-käsitteistö on juurtunut melko väistämättömäksi tavaksemme ymmärtää mieltä ja käyttäytymistä.

Representationalismin synnyinsija on hieman tulkinnanvarainen kysymys, mutta teoriaa kehitettiin merkittävästi uudella ajalla erityisesti empiristisessä filosofiassa. Kun Locke sanoi, että havaitsemme ideoita, hän tarkoitti ettemme havaitse maailmaa suoraan, vaan havaintomme ovat asiaintilojen representaatioita (Copleston, 1959, s.88–89). David Hume muotoili representationalismin vielä selvemmin, kun hän yhdisti kaikki mielentilat havaintojen kaltaisiin asiaintilojen edustuksiin tai niiden johdannaisiin (*ibid.*, s.293–305). Nämä käsitykset liittyivät kysymykseen mielen ja maailman suhteesta, täsmällisemmin sanottuna tietokykymme olemuksesta. Brittiläiset empiristit Hobbesista Humeen vaikuttivat olevan hyvin kiinnostuneita nimenomaan mentaalisten representaatioiden ja prosessien luonteesta, ja modernin representationalismin alkujuuret näyttävät palautuvan jotakuinkin tähän koulukuntaan. Humea onkin kutsuttu ensimmäiseksi kognitiotieteilijäksi. Vaikka tätä ei ehkä esitetä täysin vakavissaan, ei väite kuitenkaan ole täysin hatusta vedetty. Hume esimerkiksi esitti, että havaitsemme maailmassa kausaatiota, koska havaitsemme toistuvasti tietyn seurauksen seuraavan tiettyjä tapahtumia, mutta tosiasiassa aistien tarjoamassa materiaalissa ei ole mitään kausaalisuhdetta havaittavaksi (*ibid.*, 278–283). Tämä jossain määrin muistuttaa nykyisen kognitivismin perusajatusta, jonka mukaan mieli osallistuu aktiivisesti sekä havainnointiin että maailmaa koskevien käsitystemme muodostamiseen, eikä se ole vain passiivinen aisti-informaation vastaanotin.

Toisaalta myös uuden ajan rationalistisessa leirissä jonkinlainen representationalismi oli kova sana, ja teorian merkkipaaluna usein pidetään Descartesin *Meditationes de Prima Philosophiaa* vuodelta 1641. Esimerkiksi toisen mietiskelyn lopussa todetaan seuraavasti: "... tiedän, että itse kappaleitakaan ei varsinaisesti havaita aisteilla eikä kuvittelukyvyllä, vaan ainoastaan ymmärryksellä, ja havainto ei johdu koskettamisesta tai näkemisestä, vaan yksinomaan ymmärtämisestä, oivallan kirkkaasti, että en voi havaita mitään helpommin tai ilmeisemmin kuin oman mieleni." (Descartes, 2002, s.44.) Edellisen kaltaiset kommentit ja Descartesin radikaalin epäilyn metodi kertovat siitä filosofisesti merkittävästä oivalluksesta, että havainnot ja uskomukset ovat asiaintiloista erillisiä mielen sisäisiä ilmiöitä, joilla ei tarvitse olla vastinetta maailmassa.

Tieteellisen psykologian synty puolestaan ajoittuu suurin piirtein 1800-luvun puoliväliin, jolloin fysiologit Hermann von Helmholtz ja hänen assistenttinansa Heidelbergissä 1858–1864 toiminut Wilhelm Wundt tutkivat kokeellisesti muun muassa havainnointia ja hermoston toimintaa. Siinä missä Helmholtzin tutkimukset koskivat lähinnä aistihavaintojen fysiologiaa, ryhtyi Wundt kutsumaan tutkimustaan "fysiologiseksi psykologiaksi" ja jul-

kaisi laajasti tekstejä psykologiasta autonomisena tieteenä. (Bechtel et al. 1998, s.14–15; Gregory 1987, s.308–310,816–817.) Wundtin lisäksi huomattavaa psykologian pioneerityötä teki William James, jonka *Principles of Psychology* (1890) on varmaankin merkittävin yksittäinen teos psykologian syntyhistoriassa. Siinä missä Wundtia voinee pitää kokeellisen psykologian isänä, kuuluu vastaava status Jamesille teoreettisen psykologian piirissä (Bechtel et al. 1998, s.15; Gregory 1987, s.650–651). Heille oli yhteistä toisaalta empiirinen, vastoin kuin puhtaasti filosofinen tai spekulatiivinen, ote mielen tutkimukseen ja toisaalta taas huomattava introspektion käyttö psykologisen tiedon lähteenä. Tässä vaiheessa psykologian tehtäviin ei ainakaan selkeästi katsottu kuuluvan varsinaisen mielenteorian muodostamista, siis mentaalisten ilmiöiden naturalistista selittämistä, vaan psykologia oli mitä psykologia pohjimmiltaan on, eli mielen toiminnan empiiristä tutkimista. Tämä suhtautuminen kuitenkin alkoi muuttua vuosisadan vaihtuessa.

1900-luvun alussa tieteenfilosofinen ilmapiiri kehittyi suuntaan, jossa suhtauduttiin perin nihkeästi havainnon tavoittamattomiin teoreettisiin ilmiöihin, kuten mielentiloihin ja mentaalisiin representaatioihin. Pyrkimyksen tehdä mentalistisiin käsitteisiin pohjautuvaa empiiristä tiedettä katsottiin johtavan ylitsepääsemättömiin ongelmiin, koska selvästi mielentilat eivät ole ainakaan julkisesti havaittavissa, vaan voimme loppujen lopuksi havaita ainoastaan käyttäytymistä ja muita fysikaalisia ilmiöitä. (Watson, 1913, s.163–167.) Näin introspektiivinen metodologia joutui huonoon valoon. Mikä vielä pahempaa, 1800-luvun psykologiassa puhuttiin alitajuisista mielen toiminnoista, jotka ovat havainnoinnin mahdollisuuden ulkopuolella jopa subjektille itselleen.[9] Vakaumus, jonka mukaan sisäisiin mielentiloihin viittaava teoria ei voi olla tiukassa mielessä empiirinen, johti tieteelliseen behaviorismiin, joka käytännössä hallitsi tieteellistä psykologiaa erityisesti Pohjois-Amerikassa viime vuosisadan ensimmäisen puoliskon ajan (Bechtel et al., 1998, s.4).

Behaviorismi sisältää vähimmillään käsityksen, että kaikki tieteellisessä psykologiassa hyväksyttävä empiirinen aineisto pohjautuu organismien fysikaalisiin tiloihin ja ulkoisesti havaittavaan käyttäytymiseen. Burrhus F. Skinnerin markkinoiman radikaalin behaviorismin mukaan psykologian teorioissa on sallittua viitata ainoastaan käyttäytymiseen, mikä tarkoittaa, ettei psykologisissa selityksissä ole sallittua puhua edes niistä fysiologisiksi miellettyistä mekanismeista, jotka välittävät ärsykkeen ja reaktion välistä suhdetta. Mitä behavioristit yleisesti ajoivat takaa oli kuitenkin vain, että puhe organismin sisäisistä mentaalisista tiloista tulisi unohtaa. (Kim, 2006, s.74–78.) Yleensä kai sisäisten fysiologisten ilmiöiden, kuten aivotilojen, ei pitäisi aiheuttaa ongelmia, koska ne kuitenkin ovat empiirisin keinoin tutkittavissa. Joka tapauksessa tieteellinen behaviorismi – muodossa missä hyvänsä – tarkoitti, ettei representationaalisella mielenteorialla ollut sijaa sielutieteistä käyttäytymistieteeksi muuttuneen psykologian piirissä.

Behaviorismi ei kukoistanut vain tieteellisen psykologian alueella, vaan vaikutti vahvasti myös filosofiassa lähinnä loogisten positivistien leirissä. *Loogisen behaviorismin* mukaan jokainen mielekäs psykologinen termi voidaan täysin määritellä käyttämällä pelkästään

[9] ”Alitajuinen mielen toiminta” tuonee useimmille mieleen aktiivisen alitajunnan piilevine haluineen ja pelkoineen. Tässä ”alitajuinen” on kuitenkin huomattavasti laveampi käsite. Esimerkiksi Helmholtz tarkoitti alitajuisilla toiminnoilla muun muassa havaintapahtumiin liittyviä (aivo)prosesseja, joilla ei koskaan ole havaittavia edustuksia tietoisuudessa. (Gregory, 1987, 309) Jatkossa käsitettä *alitajuinen* käytetään tässä laveassa merkityksessä.

fysikaalisia ja käyttäytymiseen viittaavia ei-intentionalistisia käsitteitä (Hempel, 1935, s.89–90). Tämä ei enää ole vain tieteen tekemistä koskeva metodologinen oppi vaan reduktionistinen väite, jonka mukaan kaikki todet psykologiset teoriat voidaan muotoilla puhumalla pelkästään käyttäytymisestä. Carl Hempelin artikkelissa "The Logical Analysis of Psychology" (1935) esitetty argumentti loogisen behaviorismin puolesta etenee pääpirteiltään seuraavasti: 1. Kaikkien mielekkäiden väitelauseiden merkitys palautuu niihin ehtoihin, jotka täytyy varmistaa, jotta väitettä voidaan pitää totena. Tämä on positivistisen filosofian ytimeen kuulunut *merkityksen verifikaatioteesi*, joka on pitkälti sama kuin väite, että lauseen merkitys palautuu sen totuusehtoihin. 2. Väitteellä voi olla intersubjektiivista tai kommunikoituvaa sisältöä, vain jos sen verifikaatioehdot ovat julkisesti havaittavia, eli kuka tahansa voi periaatteessa empiirisesti koetella väitettä, ja 3. vain fysikaaliset ilmiöt, mukaan lukien käyttäytyminen, ovat julkisesti havaittavia. Näin ollen jokainen psykologinen väite, jota voi pitää totena, epätotena tai ylipäätään mielekkäänä, täytyy olla käännettävissä fysikaalisia ilmiöitä tai käyttäytymistä koskevaksi väitteeksi. (Hempel, 1935, s.88–92.)

Behaviorismi ei suinkaan ollut ainoa elinvoimainen psykologian suuntaus 1900-luvun alkupuoliskolla, mutta sen erityinen historiallinen merkitys piilee pitkälti siinä, että se oli ensimmäinen mielenteoria, joka tarjosi sekä reduktionistisen, naturalistisen analyysin psykologiasta – tai oikeastaan ehdotelman sellaisesta – että tutkimusmetodologian, jotka yhdessä muodostivat paradigman aikansa tieteelliselle mielentutkimukselle. Liikkeen taustalla niin filosofien kuin psykologienkin piirissä oli halu nähdä psykologia luonnontieteenä ja perusteltu kriittinen suhtautuminen introspektioon tieteellisen tiedon lähteenä. Joillekin behaviorismi tarkoitti eliminativismia,[10] mutta pääsääntöisesti behavioristiset psykologit olivat mainettaan maltillisempia. Kaikki eivät suinkaan kiistäneet mielentilojen olemassaoloa, vaikka ehkä ajattelivat niiden olevan empiirisen tutkimuksen tavoittamattomissa ja käyttäytymisen selittämisen kannalta tarpeettomia. Esimerkiksi Skinner piti sisäisiä kokemuksia kyllä todellisina, mutta hänen mukaansa ne olivat käyttäytymisen seurauksia eikä syitä ja näin ollen epäoleellisia käyttäytymisen etiologian kannalta (Gregory, 1987, s.73).

Mikäli muita tutkimusmenetelmiä ehdollisen[11] käyttäytymisen tarkastelun ja introspektion lisäksi ei näyttäisi olevan tarjolla, ensiksi mainittuun rajoittuminen ei välttämättä ole huonoin mahdollinen idea. Muun muassa behavioristisen koulukunnan perustajana pidetyn John B. Watsonin kuuluisasta artikkelista "Psychology as Behaviorist Views it" (1913) on luettavissa, että behaviorismi kumpusi nimen omaan introspektiivisen metodologian heikkouksista sekä uskosta, että käyttäytymistä voidaan tutkia ilman sisäisten tilojen tutkimusta, eikä niinkään esimerkiksi positivistisista tiedekäsityksistä. Jälkimmäistä kantaa

[10] Ks. esim. Watsonin melko intomielinen hyökkäys kaikkea ei-behavioralistista psykologiaa vastaan (Watson, 1930, s.3–18), jossa hän syytti tietoisuuden tiloihin viittaavien käsitteiden pohjautuvan uskonnollisiin ja dualistisiin harhakäsityksiin mielestä.

[11] Behavioristit eivät tutki käyttäytymistä sinänsä, vaan käyttäytymisen ja tiettyjen olosuhteiden välisiä suhteita, siksi ilmaus "ehdollinen käyttäytyminen". Ehdollisen käyttäytymisen ei kuitenkaan tarvitse olla ehdollistunutta, vaan se voi perustua esimerkiksi vieteille. Käytän tällaista laveampaa käsitettä, koska en nyt halua mennä yksityiskohtiin behavioristisen teorian muotoilussa. Joka tapauksessa kaikki behavioristit uskovat, että käyttäytyminen riippuu aina osittain ulkoisista olosuhteista ja on sikäli aina ehdollista, mutta ehdollistumisen asema teorioissa vaihtelee hieman teoreetikolta toiselle.

Watson perusteli sillä, että eläinten käyttäytymistä voidaan ennustaa ja selittää viittaamatta niiden tietoisuuden tiloihin, joten samaa metodologiaa voidaan soveltaa myös ihmisiin, jotka lopulta ovat vain eräänlaisia eläimiä (Watson, 1913, s.176–177). Watson oli vaikuttunut muun muassa Ivan Pavlovin työstä, ja behaviorismin syntyhistorian kannalta onkin sivumaininnan arvoinen huomio, että Watsonin urasta merkittävä osa koostui eläinten käyttäytymisen tutkimisesta, mitä käsitteli myös hänen 1903 julkaistu väitöskirjansa (Gregory, 1987, s.71–72).

Monet näkivät behaviorismissa mahdollisuuden naturalisoida intentionalistinen psykologia. Vaikka mentalismin ja behaviorismin yhteensovittamista pidetään erityisesti filosofisen behaviorismin keskeisenä harrasteena, pyrittiin tähän laajalti myös psykologian piirissä behaviosismin radikaalia laitaa unohtamatta.[12] Filosofisen behaviorismin puitteissä eräs kauneimmista kukkaista on Gilbert Rylen *The Concept of Mind* (1949), jossa muun muassa pyritään osoittamaan, että vaikka yleensä ajattelemme mentaalisten termien viittaavan jonkinlaisiin mielensisäisiin ilmiöihin, niin tosiasiassa näitä termejä käytetään kuvailemaan kykyjä tai systemaattisia taipumuksia toimia tietyin tavoin tietyissä olosuhteissa. Mikäli näin on, behaviorismi ei olekaan radikaali revisionistinen oppi, vaan sisäänrakennettu niin sanottuun mentalistisen kielenkäytön logiikkaan, kun oikein tarkkaan katsotaan. Tällöin kuitenkin propositionaaliset asenteet tulee tulkita organismin käyttäytymistaipumuksiksi eikä mielen sisäisiksi tiloiksi. On toisaalta hankala nähdä, miten mentaaliset representaatiot voitaisiin määritellä behavioristisesti, joten representationalismi ja behaviorismi tuntuvat sopivan huonosti yhteen. Tähän liittyvät vaikeudet osoittautuivatkin behaviorismille kohtalokkaaksi, mistä lisää myöhemmin. Mielenteorian kannalta olisi kuitenkin hyvin mielenkiintoista, mikäli uskomuksia, haluja ja muita mentaalisia ilmiöitä koskevat väitteet kyettäisiin määrittelemään behavioristisesti. Onnistuessaan tällainen projekti nimittäin osoittaisi, ettei behavioristinen näkökulma psykologiaan jätä mielenkiintoisia kysymyksiä ja ilmiöitä tutkimuksen ulkopuolelle. Eliminativismihan ei uskomusten ja halujen luonnetta tai mielen ja ruumiin suhdetta selitä, vaan hylkää koko kysymykset. Behaviorismi ei ehkä ole kovin suora reitti mentaalisuuden fysikalistiseen redusoimiseen, mutta jos mentaaliset ilmiöt osoittautuvat monimutkaisten refleksien välittämiksi käyttäytymistaipumuksiksi, ei niiden pitäisi aiheuttaa materialisteille sen kummempia käsitteellisiä ongelmia kuin vaikkapa suoliston toiminnan.

Nyt puolisen vuosisataa sen jälkeen, kun tämä keskustelu varsinaisesti oli ajankohtainen, on helppo sanoa, ettei kovin tyydyttävää behavioristista teoriaa saatu aikaa eikä tulla saamaan, koska kukaan ei ole tätä työtä enää tekemässä. On melko selkeitä syitä, miksi behaviorismi oli tuomittu epäonnistumaan. Käsitys, jonka mukaan behaviorismi upposi 1900-luvun loppupuoliskolla loogisen positivismin mukana, sisältää ehkä ripauksen totuutta. Olisi kuitenkin kummallista, jos tämä olisi lähimainkaan koko tarina, koska behaviorismi ei varsinaisesti ollut positivistien vaan psykologien projekti. Käsittääkseni behaviorismin kohtaloksi koitui yksinkertaisesti se, että psykologit saivat parempia ideoita, mutta teoriassa on myös melko helppo nähdä muutama melko ylitsepääsemätön ongelma. Ensinnäkin intentionaalisia termejä ilmeisesti on yksinkertaisesti mahdotonta määritellä pelkästään käyttäytymistaipumusten avulla, ja toiseksi teorian puitteissa on ilmeisen mahdotonta kuvata oikein käyttäytymistä, jonka kannalta useat mielentilat ovat

[12] Ks. esim. (Skinner, 1945, s.271).

samaan aikaan oleellisia. Behavioristinen malli mielestä on siis liian yksinkertainen, jotta siitä voisi johtaa arkipsykologisia lainalaisuuksia, joten uskomus–halu-yleistykset eivät palaudu käyttäytymistaipumuksiksi. Skinnerin teoria ajattelusta pääsi ehkä lähimmäksi maalia. Katsotaan aluksi miten se selviytyi päätöksen tekemisestä, päättelystä ja muista mielentilojen ketjuja edellyttävistä toiminnoista.

Ajatellaan, että henkilölle x tarjotaan kahta asuntoa A ja B. Hetken aikaa hän silmäilee niiden tietoja, raapii päätään ja lopulta sanoo "valitsen asunnon B." Yleensä silmäilyn ja pään raapimisen aikana oletamme tapahtuvan vertailua, joka koskee asuntojen hintoja, tilavuuksia, sijainteja ja niin edelleen. Päätöksenteko tai ongelmanratkaisu on prosessi, joka tapahtuu ennen päätöksen toimeenpanoa ja joka vaikuttaisi yleensä edellyttävän harkinnaksi kutsuttua mielentilojen ketjua. Skinner tiedosti tämän ja määritteli esimerkiksi ongelmanratkaisun "muuttujien manipuloimisena", missä organismin reaktio tuottaa uuden ärsykeen, joka tuottaa uuden reaktion, joka tuottaa uuden ärsykkeen ja niin edelleen, kunnes viimeinen reaktio poistaa ongelman. Lähtökohtaisesti tämä ei ole kovin huono idea. Esimerkiksi allekkain laskeminen lienee juuri tämän tyyppinen tapahtuma. Hieman yllättäen nämä muuttujat kuulemma voivatkin olla yksityisiä tapahtumia organismin sisällä, joskaan niillä ei ole mitään tekemistä sellaisten fiktiivisten prosessien kuten ajattelun kanssa. (Skinner, 1965, s.242–243,252)

Behavioristiset määritelmät mielentiloille voidaan karkeasti ottaen muodostaa seuraavan skeeman avulla: *Oliolla x on mielentila M, jos ja vain jos olosuhteissa Y x tekee teon Z todennäköisyydellä P* (Scriven, 1956, s.107,119). Toisin sanoen esimerkiksi x on *nälkäinen*, tai *haluaa syödä*, jos ja vain jos x syö korkealla todennäköisyydellä kun ruokaa on tarjolla. Skinnerille yksityiset sisäiset tilat ovat muun muassa hammassärkyjä, nälkää ja muita enemmän tai vähemmän fysiologisia ilmiöitä. Uskomukset ja arvostelmat taasen ovat parhaiten määriteltävissä kielellisten reaktioiden avulla, esimerkiksi *x uskoo, että Z, jos ja vain jos x sanoo suurella todennäköisyydellä "kyllä" kysyttäessä "päteekö Z?* Erityisesti mielentiloja koskeva puhe, olipa kyse omista tai muiden mielentiloista, on kielenkäyttöä, joka saa alkunsa julkisesta muiden ihmisten käyttäytymistä koskevasta puheesta. Mitä mielentilojen ketjut sitten ovat esimerkiksi ongelmanratkonnan yhteydessä? Eräänlaista puhetta, josta ei kuulu ääntä. Sisäisen puhunnan aktit tuottavat uuden ärsykkeen, joka tietyllä tapaa edustaa päättelyn välitulosta. Skinner siis hyväksyy tieteellisenä hypoteesina johdannossa esitetyn Sellarsin ajatuskokeen sillä erotuksella, että Skinnerille puheaktit ovat jonkinlaisia fysiologisia refleksejä jotka olio voi itse aistia, kun taas Sellarsin mukaan ne ovat intentionaalisia mentaalisia tapahtumia.[13]

Johdannossa sivulla 10 esitetyn onnettomuustilanne-esimerkin oli tarkoitus perustella, että monissa tilanteissa ihmisten käyttäytymisen ymmärtämiseksi ja ennustamiseksi on välttämätöntä ottaa huomioon hyvin monenlaisia psykologisia tekijöitä, kuten toimijoiden havaintoja, uskomuksia ja haluja. Skinner tuntui suhtautuvan loppujen lopuksi melko leväperäisesti intentionaalisen toiminnan behavioristisen kuvausten muotoilemiseen. Hän esimerkiksi esitti, että "x etsii silmälasejaan" voidaan määritellä jotakuinkin siten, että "x tekee jotain sellaista, minkä hän lopettaa löydettyään lasinsa" tai "x tekee jotain sellaista, joka aiemmin on johtanut lasien löytymiseen." Hän kyllä myönsi näiden määritelmien

[13]Skinnerin behavioristinen ajattelun teoria esitetään teoksen (Skinner, 1965) luvussa "Private events in a natural science", s.257–282.

olevan hyvin karkeita, mutta tämä johtui kuulemma vain siitä, että kuvaukset toiminnan päämääristä ja tarkoituksista ovat varsinaisten behavioraalisten määritelmien lyhennyksiä (Skinner, 1965, s.90).

Michael Scriven on huomauttanut, että tällaiset tilanteet ovat huomattavasti monimutkaisempia. Jos joku sanoo etsivänsä lasejaan, on pääteltävissä ainakin, että hän uskoo omistavansa silmälasit eikä tiedä missä ne ovat, hän haluaa löytää lasinsa ja tekee jotain sellaista, jonka hän arvelee johtavan niiden löytymiseen ja niin edelleen. Uskomiseen, arvelemiseen ja haluamiseen viittaavat mentalistiset termit eivät varmaankaan ole eliminoitavissa tällaisista kuvauksista, koska lienee käsitteellisesti mahdotonta, että joku esimerkiksi etsii lasejaan, ellei hän esimerkiksi usko omistavansa sellaisia. (Scriven, 1956, s.106) Tällaiset kuvaukset saattavat olla käännettävissä behavioralistiselle kielelle, mutta yleisesti tämä vaikuttaa epäuskottavalta erityisesti kielellisen käyttäytymisen kohdalla. Tarkastellaanpa behavioristisen yleistysten luonnetta: *Olosuhteissa Y x toimii tavalla Z todennäköisyydellä P* ja esimerkiksi haaveilua. *Haaveilun* operationalisoinnin voitaisiin ajatella olevan sellainen, että henkilö x haaveilee Y:stä, jos ja vain jos hän vastaa suurella todennäköisyydellä ”kyllä” kysyttäessä haaveiletko hän Y:stä. Mutta voidaan olettaa, että henkilö yleensä kiistää haaveilevansa sosiaalisesti tuomittavista asioista. Operationalisoidaanpa hiukan: *x haaveilee seikasta Y, jos ja vain jos hän suurella todennäköisyydellä vastaa ”kyllä” kysyttäessä haaveiletko hän seikasta Y tai ”en”, jos Y on sosiaalisesti arveluttavaa.* Edellinen tarkoittaa nyt sitä, että normaali kansalainen kiistää haaveilevansa sosiaalisesti arveluttavista asioista normaaleissa kommunikaatiotilanteissa, oli hänellä tällaisia haaveita tai ei. Näin ollen joko kukaan ei haaveile sosiaalisesti arveluttavista asioista, koska kukaan ei suurella todennäköisyydellä myönnä tätä, tai sitten haaveilua ei voi operationalisoida. Haaveilu on tässä esimerkkinä siksi, että ihmiset voivat haaveilla asioista, joita he eivät oikeastaan halua tai periaatteessa edes voi toteuttaa, joten operationalisointi ei onnistu tältäkään rintamalta.

Tästä ongelmasta on kuitenkin helppo pakotie: henkilö voi myöntää arveluttavat fantasiansa tietyissä epätyypillisissä olosuhteissa, kuten terapeutin vastaanotolla, jos ne esimerkiksi aiheuttavat hänelle ahdistusta ja hän haluaa tilanteeseen apua. Ongelmana tässä on, että ärsykeympäristönä psykologin vastaanotto voi olla täsmälleen samanlainen kuin mikä tahansa muu keskustelutilanne. Oleellista on, että x uskoo keskustelukumppaninsa olevan terapeutti, jolla on vaitiolovelvollisuus, joten arveluttavista tunnustuksista ei seuraa sosiaalista haittaa. Lisäksi x:n täytyy uskoa, että hänen tarpeensa saada apua parhaiten tyydyttyy olemalla avoin ja rehellinen. Toisaalta jos x:llä on jokin syy uskoa, että terapeutti vatiolovelvollisuudestaan huolimatta levittelee potilaittensa yksityisasioita, hän luultavasti ei paljasta salaisia haaveitaan. On ehkä mahdollista, että behavioristinen uskomusten ja halujen operationalisointi toimii joskus, mutta tällaisissa melko yksinkertaisissakin tapauksissa päädytään äkkiä tilanteeseen, jossa pelkästään ärsykkeen ja reaktion suhdetta tarkastelemalla ei voida päätellä mitä mielentiloja tarkasteltavalla henkilöllä on. Relevanttien mielentilojen lista nimittäin voi olla mielivaltaisen laaja, ja jos tietyn reaktion on tarkoitus olla empiirinen kriteeri tietylle mielentilalle, ei testi paljasta mitään, koska reaktion toteutuminen tai toteutumatta jääminen voi kertoa joko testattavasta tai sitten jostain muusta mielentilasta. Näin ollen behaviorismin oletettu vahva yhteys empiriaan näyttäisi katkeavan. Behavioristi voi tietysti todeta saman ongelman myrkyttävän

myös intentionaalista psykologiaa, mutta esimerkiksi kognitivisteille tällainen ongelma on yksinkertaisesti empiirinen, siinä missä behavioristi on käsitteellisissä vaikeuksissa. Kognitivisti voi nimittäin vedota siihen, että ärsykkeen ja reaktion välistä suhdetta välittää monimutkainen järjestelmä, missä tosiaankin monet mielentilat tekevät työtään, ja ajoittain voi olla lähes mahdotonta huomioida niitä kaikkia. Behavioristi taas ei voi vedota refleksin monimutkaisuuteen, koska ärsykkeen ja reaktion suhteen selittäminen organismin sisäisillä prosesseilla tarkoittaisi behaviorismista luopumista.

Behavioristin ehkä olisi mahdollista vedota refleksien sarjaan ja todeta, että prosessin monimutkaisuus perustuu eräänlaiseen sisäiseen puheeseen, joka muokkaa organismin ärsykeympäristöä aiemmin kuvatulla tavalla, mutta tämäkin tuntuisi johtavan umpikujaan. Jos x tietää, että seuraavalla tapaamiskerralla paikalla onkin psykologin sijasta hänen identtinen kaksoissisarensa, joka toimii pankkivirkailijana ja jolla ei ole pienintäkään velvollisuutta tai aikomusta pitää x:n syvimpiä salaisuuksia omana tietonaan, voidaan olettaa, että x:n avoimuus on huomattavasti hillitympää kuin edellisellä tapaamisella vaikka ärsykeympäristö pysyy oleellisesti samana. Behavioristi voisi ehkä vedota siihen, että ulkoiset ärsykkeet kyllä ovat samoja, mutta tässä tilanteessa mukana on vielä sisäinen ärsyke, nimittäin äänetön lausahdus ”tuo ei ole oikea terapeutti.” Mutta mikä tämän reaktion laukaisee? Syy ei voi olla ulkoisissa ärsykkeissä, jotka oletuksen mukaan pysyvät vakioisina. Syyn täytyy olla x:n sisäisessä tilassa, eli siinä, että hän *tietää*, että terapeutti on huijari. Tällöin sisäinen lause ei kuitenkaan ole tietämisen konstituentti vaan seuraus, ja tällainen ei ole behaviorismia alkuunkaan.

On varmasti mahdollista koota sinänsä tosia ”se koira älähtää, johon kalikka kalahtaa” tyyppisiä ärsyke–reaktio-yleistyksiä, mutta yleisiä psykologisia lainalaisuuksia ei liene mahdollista muotoilla viittaamatta organismin ärsykeympäristöstä riippumattomiin mielentiloihin. Itse asiassa edes relevantteja ärsykkeitä ei voi kuvailla viittaamatta intentionaalisiin tiloihin, koska oleellista käyttäytymisen kannalta ei ole organismin fyysinen ympäristö, vaan se, miten organismi ympäröivän tilanteen tulkitsee uskomustensa ja tarpeidensa pohjalta.[14] Huomautettakoon, että tämänlaiset argumentit eivät syntyneet vastareaktiona behaviorismille, vaan ovat itse asiassa sitä vanhempia. Nimittäin samanlaisia tarkasteluja löytyy esimerkiksi William Jamesin *Principles of Psychology* -teoksesta vuodelta 1890. Hän väitti, ettei esimerkiksi sammakon käyttäytymistä voi ymmärtää yksinkertaisin mekaanisin käsittein, vaan käyttäytymisen erikoislaatuisuus piilee siinä, että sitä tulee tarkastella toiminnan tarkoituksenmukaisuuden perspektiivistä (James, 1890, s.7–10).

Behaviorismissa on siis ilmeisiä teoreettisia ongelmia, joiden ylittäminen vaikuttaa melko mahdottomalta. Ohjelman empiiriset näytöt jäivät myös lopulta melko vaatimattomiksi erityisesti korkeampien mielentoimintojen saralla. Psykologian leirissä metodologista behaviorismia ei yleisesti hylätty, mutta lukuisat 40- ja 50-lukujen vaihteessa syntyneet tutkimustulokset viittasivat siihen, että behavioraalisen todistusaineiston valossa on mahdollista – ja ennen kaikkea järkevää – muodostaa teorioita, jotka perustuvat jonkinlaiseen sisäiseen mentaaliseen koneistoon. Historiallisen kehityksen kannalta on merkittävää, että

[14] Tämänkaltaisesta behaviorismin kritiikistä tarkemmin ks. esim. (Dennett, 1978b), luku 2: ”Behaviorism and Mentalism” (Fodor, 1968a, s.47–89) ja (Pylyshyn, 1984, s.7–15).

näitä tuloksia syntyi nimenomaan Pohjois-Amerikassa, jonka psykologian laitoksilla behaviorismi nautti suurta suosiota. (Bechtel et al., 1998, s.17–24.) Lopullinen kuolinisku behaviorismille oli sen ilmeinen kyvyttömyys selittää kielellistä käyttäytymistä.

Kognitiotieteen alkutaipaleella hyvin keskeinen hahmo oli amerikkalainen kielitieteilijä Noam Chomsky. Hän kirjoitti pitkänpuoleisen kritiikin Skinnerin teoksesta *Verbal Behavior*, missä hän varsin vakuuttavasti argumentoi Skinnerin käyttävän epäsuorasti mentalistista terminologiaa selittämään kielellistä käyttäytymistä näin pettäen omat behavioristiset sitoumuksensa (Chomsky, 1959, s.30–36). Kognitiotieteen synnyn kannalta merkittävämpää kuitenkin oli Chomskyn argumentti, jonka mukaan ärsyke–reaktio-malli ei riitä kielen oppimisen eikä kielellisen käyttäytymisen selittämiseen. Kokonaisuudessaan argumentti oikeastaan levittäytyy useaan julkaisuun, ja se voidaan rekonstruoida perustuen seuraaviin avainaskeliin: Muodollinen kielioppi on abstrakti kuvaus kielen käyttäjän kyvystä ymmärtää ja tuottaa lauseita sekä erotella varsinaiset lauseet epäkieliopillisista ilmauksista. Luonnollisten kielten syntaksin tutkija koittaa siis muotoilla empiiristä teoriaa kieliopista, missä tutkimusaineisto perustuu siihen, mitä lauseita kielen puhujat pitävät kieliopillisina ja mitä eivät. Näin ollen kieliopin jokseenkin abstraktista luonteesta huolimatta se ei ole riippumaton kielen tosiasiallisesta käytöstä (Chomsky, 1959, s.56). Toiseksi Chomsky oli vuosia aiemmin kirjoittanut jokseenkin huonosti tunnetun artikkelin nimeltään "Three Models for the Description of Language", jossa hän osoitti, ettei äärellisellä markovin prosessilla voida tuottaa englannin kieliopin mukaista lausejoukkoa (Chomsky, 1956, s.116). Tämä tulos oli pohjana hänen kognitiotieteen synnyn kannalta erityisen merkittävälle teokselleen *Syntactic Structures* (1957). Tässä yhteydessä riittänee todeta, että äärellinen markovin prosessi on abstrakti määritelmä äärelliselle systeemille, joka tulostaa symbolijonoja käyttämättä muistia tai käsittelemättä sisäisiä representaatioita.[15] *Syntactic Structures* puolestaan keskittyi esittelemään teoriaa, jonka mukaan englannin kielen kielioppi voidaan palauttaa joukkoon peruslausemuotoja ja muuntosääntöjä, joiden avulla kaikki kieliopilliset lauseet voidaan tuottaa (Chomsky, 1957, s.106–107).

Mikäli kielioppi kuvaa aidon psykologisen kyvyn, joka välttämättä edellyttää merkkijonojen manipulointia mielessä tietynlaisten sääntöjen avulla, seuraa tästä, että kielellisen kyvyn taustalla täytyy olla jonkinlainen symbolirakenteita käsittelevä mekanismi. Lisäksi kielen oppimisen ilmeisesti täytyy tällöin olla symbolirakenteiden käsittelyn oppimista. Alla olevassa lainauksessa Chomsky esitteleekin juuri tällaiset päätelmät:

> It is not easy to accept the view that a child is capable of constructing an extremely complex mechanism for generating a set of sentences, some of which he has heard, or that an adult can instantaneously determine whether (and if so, how) a particular item is generated by this mechanism, which has many of the properties of an abstract deductive theory. Yet this appears to be a fair description of the performance of the speaker, listener, and learner. [...] The fact that all normal children acquire essentially comparable grammars of great

[15] Chomsky määritteli markovin prosessit oleellisesti äärellisinä automaatteina sillä erotuksella, että markovin prosessit tulostavat lauseita siinä missä tilakoneet lukevat niitä. Vrt. (Chomsky, 1956, s.114) ja (Hopcroft et al., 2001, s.46). Automaatteihin palataan tarkemmin myöhemmin.

> complexity with remarkable rapidity suggests that human beings are somehow specially designed to do this, with data-handling or "hypothesis-formulating" ability of unknown character and complexity. [...] In principle it may be possible to study the problem of determining what the built-in structure of an information-processing (hypothesis-forming) system must be to enable it to arrive at the grammar of a language from the available data in the available time. (Chomsky, 1959, s.57–58.)

Chomskyn vaikutus kognitiotieteen syntyyn oli erittäin merkittävä (Boden, 2006, s.590–591), ja hänen näkemyksensä kielellisestä kyvystä syntaktisten symbolirakenteiden tiedostamattomana käsittelynä kuvaa jotakuinkin täydellisesti komputationaalisen mielenteorian perusajatuksen. Tämä ei tarkoittanut paluuta vanhaan, sillä introspektio ei tullut enää kysymykseen. Niin chomskylainen kielitiede kuin kognitivismia pohjustaneet psykologiset tulokset, kuten esimerkiksi George A. Millerin työmuistia koskevat tutkimukset (Miller, 1956), olivat empiirisesti motivoituja spekulaatioita alitajuisista, siis introspektion tavoittamattomista, mielen rakenteista. Vaikka mentaaliset representaatiot ja sen sellaiset eivät ole käytännössä suoraan havaittavissa, niin komputationaalisen mielenteorian mukaan ne kuitenkin ovat kirjaimellisesti ja konkreettisesti maailmassa, joskin jossain kallon sisällä piilossa, joten mitään periaatteellista estettä niiden tutkimiselle ei pitäisi olla. Ilmiöiden suora havainnoitavuus ei sitä paitsi voi olla kynnyskysymys, koska likipitäen kaikki tieteelliset teoriat sisältävät teoreettisia ilmiöitä, joita ei voida havaita suoraan ja joiden olemassaolo on ainoastaan pääteltävissä. Itse asiassa juuri hiukkasfysiikka, jota sentään monissa piireissä pidetään todellisuuden perimmäisen rakenteen tutkimuksena, on tästä varmaankin parhaita esimerkkejä.

Mielenteorian perspektiivistä mielen tarkasteleminen representaatioiden käsittelyjärjestelmänä ei kuitenkaan ole ongelman ratkaisemista vaan määrittelemistä, mutta 50-luvulle tultaessa oli luotu hyvä pohja tämän näkemyksen naturalistiselle ymmärtämiselle. Komputationaalinen mielenteoria, siten kuin se filosofiassa tunnetaan, ei syntynyt psykologiasta eikä kielitieteistä, vaan siitä kun moderni teoreettinen matematiikka törmäsi hieman vanhempaan filosofiseen mielenteoriaan. Ei itse asiassa ole kovinkaan kummallista, että komputationaalinen teoria kumpusi tältä rintamalta. Symbolisen matematiikan kehitys 1800-luvun loppupuolelta eteenpäin nimittäin keskittyi paljolti päättelyiden formalisoinnin ja mekanisoinnin tutkimiseen, ja tässä prosessissa päättelemisestä tehtiin sen verran abstrakti käsite, että oli helppo tulkita saatujen tulosten koskevan representaatioiden käsittelyä yleisemmin, eikä pelkästään matemaattista järkeilyä ja ongelmanratkaisua. Tämä tutkimus synnytti myös laskennan teorian ja tietokoneet, joiden mekaanisen representaatioiden käsittelyn nähtiin edustavan uutta mielen mallia.

2.2 Matemaattisen logiikan synty

Modernin logiikan isänä pidetään yleisesti saksalaista Gottlob Fregeä (1848–1925). Yksittäisen teoksen julkaisua tuskin voi pitää minkään tieteenalan varsinaisena synnyinhetkenä, mutta symbolisen logiikan esiinmarssin voisi hyvin katsoa alkaneen Fregen *Begriffssc-*

riftin julkaisemisesta vuonna 1879. Teoksen tuomista monista uudistuksista nyt kolme on erityisen oleellisia: [16] 1. Muuttujien ja vakioiden erottelu loogisissa ilmauksissa, sekä loogisen funktion käsitteen käyttöönotto, 2. kvanttorin käsitteen ja kvantifikaatioteorian keksiminen, sekä 3. aksiomaattis-deduktiivinen esitys predikaattilogiikasta.

Kaksi ensiksi mainittua uudistusta tekevät logiikasta täsmällisen ja hyvin ilmaisuvoimaisen formaalikielen. Yksinkertaisimmillaan *looginen funktio* tarkoittaa oleellisesti samaa kuin käsite. Yleisesti sillä tarkoitetaan lausemuotoa, joka ilmaisee jonkin argumentille kuuluvan ominaisuuden tai argumenttien välillä vallitsevan suhteen, missä argumentit ovat jonkinlaisia olioita tai vastaavia loogisia subjekteja. Esimerkiksi lauseesta "Vety on kevyempää kuin hiilidioksidi" voidaan analysoida esiin looginen muoto, jossa lauseen subjekti "vety" on argumentti funktiolle "x on kevyempää kuin hiilidioksidi." Toisaalta samaisesta lauseesta voidaan muotoilla funktio, jossa argumenttina on sekä lauseen subjekti "vety" että lauseen kieliopillinen objekti "hiilidioksidi".[17] Modernilla notaatiolla lause voidaan formalisoida siis $P(a)$ tai $R(a, b)$; kun a="vety", b="hiilidioksidi", $P(x)$="x on kevyempää kuin hiilidioksidi" ja $R(x, y)$="x on kevyempää kuin y". Fregen oman määritelmän mukaan looginen funktio on se osa ilmaisua, joka pysyy muuttumattomana, jos ilmaisusta vaihdetaan jokin atominen tai monimutkaisempi symboli, ja muuttuva osa ilmaisusta on puolestaan argumentti (Frege, 1970, s.13). Kvanttorien käyttöönotto taas mahdollistaa universaalisen tai epämääräisen subjektin asettamisen muuttujien paikalle. Loogisten konnektiivien kanssa tämä tekee formaalikielestä hyvin ilmaisuvoimaisen. Näin esimerkiksi loogisten funktioiden avulla voidaan ilmaista esimerkiksi seuraavat propositiot:

1. $\exists x(A(x) \wedge R(a, x))$ "On olemassa vetyä raskaampi alkuaine."
   (epämääräinen subjekti: "*Eräs* alkuaine on vetyä raskaampi.")

2. $\forall x(A(x) \wedge x \neq a \rightarrow R(a, x))$ "Vety on kaikkein kevein alkuaine."
   (universaalisubjekti: "*Kaikki* muut alkuaineet ovat vetyä raskaampia.")

Missä a ja $R(x, y)$ ovat kuten edellä ja $A(x)$ ="x on alkuaine", $\alpha \wedge \beta$ ="α ja β" sekä $\alpha \rightarrow \beta$ ="jos α, niin β". Tarkalleen ottaen esimerkiksi kohta 2. luettaisiin "Jokaiselle alkuaineelle, joka ei ole vetyä, pätee, että vety on kyseistä alkuainetta kevyempää." Selvästi molemmat suomenkieliset ilmaukset ilmaisevat saman väitelauseen, joilla on yhteinen kohdan 2. kuvaama looginen muoto.

Tämänlaisen kaavakielen suurin merkitys onkin juuri siinä, että symbolinen kieli puhdistaa luonnollisen kielen enemmän tai vähemmän mielivaltaisen pintarakenteen, ja esittää kielellisen ilmauksen sisältämän *proposition*, eli käsitteellisen sisällön muodon. Näin tämänkaltaista formaalikieltä voidaan käyttää luonnollisen kielen jatkeena, mutta – mikä on jatkon kannalta tärkeämpää – tämänlainen formalismi voi periaatteessa kokonaan korvata luonnollisen kielen ainakin propositiolauseiden osalta. Frege itse ajatteli, että hänen logiikkansa esittää puhtaan ajattelun kielen (Kneale & Kneale, 1962, s.478), mikä on nähtävissä jo teoksen *Begriffsschrift* (suom. *Käsitekirjoitus*) nimessä.

[16] Ks. esim. (Vilkko, 2005, s.28–29).

[17] Kyseessä on Fregen oma esimerkki, (Frege, 1970, s.12–13).

Frege ei suinkaan ollut ensimmäinen, joka tuli ajatelleeksi tämänlaisen kielen kehittämistä. Wilhelm Ockhamilaisen *Summa logicae* (1323) lienee ensimmäinen tunnettu teos, jossa esitetään, että ajattelu on kirjaimellisesti kielellistä, mutta kyseessä ei ole ajattelijan luonnollinen kieli vaan sen taustalla oleva ja sen mahdollistava ajattelun kieli. Wilhelmin mukaan luonnollisen puhutun ja kirjoitetun kielen merkitys palautuu ajattelun kielen ilmaisuihin, jotka ovat käsitteitä ja merkityksellisiä itsessään, ja logiikka on eräänlaista mentaalikieltä koskevaa kielitiedettä. (King, 2005, s.244–248) Uudella ajalla Leibniz työskenteli tämän ajatuksen parissa, joskin hänen tavoitteenaan oli ennemminkin keinotekoisen universaalisen formaalikielen luominen (Kneale & Kneale, 1962, s.327–330). Tämän suunnitelman taas George Boolen voidaan katsoa jossain määrin jopa toteuttaneen parisen vuosikymmentä ennen Fregeä (*ibid.*, s.404–420). Mainitussa koplassa Boole eroaa kuitenkin melko puhtaana formalistina. Hänen tärkeä havaintonsa oli, että samankaltaisia algebrallisia periaatteita voidaan soveltaa sisällöllisesti täysin erilaisiin ongelmiin, jolloin symbolien merkityksillä ei ole väliä vaan ainoastaan formaaleilla rakenteilla. Jo Leibniz oli ymmärtänyt algebran ja logiikan yhteyden, mutta Boole osoitti miten syvä tämä yhteys on. Hän osoitti, että logiikan lait ovat luonteeltaan matemaattiset ja itse asiassa lähes samat kuin luvuille 0 ja 1 pätevät algebran lait. (von Wright, 1968, s.50–60.) Boolen logiikka ei kuitenkaan sisällä esimerkiksi kvantifikaatioteoriaa vaan rajoittuu luokkien kalkyyliin ja lauselogiikkaan, joten Fregelle lankeaa kunnia olla ensimmäinen, joka käytännössä muotoili tarpeeksi ilmaisuvoimaisen formalismin, jota ehkä vakavissaan voisi pitää jonkinlaisena universaalisena käsitekielenä. On tietenkin kovin kyseenalaista, voidaanko predikaattilogiikkaa pitää yleisenä ajatusten esitysjärjestelmänä, mutta vähintään uusi logiikka näytti tarjoavan mallin tämänlaisen kielen muodostamiselle.

Logiikassa keskeistä ei kuitenkaan ole vain syntaktinen yksiselitteisyys tai pyrkimys kielen universaalisuuteen. Ensisijaisesti logiikka on päättelyn tai järjenkäytön teoriaa, ja tällöin formaalikieli on lähinnä työkalun asemassa. Siinä, missä logiikan kieli formalisoi luonnollisen kielen lauseita, aksiomaattinen looginen systeemi puolestaan mahdollistaa päättelyjen formalisoimisen. Aksiomaattinen systeemi koostuu kahdesta osasta: *aksioomista* ja *päättelysäännöistä*. Usein aksioomien sanotaan olevan itsestään selviä peruslauseita, joita ei tarvitse eikä voi todistaa. Päättelysäännöt taas ovat periaatteita, joiden avulla aksioomista tai muista lauseista voidaan johtaa uusia lauseita. Aksioomista johdettuja lauseita kutsutaan *teoreemoiksi*. Aksioomat ja päättelysäännöt määrittelevät siis teoreemojen joukon, jotka kutsutaan *loogiseksi systeemiksi*.

Itse asiassa on hieman kyseenalaista väittää, että aksioomia ei voisi todistaa, eikä ole mitenkään välttämätöntä, että ne olisivat jotenkin itsestää nselviä. Aksioomia on ehkä mielekkäintä yksinkertaisesti pitää peruslauseita, jotka päättelysääntöjen ohella määrittelevät erään loogisen systeemin. Tässä mielessä systeemin aksioomat voivat olla mielivaltaisia, joskin lienee mielekästä vaatia, että ne vähintään ovat sekä sisäisesti että keskenään ristiriidattomia. Periaatteessa päättelysäännötkin voivat olla minkälaisia vain, mutta yleensä niiltä vaaditaan ainakin totuuden säilyttävyyttä, eli että sääntöjen avulla päätelty lause on aina tosi, kun päättelyssä käytetyt premissit ovat tosia. Päättelysääntöjä voidaan kuitenkin soveltaa mihin tahansa hyvin muodostettuihin formaalikielen lauseisiin niiden totuudesta riippumatta. Näin ollen aksiomaattista päättelysysteemiä voidaan

käyttää johtamaan johtopäätöksiä mistä tahansa mielekkäistä olettamuksista ja tosista tai epätosista väittämistä.

Aksiomaattisissa systeemeissä on erityisen oleellista, että päättelysäännöt määritellään pelkästään kaavojen muodon perusteella. Esimerkiksi konjunktion tuontisääntö sanoo, että mistä tahansa kaavoista α ja β voidaan päätellä kaava $\alpha \wedge \beta$, ja eliminointisääntö taas, että kaavasta $\alpha \wedge \beta$ voidaan päätellä kumpi tahansa kaava α tai β tai molemmat. Kaavojen merkityksillä ei päättelysääntöjen soveltamisen kannalta ole mitään merkitystä. Koko aksiomatisoinnin idea on, että systeemin avulla suoritettavissa päättelyissä ei missään vaiheessa tarvitse – eikä pidä – vedota intuitioon, terveeseen järkeen, merkityksiin tai mihinkään muuhun enemmän tai vähemmän epämääräiseen periaatteeseen, vaan päättelyt tehdään täysin yksiselitteisten syntaktisten sääntöjen avulla. Aksiomaattinen systeemi on eräänlainen *kalkyyli*, joka yleisesti ottaen tarkoittaa joukkoa sallittuja formaalikielen operaatioita. Näin järkeily, tai ainakin deduktiivinen päättely, voidaan palauttaa täysin formaaliseksi toiminnaksi, joka oleellisesti on eräänlaista symbolista laskemista. Luonnollisesti päättelysysteemin laatiminen, eli aksioomien ja päättelysääntöjen valinta, vaatii jonkinlaisia rationaalisuuden kriteereitä ja intuitioita siitä, minkälaiset päättelysäännöt säilyttävät totuuden ja mitkä eivät. Toiseksi päättelysysteemin olemassaolo ei poista järjenkäytön tarvetta. Säännöt kertovat ainoastaan mitä päättelyaskelia saa tehdä, ei mitä pitää tai kannattaa tehdä. Näin ollen päättelyaskelten tekeminen ei vaadi muuta perustetta kuin askeleen sallivan säännön olemassaolon, mutta oikeiden askelten valinta voi olla kohtuullisen haastava älyllinen tehtävä.

1800- ja 1900-lukujen taitteessa formaali logiikka kehittyi voimakkaasti. Myös Frege jatkoi työtään logiikan parissa, ja muun muassa kirjoitti joukko-oppia käsittelevän kaksi osaisen teoksen *Grundgesetze der Arithmetik*, jolla oli surullisen kuuluisa kohtalo. Bertrand Russell löysi ristiriidan Fregen joukkojen teoriasta kun teoksen ensimmäinen osa oli jo painettu. Itse asiassa joukko-opin isänä pidetty Georg Cantor oli havainnut teoriassa olevan ongelmia jo kymmenisen vuotta aikaisemmin, mutta joukkojen teorian ristiriitaisuus tuli yleisesti tunnetuksi vasta kun Frege julkaisi Russellin havainnon *Grundgesetzen* toisen osan jälkikirjoituksessa vuonna 1903. (Kneale & Kneale, 1962, s.652.) Joukko-opin ongelmat olivat sikäli erityisen kiusallisia, että tuona matemaattisen logiikan voimakkaan kehityksen ajanjaksona oli voimissaan *logisistinen ohjelma*, jonka pyrkimyksenä oli palauttaa kaikki matematiikka logiikkaan ja joukko-oppiin. Nimenomaan Russell oli tämän suuntauksen merkittävä kannattaja ja pyrki Alfred Whiteheadin kanssa muotoilemaan joukko-opista ristiriidattoman version kolmiosaisessa logisismin kulmakivessä *Principia Mathematica* (1910–1913). Tutkimus johti 1930-luvulle tultaessa kanoniseen joukko-opin aksiomatisointiin, joka tunnetaan nimellä ZF, eli Zermelo–Fraenkel-systeemi (Ferreirós, 1999, s.366–369). Joukko-opin aksiomatisoinnin motiivi oli puhdistaa teoria paradokseista, jotka olivat johtuneet *joukon* täsmällisen määritelmän puuttumisesta aiemmissa yritelmissä. Zermelo itse myönsi, ettei hän pysty todistamaan teoriaansa ristiriidattomaksi, mutta ainakin aksiomatisoitu joukko-oppi vältti tunnetut paradoksit. (Kneale & Kneale, 1962, s.681–683) Tässä on itse asiassa hyvä esimerkki aksiomatisoinnin merkityksestä matematiikalle. Liian intuitiivisesti, eli leväperäisesti, määritellyt systeemit ovat vaarassa kaatua ristiriitoihin. Kun valitaan huolella pieni joukko yksinkertaisia päättelysääntöjä tai konstruktioperiaatteita, voidaan valittu systeemi pitää hallinnassa.

Logisismin lisäksi 1900-luvun alussa matematiikan perustan tutkimuksessa vaikutti vahvasti *formalistinen* ohjelma, joka yleensä yhdistetään erityisesti David Hilbertin toimintaan ajanjaksolla 1900–1930. Logisismin kantava ajatus oli osoittaa klassisen matematiikan perustuvan logiikan varmalle perustalle. (von Wright, 1968, s.64–66) Hilbertiläisen formalismin pyrkimys taas oli abstrahoida matematiikasta kaikki sisältö pois ja suhtautua siihen eräänlaisena puhtaana merkkipelinä. Formalismi vaatii kaiken matematiikan aksiomatisointia, jolloin matemaattiseen päättelyyn ja todistamiseen ei periaatteessa tarvita minkäänlaista intuitiota eikä matemaattisten väittämien sisällöllistä tarkastelua (*ibid.*, s.85–86).Viime vuosisadan alun aksiomatisointi-innosta voikin kiittää suuressa määrin Hilbertiä, mutta hänen roolinsa tässä tarinassa perustuu pitkälti hänen kuuluisaan vuonna 1928 esittämäänsä ratkeavuusongelmaan, joka kulkee usein saksankielisellä nimellään *Entscheidungsproblem*. Ratkeavuusongelmassa on kyse siitä, onko mahdollista löytää algoritmi, joka ratkaisee jokaisen matemaattisen kysymyksen, eli onko olemassa mekaaninen, aina tuloksen antava menetelmä, jonka avulla saadaan vastaus ”tosi” tai ”epätosi” jokaisen matemaattisen väitteen kohdalla (*ibid.*, s.91). Hilbertin formalistinen ohjelma ja erityisesti ratkaisuongelma motivoi matemaatikkoja tutkimusohjelmiin, joiden seurauksena muotoutui nykyinen laskennan teoria, joka puolestaan loi pohjan teoreettiselle tietojenkäsittelytieteelle ja johti tietokoneiden syntyyn. 1900-luvun alun matematiikan perustoiden – erityisesti kalkyylien ja symbolisen laskennan – tutkimuksessa, merkittävissä määrin kysymys oli matemaattisen päättelyn mekanisoinnista, jonka eräänlainen kulminaatiopiste, ja toisaalta myös kuolinisku, oli universaalin algoritmin löytäminen ja täsmällinen määrittely vuonna 1936. Asiaan palataan pian, mutta tarkastellaan ensin 30-luvulla alkunsa saanutta formaaleihin systeemeihin liittyvä tutkimussuuntaa, jota nykyään kutsutaan *malliteoriaksi*.

Malliteoria on matematiikan haara, joka pyrkii muotoilemaan tulkinnan formaaleille kielille. Siinä, missä deduktiiviset tai komputationaaliset systeemit keskittyvät kielen sisäisiin operaatioihin ja kielen kaavojen välisiin deduktiivisiin suhteisiin, malliteoriassa tutkitaan formaalikielten ja joukko-opillisten struktuurien tai eri formaalikielten välisiä suhteita. Malliteoriasta on sittemmin tullut jokseenkin itsenäinen ja laaja matematiikan haara, mutta pohjimmiltaan se perustuu Alfred Tarskin työlle ja erityisesti hänen muotoilemalleen formaalikielten totuusmääritelmälle.

Vuonna 1933 Tarski julkaisi yli satasivuisen artikkelin ”Pojęcie prawdy w językach nauk dedukcyjnych” (suom. ”Totuuden käsite formalisoiduissa kielissä”), jossa tavoitteeksi asetettiin täsmällisen, klassisen korrespondenssikäsityksen mukaisen totuuskäsitteen määrittely formaaleille kielille. (Tarski, 1956a, s.152–154) *Korrespondenssikäsitys* tässä tarkoittaa lauseen vastaavuutta todellisuuden kanssa, eli yksinkertaisesti sitä, että lause α on tosi, jos α tarkoittaa, että asiat ovat niin ja näin, ja asiat todella ovat tällä tavalla. Tarski pyrki muotoilemaan menetelmän, jonka avulla voitaisiin määritellä formaalikielen L jokaiselle lauseelle α totuuspredikaatti $T(\alpha)$ =”α on tosi.” Käytännössä tämän tuli tapahtua antamalla totuusmääritelmä kieltä L rikkaammassa metakielessä siten, että kaikki seuraavan skeeman (T) intuitiivisesti todet instanssit voidaan johtaa metakielessä: $T(\alpha)$ *jos ja vain jos* $I(\alpha)$, missä $I(\alpha)$ on kaavan α käännös metakieleen. Usein filosofisesti suuntautuneemmissa teksteissä metakieli yksinkertaisesti määritellään olevan mikä tahansa kieli, jolla totuusehdot ilmaistaan, eli oikeastaan se jätetään määrittelemättä tai sen oletetaan

olevan luonnollinen kieli. Mutta Tarski määritteli metakielen olevan myös formaalikieli, joka sisältää kielen L, totuuspredikaatin ja joukko-opin. Lisäksi metakieli sisältää loogiset ilmaukset, kuten *ja*, *tai*, *jos ja vain jos*, ja niin edelleen sekä aksioomat, joiden avulla skeeman (T) instanssit voidaan johtaa. (Tarski, 1956a, 165–172.) Ongelma tässä on siis löytää menetelmä tuottaa ehtojen mukainen määritelmä tulkinnalle I. Filosofeille tutumpi teksti "The Semantic Conception of Truth" (Tarski, 1944) sisältää tiivistettynä vuoden 1933 artikkelin idean ilman teknistä matemaattista käsittelyä.

Nykyinen malliteoreettinen totuusmääritelmä löytyy selkeästi ilmaistuna kenties ensimmäisen kerran Tarskin ja Robert Vaughtin artikkelista "Arithmetical Extensions of Relational Systems" (1952, s.84–85), joskin malliteorian keskeisiä ideoita on hahmoteltu jo Tarskin artikkelissa "On the Concept of Logical Consequence" vuodelta 1936. Malliteoreettisen totuusmääritelmän idea on, että totuuspredikaattia ei pyritä määrittelemään suoraan, vaan antamalla induktiivinen määritelmä kaavan totuudelle suhteessa johonkin joukko-opilliseen struktuuriin. Mallin täsmällinen määritelmä riippuu tietenkin siitä formaalikielestä, jonka malleja ollaan rakentamassa, mutta yleisesti malliteoreettiset kielet koostuvat kolmenlaisista symboleista: loogisista vakioista, ei-loogisista vakioista ja muuttujista. Nykyään malli M määritellään yleensä *perusjoukon* X ja *tulkintafunktion* I parina. Perusjoukko on kokoelmasta mitä tahansa olioita, ja tulkintafunktio poimii formaalikielen ei-loogisten vakioiden viittauskohteet perusjoukosta.[18] Ei-loogisia vakioita ovat olioihin viittaavat yksilövakiot, ominaisuuksia puolestaan vastaavat joukot ja olioiden välisiä suhteita kaksi- tai useampipaikkaiset relaatiot. Funktio I siis oleellisesti määrittelee formaalikielissä olevien nimien ja käsitteiden tulkinnan. Jos siis esimerkiksi a on kielen yksilötermi ja P preditaatti, niin $I(a)$ poimii mallista jonkin olion ja $I(P)$ joukon. Loogiset vakiot, kuten konnektiivit "ja" sekä "tai", kvanttorit ja sen sellaiset, saavat vakioiset tulkinnat ja muuttujia ei tulkita.

Tarskilaisten totuusmääritelmien idea on suorastaan banaalin yksinkertainen. Ne sanovat suurinpiirtein jotain sellaista, että esimerkiksi lause $P(a)$ on totta, jos formaalikielen termiä a vastaava olio $I(a)$ löytyy ominaisuutta P vastaavasta joukosta $I(P)$. Luonnolliseen kieleen sovellettuna tämä käytännössä tarkoittaa, että esimerkiksi lause "lumi on valkoista" on totta, jos ja vain jos sanan "lumi" viittauskohde kuuluu niiden asioiden joukkoon, johon sana "valkoinen" viittaa. Siinäpä totuuden korrespondenssiteoria pähkinänkuoressa. Syy olla innoissaan tällaisesta innovaatiosta on monitahoinen. Ensinnäkin formaalikielen kaavoja ei tarvitse ajatella itsessään merkityksellisenä, mutta merkitystä tai viittaussuhdetta ei tarvitse ajatella mitenkään intuitiivisesti, vaan se voidaan täsmällisesti määritellä suhteessa malleihin. Lisäksi tästä perustasta käsin voidaan täsmällisesti määritellä muita mielenkiintoisia käsitteitä.

Ensinnäkin totuuden malliteoreettisesta määritelmästä saadaan helposti yleinen totuusmääritelmä seuraavalla tavalla: kielen L lause α on yleisesti, siis aina tai välttämättä, tosi, jos se on tosi kaikissa kielen L malleissa. Myös looginen seuraus voidaan määritellä mallien avulla: jos α on tosi jokaisessa mallissa, jossa kaavajoukko T on tosi, niin kaava α seuraa loogisesti kaavajoukosta T. Toiseksi edellä olevan menetelmän avulla voidaan mää-

[18] Tämä määritelmä on predikaattilogiikan kielelle, mutta käytännössä annettu määritelmä toimii, enemmän tai vähemmän merkittävin muutoksin, mille tahansa malliteoreettiselle kielelle.

ritellä täsmälleen, mitä tarkoitetaan teorian mallilla. Voidaan ajatella, että ainakin ideaalitapauksessa tieteellisten teorioiden käsitteet ovat niin hyvin määriteltyjä, että voidaan puhua esimerkiksi fysiikan kielestä, psykologian kielestä ja niin edelleen. Mikäli jonkin tieteenalan kieli on hyvin määritelty siinä mielessä, että se muodostaa jonkinlaisen malliteoreettisen kielen L, niin kyseisen tieteenalan teoriat voidaan määritellä lausejoukkona T, joka koostuu teorian mukaan tosista kielen L lauseista.[19] Tämä on pitkään ollutkin analyyttisen filosofian piirissä eräänlainen vakiokäsitys teorioiden luonteesta, mille johdannossa käsitellyn reduktiomallin idea pohjautuu. Tällöin M on teorian T malli, jos jokainen teorian lause on tosi mallissa M.

Edellä käsitettä *teoria* käytetään hieman laajemmassa merkityksessä kuin yleensä puhuttaessa tieteellisistä teorioista. Esimerkiksi Hempelin mukaan pelkkä mielivaltainen lausejoukko ei ole tieteellinen teoria, vaan teorian tieteellisyys edellyttää eräitä epistemologisia kriteereitä. Esimerkiksi, että teorian peruslauseista tulee voida johtaa ennusteita tiettyjen reunaehtojen vallitessa, lauseiden pitää olla empiirisesti testattavissa ja niin edelleen (Hempel, 1966, s.70–72). Jatkon kannalta teorian käsite on kuitenkin tarkoituksenmukaista määritellä varsin löyhästi. Kognitivismin perusajatuksiin kuuluu, että että esimerkiksi ihmisten käsitykset maailmasta ovat tämänlaisia – enemmän tai vähemmän kirjaimellisesti – päässä sijaitsevia lausejoukkoja. Tämä on tietysti suora seuraus oletuksesta, että mentaaliset representaatiot ovat lauseidenkaltaisia. Koska tällaiset lausejoukot voivat olla epistemologisesti miten tahansa muodostuneita, viitataan niihin yleensä tällä lailla löyhästi ja teknisesti määritellyllä teorian käsitteellä. Teorialla tässä mielessä voidaan tarkoittaa myös esimerkiksi tietokantaa, joka yksinkertaisesti luettelee joitain mitä tahansa asioita. Alla on esimerkki formaalikielisestä teoriasta, joka kertoo eräitä tosiasioita aurinkokunnasta:

Olkoon $a, b, c, d, e, f, g, P(x), S(x)$ ja $L(x, y)$ formalisointeja seuraaville ilmauksille:

a ="Aurinko" b ="Merkurius" c ="Venus" d ="Maa" e ="Mars"
f ="Aamutähti g ="Iltatähti"
$P(x)$ ="x on planeetta", $S(x)$ ="x on sisäplaneetta"
$L(x, y)$ ="x on lähempänä aurinkoa kuin y"

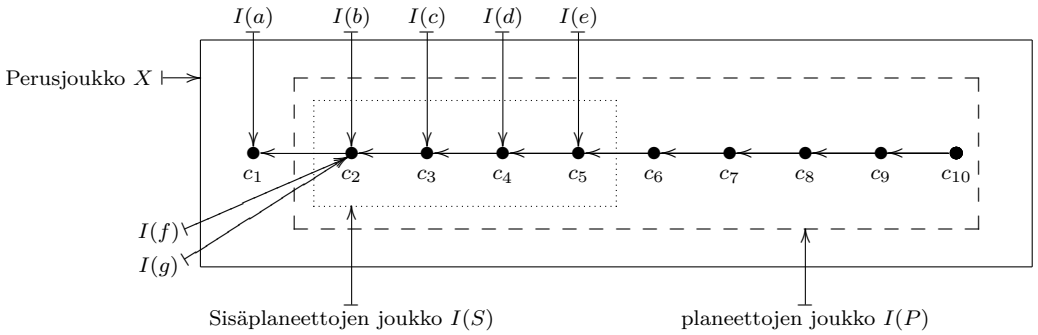
Olkoon sitten teoria T seuraava joukko lauseita, missä normaaliin tapaan $\wedge$="ja", $\vee$="tai", $\exists x$="On olemassa x siten, että ..." ja $\forall x$="Jokaiselle x pätee, että ..."

1. $\exists x_1, \ldots, x_{10} : \forall y : (y = x_1 \vee y = x_2 \vee, \ldots, \vee y = x_{10})$
2. $\neg P(a)\ \&\ \forall x : (x \neq a \rightarrow P(x))$
3. $f = g \wedge g = c$
4. $L(c, d) \wedge \neg\exists x : (x \neq c \wedge L(x, d) \wedge \neg L(x, c))$
5. $S(b) \wedge S(c) \wedge S(d) \wedge S(e)$

Teoria siis sanoo, että aurinkokunnassa on korkeintaan kymmenen kappaletta (lause 1), jotka ovat aurinkoa lukuun ottamatta planeettoja (lause 2). Lisäksi Iltatähti ja Aamutähti ovat sama taivaankappale, nimittäin planeetta Venus (lause 3), joka puolestaan on aurinkoa lähempänä oleva Maan naapuriplaneetta (lause 4). Lause taas 5 toteaa, että Merkurius, Venus, Maa ja Mars ovat sisäplaneettoja.

[19] Ks. esim. (Hempel, 1966) luvut 5. ja 6. (s.47–84), ja erityisesti Carnap (1956, s.46–76).

Alla on eräs teorian T malli, missä perusjoukko $X = \{c_1, \ldots, c_{10}\}$ ja tulkinta I näkyy kaaviosta. Alkiosta c_{10} alkioon c_1 kulkevat nuolet kuvaavat relaatiota $L(x, y)$.



Yleensä teoria sallii useita erilaisia malleja, eli ei ole olemassa vain yhtä mallia, jossa teoria on tosi. Näin on asian laita myös yllä olevassa tapauksessa. Teoria sallii muun muassa vähemmän planeettoja ja enemmän sisäplaneettoja sekä, että esimerkiksi ilmaukset ”Merkurius” ja ”Mars” viittaavat samaan planeettaan kuten ilmaukset ”Iltatähti” ja ”Aamutähti”. Myöskään relaatiota L ei ole juurikaan määritelty ja periaatteessa sen voi määritellä mallissa miten vain, kunhan se toteuttaa lauseen 3. Lisäämällä peruslauseita voidaan teoriaa tarkentaa siten, että se sallii vähemmän malleja. Mallien joukkoa voidaan rajoittaa myös teoriaan kajoamatta tarkastelemalla malleja, joissa sekä teoria T että joitain reunaehtoja määrittävä lausejoukko ovat molemmat tosia.

Jatkossa *formaali systeemi* viittaa systeemiin, joka koostuu formaalikielen L lisäksi kalkyylistä tai formaalista semantiikasta, joka määrittelee yksiselitteisesti L:n kaavojen tulkinnat tai totuusehdot. Nämä ovat kaksi tapaa lähestyä formaaleja systeemejä, eivätkä ne ole toisiaan poissulkevia. Logiikassa kalkyyli koostuu päättelysäännöistä ja aksioomista ja malliteoria määrittelee kaavojen tulkinnat. Itse asiassa logiikan tapauksessa on aivan samantekevää käytetäänkö systeemin määrittelemiseen malliteoriaa vaiko deduktiosysteemiä. Teorioiden loogiset seuraukset ovat samat molemmilla tavoilla tarkasteltuna. Vuonna 1930 Kurt Gödel todisti, että predikaattilogiikalle on täydellinen aksiomatisointi, eli että jokainen loogisesti tosi ensimmäisen kertaluvun lause voidaan johtaa äärellisestä joukosta aksioomia ja päättelysääntöjä (Kneale & Kneale, 1962, s.703). Tämä tarkoittaa, että jokainen pätevä päättely, joka voidaan formalisoida ensimmäisen kertaluvun predikaattilogiikan kielellä, voidaan suorittaa täysin syntaktis-deduktiivisesti. Tämä tulos on varmaankin metalogiikan ja formalistisen ohjelman tärkeimpiä yksittäisiä tuloksia ja nykyään se on tapana tulkita siten, että teoriasta T voidaan deduktiivisesti johtaa täsmälleen ne kaavat, jotka ovat tosia kaikissa teorian T malleissa.

Mutta palataanpa tarinassa vielä hieman taaksepäin. Kuinka kävikään formalistisen ohjelman ja itse Hilbertin? Heti predikaattilogiikan täydellisyystulosta seuraavana vuotena, eli 1931, Gödel julkisti vielä mullistavammat tuloksensa artikkelissaan ”Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I” (suom. ”*Principia mathematican* ja vastaavien systeemeiden formaalisesti ratkeamattomista lauseista I”). Artikkeli sisälsi Gödelin kuuluisat epätäydellisyyslauseet: 1. Ei ole mahdollista muotoilla äärellistä ja ristiriidatonta aksioomasysteemiä, jonka avulla voitai-

siin johtaa kaikki todet luonnollisia lukuja koskevat väitteet. 2. Mikä tahansa äärellinen teoria, joka on ilmaisuvoimaltaan tarpeeksi vahva ilmaisemaan oman ristiriidattomuutensa, sisältää teoreemanaan oman ristiriidattomuutensa toteavan lauseen jos ja vain jos se on ristiriitainen. (Nagel & Newman, 2001, s.92–94.)

Formalistinen ohjelma kaatuu siis ensimmäiseen epätäydellisyyslauseen, koska sen mukaan lukuteoriaa ei voi aksiomatisoida. Toinen epätäydellisyyslause taas tarkoittaa, että yksikään ristiriidaton formaali systeemi ei voi osoittaa omaa ristiriidattomuuttaan. Tietyin varauksin tämän voi katsoa tarkoittavan, ettei matematiikalla voi olla ristiriidattomaksi osoitettavaa perustaa. Nimittäin jos kaikki matematiikka palautettaisiin johonkin perusteoriaan, esimerkiksi logisistisen ohjelman mukaisesti joukko-oppiin, tämän perustan voisi todistaa ristiriidattomaksi vain suhteessa johonkin toiseen systeemiin. Mutta koska oletuksen perusteella tämä ensimmäinen systeemi on kaiken matematiikan perusta, on hieman hankalaa nähdä, mitä tämänlaisella ristiriidattomuustodistuksella saavutettaisiin. 1930-luvun aikana luvassa oli vielä lisää merkittäviä formaalien systeemien periaatteellisia rajoituksia koskevia tuloksia.

2.3 Laskennan malleista tietokoneisiin

Alonzo Church kehitteli 1930-luvulla Stephen Kleenen kanssa λ-kalkyylinä tunnetun formalismin, ja esitteli siihen sisältyvän λ-määritelteltävyyden käsitteen 1936 julkaistussa artikkelissa "An Unsolvable Problem of Elementary Number Theory". Tekstin tarkoitus oli otta kantaa Hilbertin ratkaisuongelmaan määrittelemällä mekaanisen laskettavuuden käsite mahdollisimman täsmällisesti, sillä ennen kuin ratkaisuongelma voidaan ratkaista, tulee laskettavuuden käsite jotenkin määritellä, jotta voidaan osoittaa mitä mekaanisen laskennan avulla voidaan periaatteessa ratkaista ja mitä ei. Kyseisessä artikkelissa esitettiin myös niin sanottu μ-*rekursiivisten* (jatkossa vain *rekursiivisten*) funktioiden joukko, ja osoitettiin, että λ-määriteltävät funktiot muodostavat täsmälleen saman luokan kuin rekursiiviset, jotka voidaan muodostaa neljästä hyvin yksinkertaisesta funktiosta:[20]

$Z(x) = 0$, kaikilla $x \in \mathbb{N}$
$S(x) = x + 1$
$U_i^n(x_1, \ldots, x_n) = x_i$
$\mu y(f(x, y) = 0) =$ pienin y jolle pätee $f(x, y) = 0$, jos $f(x, z)$ on määritelty $\forall z \leq y$;
muussa tapauksessa $\mu y(f(x, y) = 0)$ ei ole määritelty.

Lisäksi rekursiivisten funktioiden määritelmä sisältää kaksi, tässä sivuutettavaa, periaatetta, joiden avulla yllä olevista voidaan rakentaa uusia funktioita. Minimalisaatiofunktio μ ei ehkä avaudu ensimmäisellä vilkaisulla, mutta pienen perehtymisen jälkeen pitäisi olla selvää, että yllä olevat funktiot ovat laskettavia missä tahansa kiinnostavassa laskettavuuden mielessä.

[20] Artikkelissa annettu μ-rekursiivisten funktioiden määritelmä on itseasiassa Gödelin primitiivirekursiivisten funktioiden joukko, johon on lisätty Kleenen niin sanottu *minimalisaatiofuntkio*, toiselta nimeltään Kleenen μ-funktio, ks. (Church, 1936a, s.351 ja 353 alaviitteet 9 ja 12), joten oleellisesti Church vain veti yhteen olemassa olevia tuloksia. Alla oleva määritelmä on teoksesta (Cutland, 1980, s.25–43), josta löytyy myös rekursiivisten funktioiden muodostustamisperiaatteet.

Church argumentoi, että mikä tahansa algoritmi ja aksiomaattinen systeemi voidaan esittää rekursiivisten funktioiden avulla, ja johtopäätöksenä tästä hän esitti *Churchin teesinä* tunnetun väitteen: *Jokainen laskettava funktio on rekursiivinen.* (Church, 1936a, s.356–356.) Se, että rekursiiviset funktiot ovat määritelty luonnollisille luvuille, ei rajoita teesin yleisyyttä. Osa Churchin argumenttia on Gödelin vuoden 1931 artikkelista periytyvä menetelmä, jonka avulla mikä tahansa formaalikieli voidaan koodata luonnollisiksi luvuiksi niin sanotun Gödel-numeroinnin avulla ja jokainen formaalin systeemin operaatio esittää luonnollisten lukujen funktionta. (*ibid.*, s.349–359,) Toisin sanoen mikä tahansa algoritmi tai päättelysysteemi voidaan muodostaa lukujen ja numeeristen funktioiden avulla.

Churchin tähtäimessä oli siis ratkaista ratkeavuusongelma, mitä varten hän tarvitsi määritelmän laskettavuudelle, jonka rekursiiviset funktiot puolestaan tarjosivat. Tässä tehtävässään hän myös onnistui muotoilemalla λ-kalkyylin avulla erään ei-rekursiivisen funktion (Church, 1936a, s.361–363). Churchin teesistä tällöin seuraa, että vaikka kyseinen funktio on hyvin määritelty, niin yleistä menetelmää sen arvon laskemiseksi kaikilla mahdollisilla argumenteilla ei ole. Tulos koski lukuteoriaa, mutta vielä samana vuonna hän osoitti ratkeamattomuuden pätevän myös predikaattilogiikassa, eli vaikka predikaattilogiikalle on täydellinen aksiomatisointi, ei ole mahdollista laatia menetelmää, joka rakaisisi varmuudella mielivaltaisesta kaavasta, onko se systeemin teoreema vai ei (Church, 1936b).

Samana vuonna myös matemaatikko Alan Turing julkaisi oman tutkielmansa ratkeavuusongelmasta. Siinä missä Church pyrki määrittelemään laskettavan funktion käsitteen, oli Turingin mielessä universaali algoritmi, eli mekaaninen menetelmä minkä tahansa matemaattisesti hyvin määritellyn ongelman ratkaisemiseksi, vaikkakin suoranaisesti Turingin artikkeli "On Computable Numbers, with an Application to the Entscheidungsproblem" rajoittuu laskettavien lukujen käsittelyyn. Artikkelissa laskettava luku määritellään sellaiseksi, jonka desimaalikehitelmän binaariesitys voidaan tuottaa symboleita 0 ja 1 tulostavalla automaattisella koneella. Jokainen äärellinen sekvenssi, eli bittijono, voidaan tietenkin aina tuottaa koneellisesti, sillä mikä tahansa äärellinen symbolijono voidaan yksinkertaisesti kirjoittaa ohjelmaan ja laittaa kone tulostamaan sen. Sekvenssin laskettavuuden ongelma syntyy, kun desimaalikehitelmä on ääretön. Tällöin sitä ei voida koodata suoraan ohjelmaan, vaan tarvitaan jokin menetelmä sen muodostamiseksi.[21] (Turing, 1936, s.230–232.) Artikkelissa esitellyt koneet voidaan kuitenkin melko suoraviivaisesti yleistää minkä tahansa laskennan malliksi.

Kun siis jatkossa on puhetta laskettavista funktioista Turingin vuoden 1936 artikkelin yhteydessä, mainittakoon täsmällisyyden vuoksi, että Turing itse puhui laskettavista sekvensseistä. "On Computable Numbers" -artikkelin varsinaiset tulokset olivat, että kaikki sekvenssit eivät ole algortmisesti muodostettavissa ja, että predikaattilogiikka ei ole algoritmisesti ratkeava (Turing, 1936, s.236–263). Jälkimäinen tulos on siis oleellisesti sama, minkä Church oli osoittanut jo hieman aiemmin. Tulokset eroavat kuitenkin siten, että Turing muotoili todistuksensa suhteessa kehittelemiinsä abstrakteihin koneisiin, sii-

[21] Käytännössä luvun laskettavuus tarkoittaa, että kone voi tulostaa sen desimaalikehitelmästä mielivaltaisen mittaisen alkusegmentin. Äärettömän pitkää sybolijonoa ei tietenkään ole mahdollista tulostaa äärellisessä ajassa, joten jokainen varsinainen tuloste on luonnollisesti äärellisen mittainen. Kuitenkin jos kone jatkaisi äärettömän kauan, se periaatteessa lopulta tulostaisi koko desimaalikehitelmän.

nä missä Church ratkaisi ratkeavuusongelman rekursiivisten funktioiden avulla. Turingin artikkelin erityinen merkitys onkin juuri näiden abstraktien koneiden esittelemisessä.

Tarkastellaan konetta, joka koostuu nauhasta, lukulaitteesta ja prosessorista. Nauha on jaettu ruutuihin, jotka voivat sisältää jonkin symbolin rajoitetusta joukosta. Yleisyyttä rajoittamatta joukon voidaan olettaa olevan $\{0, 1\}$, mutta kyseessä voi olla mikä tahansa muukin äärellinen kokoelma symboleita. Prosessori puolestaan koostuu äärellisestä määrästä tiloja. Kone lukee nauhaa, jota syötetään lukulaitteeseen ruutu kerrallaan. Systeemi on suunniteltu siten, että aina symbolin luettuaan prosessori vaihtaa tilasta toiseen, joka määräytyy yksiselitteisesti edeltävän tilan ja koneen lukeman symbolin perusteella. Tätä jatketaan, kunnes koko syöte on luettu. Koneen vaste nauhalla olevaan syötteeseen on se tila, missä se on syötteen loputtua. Tämänlaista konetta kutsutaan *äärelliseksi automaatiksi*. (Hopcroft et al., 2001, s.46.)

Äärellinen automaatti koostuu 1. syöteaakkostosta Σ , joka sisältää nauhan mahdolliset symbolit, joilla syöte koodataan, 2. tilojen joukosta Q, joka sisältää erityisen aloitustilan $q_0 \in Q$ sekä hyväksymistilojen joukon $F \subseteq Q$, ja 3. siirtymäfunktiosta $\delta : Q \times \Sigma \to Q$; eli δ määrää koneen seuraavan tilan $q_i \in Q$ koneen edeltävän tilan $q_j \in Q$ ja luetun symbolin $a \in \Sigma$ muodostaman parin perusteella.

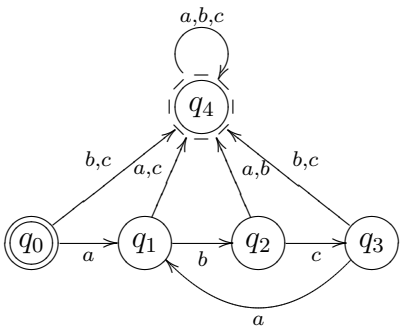
Seuraava äärellinen automaatti hyväksyy syötteenä sekvenssit, jotka voidaan muodostaa laittamalla peräkkäin symboleita a, b ja c, tässä järjestyksessä, eli jonot a, ab, abc, $abca$, $abcab$, $abcabc$, $abcabca$,... Funktio δ on esitetty alla selkeyden vuoksi taulukkona:

$$\Sigma = \{a, b, c\}$$
$$Q = \{q_0, q_1, q_2, q_3, q_4\}$$
$$F = \{q_1, q_2, q_3\}$$

$$\delta(q, x) =$$

	a	b	c
q_0	q_1	q_4	q_4
q_1	q_4	q_3	q_4
q_2	q_4	q_4	q_3
q_3	q_1	q_4	q_4
q_4	q_4	q_4	q_4

Tämänlaisten koneen rakenteesta ja toiminnasta saa paremmin selvää esittämällä se seuraavankaltaisena tiladiagrammina.



Kone tarvitsee syötteen a siirtyäkseen hyväksymistilaan q_1 alkutilasta q_0. Tilasta q_1 kone siirtyy hyväksymistilaan q_2 vain syötteellä b, ja muuten hylkäystilaan q_4, jossa se pysyy riippumatta siitä miten syöte jatkuu, ja niin edelleen. Kone siis pyörii hyväksymistilojen silmukassa, kun syöte on muotoa a, ab, abc, $abca$,... ja siirtyy hylkäystilaan q_4, jos jokin säännön vastainen symboli katkaisee tämän sekvenssin. Esimerkiksi syöte *abcab***a***cabcabc* johtaa hylkäystilaan lihavoidun a:n kohdalla.

Äärelliset automaatit eivät siis tulosta mitään, vaan niiden vaste on yksinkertaisesti se tila, johon kone päätyy syötteen luettuaan. Tyypillisesti tarkastellaan vain sitä, onko kone lopettaessaan hyväksymistilassa vaiko ei. Yleensä automaattien ajatellaan tunnistavan jonkin formaalikielen L. Tässä yhteydessä kieli määritellään siten, että aakkoston Σ kieli L on mikä tahansa kokoelma joukkoon Σ^* kuuluvia symbolijonoja, eli $L \subseteq \Sigma^*$, missä Σ on äärellinen aakkosto ja Σ^* sisältää kaikki jonot, jotka voidaan muodostaa joukon Σ symboleista, mukaan lukien tyhjä jono. Siis jos esimerkiksi $\Sigma = \{0, 1\}$, niin $\Sigma^* = \{\lambda, 0, 1, 00, 01, 10, 11, 001, 011, 100 \ldots\}$, missä λ on tyhjä merkki.[22] Joukon L alkioita kutsutaan *kielen L kaavoiksi. Automaatti tunnistaa* tai *määrittelee kielen L*, mikäli se päätyy jokaisella syötteellä α hyväksymistilaan jos ja vain jos α kuuluu kieleen L. Edelliseen lukuun (s.24) viitaten, Chomsky oleellisesti osoitti vuonna 1956, että tällaisella automaatilla ei voida erotella oikeinmuodostettuja englannin kielen lauseita epäkieliopillisista ilmauksista, eli ettei ole äärellistä automaattia, joka tunnistaisi englannin kielen.

Yleensä formaalikielet laaditaan jotain sellaista tarkoitusta varten, missä symboleilla on jokin tulkinta, kuten loogiset kielet, ohjelmointikielet, ja niin edelleen, mutta automaattien yhteydessä formaalikieli on usein hyvin abstrakti käsite. Aiemmin logiikan yhteydessä puhuttiin malliteoreettisista kielistä, mutta automaattien tapauksessa kielen mahdollisesta tulkinnasta ei välttämättä olla ollenkaan kiinnostuneita. Tällöin voidaan puhua myös *syntaktisista kielistä*, joita tyypillisesti käytetään nimenomaan automaattien tai abstraktien kielioppien ominaisuuksien tutkimiseen.

Toisaalta automaatteja ja syntaktisia kieliä voidan pitää myös normaaliin tapaan formaaleina systeemeinä: automaattia A voi pitää kielen Σ^* formaalina semantiikkana siten, että A jakaa kielen kaavat predikaatin $L(x)$ suhteen kahteen luokkaan: $L(\alpha)$ jos ja vain jos $\alpha \in L$, missä L on automaatin tunnistama formaalikieli. Lisäksi lopputiloja voidaan käyttää predikaatioina: $Q(\alpha)$ jos ja vain jos automaatti on tilassa q luettuaan syötteen α.

Äärellistä automaattia voidaan laajentaa siten, että se voi itse sekä liikutella nauhaa että tulostaa sille symboleja. Näin saadaan varsinainen *Turing-kone*. Nykyään Turing-koneet määritellään koostuvan seuraavista elementeistä (Hopcroft et al., 2001, s.318–326):

1. Äärellinen nauha-aakkosto Γ, joka sisältää kaikki symbolit joita nauha voi sisältää, ja erityisesti se sisältää ainakin
2. syöteaakkoston Σ, jolla on sama rooli kuin äärellisissä automaateissa, sekä
3. tyhjän merkin β, joka esiintyy ruuduissa, joissa ei ole muuta symbolia.
4. Tilojen joukko Q, joka sisältää
5. hyväksymistilojen joukon F ja
6. alkutilan q_0.
7. Siirtymäfunktio δ on äärellisten automaattien vastaavaa monimutkaisempi: $\delta : Q \times \Gamma \to Q \times \Gamma \times \{L, R\}$.

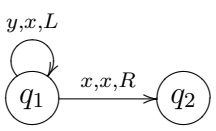
Nauha-aakkoston symbolit määrittelevät mitä nauhalla ylipäätään voi olla. Koneen varsinainen syöte taas koodataan syöteaakkoston avulla, kuten äärellisissä automaateissa, ja

[22] Tyhjällä merkillä on tiettyä teknistä merkitystä kieliä määriteltäessä ja se toimii siten, että jos α on symbolien jono, niin $\alpha\lambda = \alpha$. Formaalikielistä ja automaateista tarkemmin ks. esim. Hopcroft et al. (2001) luku 3. "Regular Expressions and Languages" s.83–121.

koneen tulosteesta luetaan ainoastaan syöteaakkoston merkkejä. Tyhjä merkki ei kuulu syöteaakkostoon, vaan se merkitsee nauhalla tyhjää ruutua. Tämä symboli tarvitaan, jotta koneen toiminta voidaan määritellä, mikäli se siirtyy nauhalla tyhjään kohtaan. Γ on siis ainakin β:n verran laajempi kuin Σ, mutta se voi sisältää muitakin merkkejä, joiden avulla kone voi tehdä nauhalle eräänlaisia muistiinpanoja. Ylimääräisiä nauha-aakkosia voidaan myös kirjoittaa syötteen joukkoon ohjaamaan koneen toimintaa.

Tilat ovat vastaavia kuin äärellisessä automaatissa, mutta siirtymäfunktio on selvästi monimutkaisempi. Argumentiksi siirtymäfunktio ottaa tilan ja nauhalla olevan symbolin muodostaman parin (q, γ), kuten äärellinen automaattikin, tässä vain symboli γ ei välttämättä ole syöteaakkonen. Mutta tämä pari ei määrää vain seuraavaa tilaa, vaan δ poimii nauhan ruutuun uuden symbolin tai pitää siinä jo olevan ennallaan, määrää seuraavan tilan, ja siirtää nauhaa vasemmalle (L) tai oikealle (R). Käytännössä Turing-kone siis toimii siten, että se lukee nauhalta symbolin ja mahdollisesti korvaa sen toisella. Tämän jälkeen kone vaihtaa tilaa ja siirtää nauhaa. Tätä jatketaan niin kauan kunnes kone pysähtyy, jonka jälkeen tuloste, eli nauhalla olevat syöteaakkoset, luetaan. Kone pysähtyy, kun siirtymäfunktiota ei ole määritelty, eli jos kone on tilassa q ja se lukee nauhalta symbolin γ, niin laskenta päättyy jos $\delta(q, \gamma)$ ei ole määritelty.

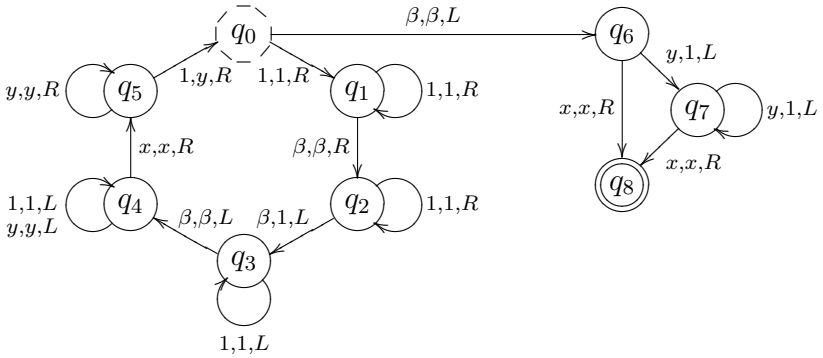
Tarkastellaan seuraavaksi esimerkin vuoksi Turing-konetta, joka laskee funktion $f(x) = 2x$, missä x on luonnollinen luku. Tässä syöteaakkosto koostuu pelkästään merkistä 1, jolla luonnolliset luvut koodataan yksinkertaisella tavalla: nauhalle laitetaan aluksi x:n mittainen jono ykkösiä, eli esimerkiksi luku 3 koodataan jonoksi 111. Koneen pysähdyttyä nauhalta luetaan ykkösten lukumäärä. Esimerkin helpottamiseksi oletetaan, että ykkösten ei tarvitse muodostaa jonoa. Nauhalta siis pitäisi koneen pysähdyttyä löytyä ykkösiä kaksinkertainen määrä kuin mitä oli alkuperäisessä syötejonossa. Lisäksi nauha-aakkostoon kuuluu merkki x, joka asetetaan syötejonon alkuun merkitsemään koneelle, että tämä on syötejonon vasemmanpuoleisin reuna, sekä merkki y, jota kone käyttää muistiinpanojen tekemiseen nauhalle. Nämä merkit toimivat siis koneelle ohjeina laskennan edetessä, ja näin siis, kuten Turing-koneissa yleensä, nauhaa käytetään sekä syötteen antamiseen ja tulosteen lukemiseen että työmuistina näiden tapahtumien välissä. Kone koostuu yhdeksästä tilasta, joista q_0 on alkutila ja q_8 hyväksymistila.[23] Alla on esimerkkikoneen tilasiirtymäkaavio, jota luetaan siten, että kun esimerkiksi tilasta q_2 tilaan q_3 menee nuoli, jonka vieressä lukee $\beta, 1, L$, niin kone vaihtaa tilasta q_2 tilaan q_3, kirjoittaa nauhalle merkin 1, ja lukupää siirtyy nauhalla vasemmalle, mikäli kone on tilassa q_2 ja lukee merkin β.



Eli tämänlainen kaavio vastaa ohjeita: $\delta(q_1, y) = (q_1, x, L)$ ja $\delta(q_1, x) = (q_2, x, R)$, eli "Tilassa q_1: jos nauhalla on y, kirjoita sen paikalle x, pysy samassa tilassa ja siirry nauhalla vasemmalle, jos taas nauhalla on x, älä muuta symbolia vaan vaihda tilaan q_2 ja siirry nauhalla oikealle.

[23] Itse asiassa kun Turing-konetta käytetään funktioiden laskemiseen, hyväksymistiloja ei tarvita, koska koneen vaste syötteeseen ei ole lopputila, vaan nauhan sisältö. Tämänlaisissakin koneissa lopputila voi olla kuitenkin kätevä esimerkiksi varmistamaan, että kone pysähtyy syötteen hyväksymisen ja loppuun suoritetun laskennan johdosta, eikä esimerkiksi väärin koodatun syötteen tai huonosti muodostetun siirtymäfunktion takia.

Tilasiirtymäkaavion alla on vielä esimerkki nauhan ja lukupään toiminnasta, kun koneelle syötetään luku 2, joka siis on koodattu nauhalle merkkijonona $x11\beta\beta\beta\ldots$



1. q_0 :	$x\ 1\ 1\ \beta\ \beta\ \beta\ \beta\ldots$	8. q_4 :	$x\ 1\ 1\ \beta\ 1\ \beta\ \beta\ldots$	15. q_3 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$	22. q_6 :	$x\ y\ y\ \beta\ 1\ 1\ \beta\ldots$
	$\triangle$		$\triangle$		$\triangle$		$\triangle$
2. q_1 :	$x\ 1\ 1\ \beta\ \beta\ \beta\ \beta\ldots$	9. q_5 :	$x\ 1\ 1\ \beta\ 1\ \beta\ \beta\ldots$	16. q_4 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$	23. q_7 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$
	$\triangle$		$\triangle$		$\triangle$		$\triangle$
3. q_1 :	$x\ 1\ 1\ \beta\ \beta\ \beta\ \beta\ldots$	10. q_0 :	$x\ y\ 1\ \beta\ 1\ \beta\ \beta\ldots$	17. q_4 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$	24. q_7 :	$x\ 1\ 1\ \beta\ 1\ 1\ \beta\ldots$
	$\triangle$		$\triangle$		$\triangle$		$\triangle$
4. q_2 :	$x\ 1\ 1\ \beta\ \beta\ \beta\ \beta\ldots$	11. q_1 :	$x\ y\ 1\ \beta\ 1\ \beta\ \beta\ldots$	18. q_4 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$	24. q_8 :	$x\ 1\ 1\ \beta\ 1\ 1\ \beta\ldots$
	$\triangle$		$\triangle$		$\triangle$		$\triangle$
5. q_3 :	$x\ 1\ 1\ \beta\ 1\ \beta\ \beta\ldots$	12. q_2 :	$x\ y\ 1\ \beta\ 1\ \beta\ \beta\ldots$	19. q_5 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$		
	$\triangle$		$\triangle$		$\triangle$		
6. q_4 :	$x\ 1\ 1\ \beta\ 1\ \beta\ \beta\ldots$	13. q_2 :	$x\ y\ 1\ \beta\ 1\ \beta\ \beta\ldots$	20. q_5 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$		
	$\triangle$		$\triangle$		$\triangle$		
7. q_4 :	$x\ 1\ 1\ \beta\ 1\ \beta\ \beta\ldots$	14. q_3 :	$x\ y\ 1\ \beta\ 1\ 1\ \beta\ldots$	21. q_0 :	$x\ y\ y\ \beta\ 1\ 1\ \beta\ldots$		
	$\triangle$		$\triangle$		$\triangle$		

Yllä oleva kone toimii siis siten, että jos nauhalla ei ole yhtään ykköstä, se siirtyy lopetusosaan (oikealla), siirtää lukupään nauhan alkuun ja jättää nauhalle tuloksen $x\beta\beta\ldots$ Tämä vastaa tilannetta, jossa syöte on 0, ja kone siis laskee $f(0) = 2 \times 0 = 0$, kuten pitääkin. Jos nauhalla on syötteenä ykkösiä, lukupää vaeltaa syötteen loppuun ja lisää ensimmäiseen tyhjän ruutuun symbolin 1. Tämän jälkeen kone palaa syötteen alkuun, jonka se tunnistaa symbolista x, ja merkitsee ensimmäisen ykkösen käsitellyksi tulostamalla sen päälle symbolin y. Kone jatkaa tätä silmukkaa, kunnes kaikki syötteenä olevat ykköset on käsitelty. Tämän jälkeen kone siirtyy lopetusohjelmaan, joka korvaa kaikki symbolit y symboleilla 1, palaa syötteen alkuun ja lopettaa.

On mahdollista laatia kone, joka voi suorittaa minkä tahansa Turing-koneen tekemän laskennan (Turing, 1936, s.243–246). Tämänlaista konetta kutsutaan *univeraaliseksi Turing-koneeksi* (UTM). Jokainen Turing-kone voidaan koodata UTM:n syöteaakkoston Σ avulla siten, että koneen M representaatio on jokin Σ-symbolien jono z. Tällöin $UTM(z, x) \downarrow y$ jos ja vain jos $M(x) \downarrow y$, missä merkintä $M(x) \downarrow y$ tarkoittaa, että Turing-kone M pysähtyy syötteellä x siten, että nauhalle jää tuloste y. Toisin sanoen kun UTM:lle syötetään koneen M koodaus, se tulostaa syötteellä x saman tulosteen kuin M.

”On Computable Numbers” -artikkelissaan Turing lisäksi esitteli argumentteja sen puolesta, että mikä tahansa mekaaninen laskenta on mahdollista suorittaa Turing-koneilla (*ibid.*, s.249–258). Koska Turing-koneet itsessään ovat laskennan malleja, ja toisaalta selvästi jokainen tällainen kone on eräänlainen algoritmi, tuli Turing määritelleeksi täsmällisesti

(universaalin) algoritmin käsitteen. Näin muodostui oleellisesti Churchin teesiä vastaava *Turingin teesi*: Olkoon f mikä tahansa osittainen tai totaalinen funktio. Tällöin f on laskettava, jos ja vain jos on olemassa Turing-kone M siten, että M laskee funktion arvon kaikilla argumenteilla, joilla f on määritelty.

Artikkelin lopusta löytyy vielä liite, jossa osoitetaan, että Turing-laskettavat funktiot muodostavat täsmälleen saman luokan kuin λ-määriteltävät, jotka siis edelleen muodostavat täsmälleen saman luokan kuin rekursiiviset funktiot (Turing, 1936, s.263–265). Tämä on itseasiassa Turingin esittämistä teesiään puolustavista argumenteista ehkä merkittävin, eli että Turingin teesi sanoo toisella tavalla muotoiltuna täsmälleen saman asian kuin Churchin teesi. Näin ollen nämä väitteet tuovat verrattain vahvaa intuitiivista tukea toisilleen. Church ja Turing julkaisivat tuloksensa lähes samaan aikaan ja päätyivät tuloksiinsa toisistaan riippumatta. Siksi tätä väitettä nykyään kutsutaan *Church–Turing-teesiksi*, joka yleensä ilmaistaan epämuodollisesti lähes samoin kuin Turingin teesi yllä:

Church–Turing-teesi (CT-teesi): Jokainen periaatteessa laskettava funktio voidaan laskea Turing-koneella.

Määritelläänpä tässä yhteydessä vielä kolme formaaleihin systeemeihin liittyvää käsitettä, joihin törmätään vielä myöhemmin:

Turing-täydellisyys: Formaali systeemi T on Turing-täydellinen, jos jokainen Turing-koneella laskettava funktio on ratkeava tai laskettavissa T:ssä.

Turing-ekvivalenssi: Formaali systeemi T on Turing-ekvivalentti, jos siinä ratkeavat funktiot muodostavat täsmälleen Turing-laskettavien funktioiden luokan.

Komputationaalinen universaalisuus: Formaali systeemi T on universaali suhteessa formaalien systeemien luokkaan $\mathcal{C}$, jos jokainen funktio, joka on ratkaistavissa jollain luokkaan $\mathcal{C}$ kuuluvalla formalismilla, on aina ratkaistavissa T:ssä.

Nyt CT-teesi voidaan ilmaista myös siten, että jokainen Turing-täydellinen systeemi on Turing-ekvivalentti, tai että Turing-arkkitehtuuri on universaali suhteessa kaikkiin mahdollisiin formaaleihin systeemeihin.

Church–Turing-teesistä seuraa, että jokainen formaalisti hyvin määritelty ongelma voidaan ratkaista Turing-koneella, olettaen että ongelma on ylipäätään ratkaistavissa. Olkoon T jokin formaali systeemi, joka yksikäsitteisesti määrittelee, päteekö jokin ominaisuus P teorian mielivaltaiselle kaavalle α tai termille x. Tällöin ongelma $P(\alpha)$?, eli *päteekö P kaavalle α?* – ja vastaavasti $P(x)$? – on hyvin määritelty. Huomaa, että jos P koskee kaavoja, se itse ei kuulu teoriaan T, vaan kyseessä on teorian lauseita koskeva metakielenpredikaatti. Teorian omat predikaatit taas koskevat sitä, mitä teoria nyt sattuu käsittelemään. Esimerkiksi jos T on predikaattilogiikka, jokaiselle predikaattilogiikan kaavalle predikaatti $P(\alpha) =$"α on teoreema" on hyvin määritelty. Vastaavasti hyvin määriteltyjä lukuteoreettisia predikaatteja on esimerkiksi $A(x) =$"x on alkuluku", $P(x) =$"x on parillinen" ja näiden avulla määritelty $G(x) =$"Jos $P(x)$ ja $x > 2$, niin $\exists y, z : A(y)$ ja $A(z)$ ja $z + y = x$". Goldbachin konjenktuurin mukaan $G(x)$ pätee kaikille luonnollisille luvuille.

Huomaa, että teorian ei tarvitse olla matematiikan alaan kuuluva. Periaatteessa jos joku aksiomatisoisi sosiologian, niin väite, että kapitalismi johtaa väistämättä luokkasotaan, voisi olla tässä teoriassa muotoiltavissa ja hyvin määritelty. Toisaalta hyvin määritelty ongelma ei aina vaadi teoriaa. Jos aion ratkaista montako kirjainta on jonossa *abc*, en taida tarvita erityistä teoriaa kertomaan mitä tämä kysymys tarkoittaa. Joka tapauksessa kysymystä voidaan pitää hyvin määriteltynä, jos voidaan ajatella, että on mielekkäästi muotoiltavissa formaali teoria, jossa kysymys on hyvin määriteltävissä.

Turing-koneiden kyky ratkaista hyvin määriteltyjä ongelmia perustuu siihen, että niitä voidaan käyttää hyväksymään ja hylkäämään syötteitä samalla tavalla kuin äärellisiä automaatteja. Tällöin koneen vastetta ei lueta nauhalta, vaan, samoin kuin automaattien tapauksessa, syöte katsotaan hyväksytyksi, jos kone pysähtyy hyväksymistilaan. Tällöin nauhaa käytetään syötteen antamiseen, mutta ei siis vasteen lukemiseen. Kone kuitenkin yleensä käyttää nauhaa muistilaitteena kirjoittamalla siihen merkkejä, joita se käyttää muistiinpanoina ohjaamaan syötteen käsittelyä. Tämä muistin käyttö mahdollistaa huomattavasti automaatteja laajemman kieltenluokittelukyvyn. Automaatit ovat käteviä, mutta niillä on pahoja rajoituksia. Äärellisillä automaateilla ei esimerkiksi voida erotella bittijonojen $\{0^n1^n \mid n \leq 1\} = \{01, 0011, 000111, \ldots\}$ joukkoa kaikista muista bittijonoista (Hopcroft et al., 2001, s.126–127). Turing-koneilla tämä ei ole isokaan ongelma. Jos *hyvin määritelty ongelma* määritellään kuten edellä, väite että jokainen hyvin määritelty ongelma on ratkeava jollain Turing-koneella, seuraa suoraan CT-teesistä: teorian T ongelma $P(\alpha)$? on ratkaistavissa, jos on olemassa Turing-kone, joka hyväksyy syötteen α jos ja vain jos $P(\alpha)$ pätee.[24] Tällöin sanotaan, että predikaatti P on ratkeava. Esimerkiksi predikaatti $P_{PL}(\alpha) =$"α on loogisesti tosi", missä α on lauselogiikan kaava, on hyvin määritelty ja ratkeava. Kysymys voidaan ratkaista mekaanisesti totuustaulumenetelmän avulla, ja CT-teesin perusteella tällöin jollain Turing-koneella. Toisaalta $P_{FOL}(\alpha) =$"α on loogisesti tosi", missä α on predikaattilogiikan kaava, on hyvin määritelty, mutta *ei* ratkeava. Tämä tulos on juurikin sekä Churchin "A Note on the Entscheidungsproblem" -artikkelin että Turingin "On Computable Numbersin" ydin.

Todettakoon, että Church–Turing-teesi ei ole matemaattisesti hyvin määritelty väite. Sitä ei siis voi todistaa missään matemaattisen todistamisen mielessä. Tämä johtuu siitä, että mekaanisuus ja laskettavuus ovat intuitiivisia ja epätäsmällisiä käsitteitä, joten niitä koskevia väitteitä ei voi todistaa, koska niitä koskevia väitteitä ei voi edes yksiselitteisesti formalisoida.[25] Tässä mielessä CT-teesiä voi pitää nimenomaan mekaanisen laskettavuuden matemaattisena määritelmänä, tai hieman heikommin tulkittuna tämän käsitteen teknisenä täsmentämisyrityksenä. Kuitenkin teesillä on selvästi hyvin vahva sisältö, ja sen kyllä voi osoittaa epätodeksi antamalla mikä tahansa selvästi mekaaninen tapa laskea jokin ei-rekursiivinen funktio tai esittämällä yleinen ratkaisumenetelmä hyvin määritellylle ongelmalle, joka ei kuitenkaan ole Turing-ratkeava. Toisekseen kolme erilaista formaalia systeemiä, λ-määriteltävyys, rekursiiviset funktiot ja Turing-koneet, määrittelevät täsmälleen saman laskettavien funktioiden luokan. Lisäksi Church ja Turing molemmat esittivät artikkeleissaan kohtuullisen vahvoja argumentteja vakuuttaakseen, että yleisempää las-

[24] Ks. esim. Cutland (1980, s.22–23,100–101).

[25] Tämän seikan sekä Church että Turing ymmärsivät (Church 1936a, s.356; Turing 1936, s.249).

kettavuuden määritelmää, ja siis laajempaa laskettavien funktioiden luokkaa, tuskin on olemassa (Church 1936a, s.356–358; Turing 1936, s.249–258). Näin jälkikäteen todettuna teesiä edelleen tukee se, että muitakin varsin erityyppisiä formaaleja systeemeitä on aikain saatossa kehitetty, ja kaikki Turing-täydelliset ovat osoittautuneet Turing-ekvivalenteiksi, kuten esimerkiksi Postin produktiosysteemit ja rekisterikoneet.[26]

On huomautettava, että vaikka Turing puhui koneesta, nauhasta, luku- ja kirjoituslaitteesta ja sen sellaisesta, Turing-koneet ovat kuitenkin abstrakteja laskennan malleja. Missään kohtaa vuoden 1936 artikkelia hän ei esitä, miten tämänlainen kone käytännössä voitaisiin rakentaa, eikä myöskään väitä, että universaali Turing-kone ylipäätään olisi rakennettavissa. Toisaalta Turing-koneiden hahmotelma antaa selvästi karkean suunnitelman konkreettisen koneen laatimiseksi. Ainoa epärealistinen oletus koneissa on äärettömän pitkä nauha. Tätä voi pitää joko välttämättömyytenä tai yksinkertaistuksena. Välttämättömänä siksi, että kone ei voi laskea yhtäkään totaalista funktiota $\mathbb{N} \to \mathbb{N}$, jos nauha on tiettyä äärellistä pituutta, koska tarpeeksi suuria lukuja ei voi mitenkään koodata äärelliselle nauhalle.[27] Kuitenkin jokainen yksittäinen pysähtyvä laskenta kestää vain äärellisen määrän askelia, joten kone käy vain äärellisessä määrässä ruutuja. Näin ollen mikä tahansa yksittäinen laskenta tarvitsee ainoastaan äärellistä nauhaa. Eli vaikka mikään äärellinen kone ei voi suorittaa kaikkia funktion $f : \mathbb{N} \to \mathbb{N}$ laskentoja, niin jokaista laskentaa kohti on äärellinen kone joka sen voi suorittaa. Ongelma on, ettei ole yleistä tapaa selvittää miten paljon laskenta vaatii nauhaa. Näin ollen matemaattisia tarkasteluja varten on merkittävä yksinkertaistus olettaa nauhan olevan ääretön. Tämä ei tee Turing-koneista kuitenkaan täysin epärealistisia konstruktioita. Voidaan nimittäin ajatella esimerkiksi, että koneen käyttäjä teippaa uuden rullan edellisen perään, mikäli nauha pääsee loppumaan.

Mekaaninen laskenta ei tietenkään ollut aivan tuorein keksintö edes 1930-luvulla. Periaatteessa helmitaulu on eräänlainen mekaaninen laskukone, joskaan ei sikäli automaattinen, että helmiä täytyy jonkun siirrellä ja vieläpä oikealla tavalla. Kuitenkin jo 1800-luvulla Charles Babbage suunnitteli ja osittain jopa rakensi universaalin laskentakoneen. Tämä keksintö kuitenkin vaikuttaa jostain syystä painuneen unohduksiin verrattain pitkäksi ajaksi. Esimerkiksi vaikka ensimmäisen universaalin elektromekaanisen laskukoneen pääinsinööri Howard Aiken katsoi nimenomaan vieneensä Babbagen projektin päätökseen ja tapasi ylistää Babbagen nerokkuutta aina kun tähän tarjoutui mahdollisuus, niin tosiasiassa hänkään ei ilmeisesti tuntenut kovin hyvin mitä Babbage itse asiassa oli tehnyt (Cohen, 1988). 1930–40-lukujen taitteessa oli käytössä useita elektromekaanisia laskukoneita, mutta Turingin hahmotelmista huolimatta vielä 40-luvun alussa ei ollut mitenkään selvää, että käyttökelpoista universaalia konetta voitaisiin rakentaa. Kehitys oli kuitenkin kohtuullisen nopeaa. Vuonna 1946 valmistui ensimmäinen Turing-täydellinen elektroninen kone ENIAC (Electronic Numerical Integrator And Computer) ja viisi vuotta myöhemmin nykyisenkaltainen ohjelmoitava tietokone EDVAC (Electronic Discrete Variable Automatic Computer) (Shurkin, 1996, s.166,342–343).

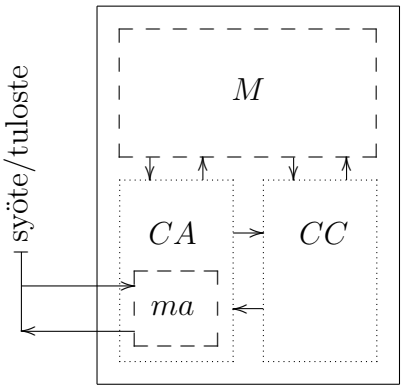
[26] Ks. (Cutland, 1980); rekisterikoneiden määritelmästä s.9–14, hahmotelma ekvivalenssitodistuksesta s.57, ja produktiosysteemeiden määritelmästä sekä Turing-ekvivalenssin todistuksesta s.59–64.

[27] Tarkemmin sanoen äärelliseen määrään ruutuja voidaan kirjoittaa vain äärellinen määrä äärellisestä aakkostosta koostuvia symbolijonoja. Näin ollen jos nauha on tietyn mittainen, on aina olemassa luku, jonka koodaus mahdu nauhalle varsinaisesti riippumatta kyseisen luvun suuruudesta.

EDVAC oli ensimmäinen kone, joka perustui niin sanottuun *von Neumann* -arkkitehtuuriin (vN), johon likipitäen kaikki nykyiset tietokoneet perustuvat (Ceruzzi, 2003, s.6,21). Von Neumann -arkkitehtuuri syntyi 1940-luvulla Pennsylvanian yliopiston *Moore Schoolissa* suunnitelmasta rakentaa toimiva, käytännössä universaali elektroninen laskija. Ongelma esimerkiksi ENIACissa oli, että vaikka kone oli periaatteessa universaalinen, käytännössä kuitenkin suoritettavan ohjelman, eli laskettavan funktion, vaihtaminen vaati koneen purkamista osiin ja sen fyysisen rakenteen muutamista. ENIACin ohjelman vaihtaminen saattoi viedä aikaa päivätolkulla (*ibid.*, s.20–21.).

Kuten aiemmin on todettu, universaali Turing-kone ei sisällä varsinaista ohjelmaa, vaan nauhalla on sekä jonkin Turing-koneen koodaus että koneen varsinainen syöte. Täten käsiteltävä data ja ohjelma eivät ole täysin erillisiä, vaan syöte koostuu aina molemmista. Von Neumann -arkkitehtuurin tuoma uudistus oli lisätä systeemiin muistilaite ja erotella muistissa oleva data ohjelmakoodista. Näin kone ensinnäkin voi varastoida syötteitä, mutta ennen kaikkea sisältää useita ohjelmia, eikä ohjelmien lisääminen vaadi koneen purkamista ja uudelleen kasaamista, vaan ainoastaan uuden ohjelman syöttämistä muistilaitteeseen. vN-arkkitehtuurin yleisrakenne löytyy John von Neumannin vuonna 1945 laatimasta raportista "First Draft of a Report on the EDVAC" (von Neumann, 1945, s.33–36). Myöhemmin rakennettu varsinainen EDVAC-kone poikkesi teknisiltä ratkaisuiltaan von Neumannin suunnitelmista, mutta yleinen vN-arkkitehtuuri on pysynyt jotakuinkin alkuperäisenä.[28]

vN-kone koostuu muistilaitteesta M, aritmeettisesta yksiköstä CA ja loogisesta kontrollista CC. Aritmeettinen yksikkö suorittaa varsinaisen laskennan, ja muistilaite pitää sisällään ohjelmakoodia sekä mahdollisesti dataa. Kontrollilaite puolestaan sisältää ohjelman, joka ohjaa koneen toimintaa. Kontrollilaitteen ohjelma on kiinteä ja sen tehtävä on huolehtia, että kone suorittaa sille annetut mitkä tahansa ohjeet. Käyttäjä on suoraan yhteydessä ainoastaan syöte- ja tulostelaitteeseen, joka on kytketty suoraan aritmeettiseen yksikköön sijoitetun akkumulaattoriin ma, joka puolestaan on eräänlainen aritmeettisen yksikön välimuisti.



Usein kognition filosofian piiriin kuuluvissa teksteissä puhutaan Turing-koneista kun käsitellään fysikaalisia symbolisysteemeitä tai ajattelun tietokonemalleja. Kuitenkin huomattavasti asiallisempi analogia on vN-koneet. Käytännössä CC ja CA muodostavat universaalin Turing-koneen, jossa CA lisäksi sisältää suoraan rautaan upotettuna aritmeettisia operaatioita suorittavan ohjelmiston.[29] Äkkiseltään katsottuna muistilaite lisää tähän pa-

[28] Von Neumannin erotti koneen loogisen rakenteen sen teknisestä toteutuksesta. Vaikka von Neumannin alkuperäinen raportti sisälsi teknisen suunitelman koneelle, jota ei koskaan rakennettu, tosiasiallisesti rakennettu EDVAC säilytti pitkälti alkuperäisen suunnitelman mukaisen yleisen rakenteen. (Godfrey & Hendry, 1993)

[29] Tarkemmin ottaen vN-koneet ovat itse asiassa rekisterikoneita eivätkä Turing-koneita; vrt. rekisterikoneiden määritelmä (Cutland, 1980, s.9–14) ja von Neumannin suunnittelema kontrolliohjelma (Knuth, 1970). Tämä on kuitenkin jatkossa sanotun kannalta lähinnä tekninen yksityiskohta. Turing-koneeseen

kettiin vain käytännöllisen tavan ohjelmoida kone suorittamaan erilaisia laskentoja. Käytännössä kuitenkin muistilaitteen lisääminen muuttaa koneen luonnetta merkittävästi.

Muistilaite mahdollistaa korkean tason ohjelmointikielten käyttämisen ja interaktiivisen toiminnan koneen kanssa, missä syötettä voidaan käsitellä ja lisätä laskennan aikana, mikä on edellytys useimmille tietokonesovelluksille, kuten käyttöjärjestelmille ja tekstinkäsittelyohjelmille. Korkeamman tason ohjelmointikielet taas vapauttavat ohjelmoijan pelkän konekielen käytöstä ja mahdollistavat koneen ohjelmoimisen periaatteessa millä tahansa formaalikielellä. Koneeseen voidaan tällöin tallentaa ohjelmia ja tietoa periaatteessa mitä tahansa tarkoituksenmukaista kieltä käyttäen. Muistilaite ja datan erottaminen ohjelmakoodista ovat siis käytännössä mahdollistaneet tietokoneiden käyttämisen äärimmäisen monipuolisilla tavoilla, vaikka syöte–tuloste-järjestelmästä erillinen muistilaite ei varsinaista laskentakapasiteettia lisääkään.

Edellä mainitut toiminnalliset uudistukset perustuvat *virtuaalikoneiden* toteuttamiseen vN-koneilla. Virtuaalikoneet ovat koneilla toteutettuja koneita, tai kuten yleensä sanotaan ”koneiden simulaatioita”. Tosiasiassa kuitenkin simulaatiosta puhuminen on usein harhaanjohtavaa. Virtuaalikoneena voidaan pitää mitä tahansa symboleja tai representaatioita käsittelevää systeemiä, joka koostuu ohjeista, jotka määräävät miten symbolirakenteiden manipulointi tapahtuu. Virtuaalikone on siis eräänlainen formaali systeemi, joka koostuu formaalikielestä ja *proseduraalisista ohjeista*.[30] Proseduraaliset ohjeet ovat ikään kuin päättelysääntöjä, jotka kuitenkin kuvaavat miten kone symbolirakenteita käyttää, vastoin kuin varsinaiset päättelysäännöt, jotka kuvaavat mitä symboleilla on sallittua tehdä. Proseduraaliset ohjeet ovat siis ennemminkin virtuaalikoneen toiminnan kausaalisia kuvauksia, siinä missä päättelysäännöt kuvaavat formaalikielen kaavojen välisiä loogisia suhteita. Esimerkiksi Turing-koneet ovat virtuaalikoneita mutta predikaattilogiikan päättelysysteemit eivät. Rekursiiviset funktiot ja ohjelmointikielet ovat ehkä jonkinlaisia virtuaalikonemaisia rajatapauksia.

Esimerkiksi ohjelmointikielten tulkit muodostavat erään mielenkiintoisten ja hyvin oleellisen virtuaalikoneiden luokan. Olkoon M vN-kone ja T mainitulla koneella suoritettava ohjelmointikielen L tulkki. Yleisesti ottaen ohjelmat ovat jonkinlaisia, yleensä tiettyä tarkoitusta varten laadittuja, virtuaalikoneita. Tulkki puolestaan on M:n konekielinen ohjelma, joka suorittaa kielellä L kirjoitettuja ohjelmia kääntämällä ne M:n konekielisten komentojen sarjaksi. Puhtaasti formaalista perspektiivistä katsottuna ei ole mitään hyötyä suorittaa Turing-täydellistä ohjelmointikieltä Turing-täydellisellä koneella M. Ohjelmointikielen käyttöönotto tulkin avulla ei millään tavalla lisää koneen laskentavoimaa, koska koneen Turing-täydellisyydestä seuraa CT-teesin perusteella, että tulkin ajama oh-

verrattuna rekisterikone on vain hieman erilainen ja perusformalismiltaan enemmän numeerisen laskennan kaltainen tapa muotoilla universaali symbolisen laskennan malli.

[30] Alunperin tietotekniikan yhteydessä virtuaalikoneen käsite viittasi juuri todellisten tietokonearkkitehtuurien ohjelmalliseen simulaatioon, ks. esim. (Goldberg, 1974) ja (Popek & Goldberg, 1974). Nykyisessä käytössä termi *virtuaalikone* ei välttämättä tarkoita tätä, joskin virtuaalikoneilla ajatellaan yleensä olevan samanlainen rooli kuin varsinaisilla koneilla, eli ne ovat alustoja esimerkiksi käyttöjärjestelmien suorittamiselle. Virtuaalikoneiden nykykäytöstä ks. esim. (Smith & Nair, 2005). Käytän tässä virtuaalikoneen käsitettä vielä yleisemmässä, kappaleessa määritellyssä merkityksessä, joka vastaa suunnilleen tapaa miten esimerkiksi John Haugeland sitä käyttää (Haugeland, 1981a, s.10–15).

jelma voitaisiin joka tapauksessa muodostaa käyttämällä koneen M kieltä. Oikeastaan tämä on selvää jo sen perusteella, että M itse asiassa suorittaa ohjelmaa tulkin välityksellä. Tulkki on siis teoriassa tarpeeton välikäsi. Käytännössä tilanne on kuitenkin toinen. Ohjelmointikielten ja -tekniikoiden kehitys on ollut tietotekniikan kannalta äärimmäisen merkittävää, koska ilman tulkkeja ja kääntäjiä, jotka muodostavat ohjelmakoodista suoraan koneella ajettavia ohjelmia, kaikki ohjelmat pitäisi kirjoittaa suoraan konekielellä, mikä yleensä on käytännössä likipitäen mahdotonta (Boden, 2006, s.779).

Tyypillisiä vN-koneiden perusoperaatioita ovat muistirekisterien lukeminen ja manipuloiminen esimerkiksi lisäämällä tai kopioimalla rekisteriin luvun toisesta.[31] Kone voidaan ohjelmoida suorittamaan huomattavan monimutkaisia toimintoja – siis mitä tahansa laskettavuuden rajoissa – ketjuttamalla näitä perusoperaatioita sopivasti. Kuitenkin kohtuullisen yksinkertaisetkin ohjelmat voivat konekielellä kirjoitettuna äärimmäisen monimutkaisia. Ohjelmointikielet helpottavat tätä ongelmaa määrittelemällä joukon uusia perusoperaatioita, joilla monimutkaisempia ohjelmia on helpompi kirjoittaa. Tulkin tehtävä on sitten tulkita ohjelmointikielen ilmaisut koneen käyttämän formaalikielen ohjeiden ketjuiksi. Lisäksi korkean tason ohjelmointikielet tyypillisesti tarjoavat yksinkertaisen tavan määritellä uusia perusoperaatioita ketjuttamalla ohjelmointikielen ilmaisuja. Käytännössä nämä ketjut ovat siis itse enemmän tai vähemmän yksinkertaisia aliohjelmia. Tällaisista aliohjelmista koostuvat kirjastot puolestaan mahdollistavat ohjelmointitekniikoiden kumuloitumisen, joka edelleen helpottaa ohjelmointityötä suunnattomasti.

Itse asiassa myös vN-koneet ovat eräänlaisia virtuaalikoneita. Vaikka tietokoneita on yleensä tapana ajatella konkreettisina laitteina, ne ovat toisaalta myös täysverisiä formaaleita systeemeitä. Kun tietokoneita tarkastellaan fyysisinä laitteina, ne koostuvat piisiruista, kondensaattoreista, ruuveista ja sen sellaisista, mutta tietokoneita voi tarkastella myös formaaleina systeemeinä, jolloin ne koostuvat konekieleksi kutsutusta formaalikielestä ja proseduraalisista ohjeista, jotka määrittävät mitä prosessori konekielen lauseilla tekee. Tämä on eräs seikka, joka tekee tietokoneista filosofisesti mielenkiintoisia: tietokoneet ovat formaaleita systeemeitä, joita ei tarvitse tulkita. Konekielellä on aina fyysinen representaatio esimerkiksi elektronivirtana johtimissa ja konekielen operaatiot toteutetaan fyysisesti koneen prosessorissa. Tietokoneet ovat siis rakennettu siten, että tietyt fysikaaliset tapahtumat vastaavat enemmän tai vähemmän yksi yhteen tiettyjä konekielen operaatioita. Näin ollen kone ei tarvitse erillistä komponenttia tulkitsemaan konekielen käskyjä toimintaohjeiksi, koska perusoperaatiot tapahtuvat koneessa fysikaalisina prosesseina fysikaalisten kausaalilakien määräämällä tavalla. Toisin sanoen vN-koneen ei tarvitse tulkita käyttämäänsä kieltä sen enempää kuin höyrykoneen tarvitsee miettiä mitä se höyryllä tekisi.

Luonnollisesti virtuaalikoneilla voidaan suorittaa toisia. Tätähän tulkit juuri tekevät. Virtuaalikoneita on siis mahdollista ikään kuin pinota siten, että virtuaalikone M_1 implementoi koneen M_2, joka puolestaan voi implementoida koneen M_3 ja niin edelleen periaatteessa rajattomasti. Fysikaalisesti implementoitua formaalia systeemiä M_1, joka koostuu konekielestä ja tietokoneen alkeisoperaatioista, kutsun jatkossa ensimmäisen kertaluvun virtuaalikoneeksi, konekielellä M_1 toteutettua virtuaalikonetta M_2 toisen kertaluvun virtu-

[31] Ks. esim. (Knuth, 1970).

aalikoneeksi ja niin edelleen. Ainakin periaatteessa korkeamman kertaluvun virtuaalikone voitaisiin aina implementoida fyysisesti ja palauttaa se ensimmäisen kertaluvun koneeksi. Tämän tyyppisiä ratkaisuja on jonkin verran toteutettukin.[32] Toisaalta CT-teesin perusteella M_1 voitaisiin implementoida millä tahansa Turing-täydellisellä formalismilla. Virtuaalikoneen kertaluku ei siis ole formaaleihin systeemeihin mitenkään olemuksellisesti liittyvä ominaisuus, mutta se on mielekäs ja hyödyllinen käsite tarkasteltaessa konkreettisia informaatiota käsitteleviä systeemejä, joilla on hierarkkinen organisaatio. Yleensä kognitivistisen teorian mukaan mieli on juurikin tällainen systeemi, josta lisää seuraavassa luvussa.

Virtuaalikoneet voivat olla varsinaisten fyysisten koneiden simulaatioita, mutta komputationaalisen mielenteorian kannalta mielenkiintoisempaa on tarkastella virtuaalikoneita, jotka ovat jonkinlaisen representaatiosysteemin implementaatioita. *Implementaatio* tarkoittaa oleellisesti toteutusta, mutta toteutuksen ei tarvitse olla fyysinen. Tietokoneen prosessori on vN-koneen loogisen kontrollin ja aritmeettisen yksikön fyysinen implementaatio ja kovalevy vastaavasti muistilaitteen fyysinen implementaatio. Esimerkiksi Lispillä kirjoitettu tietokanta taas on abstraktin tietorakenteen ei-fyysinen implementaatio, joka voidaan implementoidaan fysikaalisesti muistilaitteessa. Tästä voi edelleen jatkaa abstraktiosta ylöspäin: shakkiohjelman kanssa pelattu peli on shakkipelin implementaatio, sillä pelin kannalta oleellista on noudattaa sääntöjä, ei liikutella fyysisiä kappaleita laudalla. Kaksi ihmistä voi tietenkin periaatteessa pelata shakkia ilman mitään fyysisiä kappaleita, jos pelaajat kertovat toisilleen tekemänsä siirrot ja pitävät mielessään laudan tilanteen ja siinä tapahtuvat muutokset. Tämäkin on eräänlainen shakkipelin implementaatio. Oleellista on kiinnittää huomiota siihen, mitä virtuaalikoneilla voidaan implementoida ja mitä ei. Periaatteessa tietokoneilla voidaan simuloida lähes mitä tahansa, mutta implementaation kannalta välttämätön ehto on, että mallinnettava systeemi on olemukseltaan symbolinen tai muuten abstrakti. Esimerkiksi sään simulaatio ei tietenkään ole implementaatio, olipa kyseessä miten tarkka kuvaus tahansa. Sateen simulaatio kun ei kastele mitään.

Menemättä tässä vaiheessa sen syvällisemmin implementaation, simulaation ja virtuaalikoneiden metafysiikkaan, muistilaitteen käyttöönotto mahdollistaa myös erään ilmeisen mutta mielenkiintoisen seikan, nimittäin teorioiden ja mallien upottamisen koneisiin. Mallit voidaan syöttää koneen muistiin asiaan soveltuvaa formaalikieltä käyttäen, ja vastaavasti teorian lait voidaan kapseloida koneeseen tietorakenteena tai ohjelmana. Tällöin laskukone muuttuu varsinaiseksi *tieto*koneeksi, ainakin tiedon käsitettä hieman vapaamielisesti käyttäen. Näin kone voi sisältää teorioita, malleja ja muita asiaintilojen representaatioita. Sopivasti ohjelmoituna se voi myös käyttää sisältämäänsä tietoa mielekkäällä tavalla. Koneelle voitaisiin antaa esimerkiksi jokin malli ja teoria sekä ohjelmoida se ratkaisemaan niitä koskevia ongelmia, esimerkiksi onko teoria mallissa tosi, pätevätkö niissä sellaiset ja tällaiset väitteet ja niin edelleen. Näin ollen on kovin harhaanjohtavaa suhtautua koneisiin pelkästään bittijonoja muuntelevina laskukoneina. Ennen kaikkea tietokoneet ovat konkreettisia, representaatioita käsitteleviä kausaalisia systeemeitä.

[32] Ks. esim. (McGhan & O'Connor, 1998).

2.4 Tietokoneet ja ajattelun implementointi

Vuoden 1950 elokuussa julkaistiin Alan Turingin ehkä kuuluisin artikkeli ”Computing Machinery and Intelligence”, jossa esitetään *Turingin testinä* tunnettu ajatuskoe. Turingin mukaan kysymykseen *voiko kone ajatella?* ei voida antaa mielekästä vastausta, koska ei tunnu olevan mitään yleisesti hyväksyttyä käsitystä siitä, mitä ajattelu on, joten on epäselvää mitä koko kysymys ylipäätään tarkoittaa. Turing esitti tämän kysymyksen korvaamista käsitteellisesti yksiselitteisellä kokeella: Laitetaan henkilö A kuulustelemaan tekstiterminaalin välityksellä henkilöitä B ja C, joista toinen on nainen ja toinen mies. Kuulustelijan tehtävä on selvittää B:n ja C:n sukupuoli esittämällä kysymyksiä. Kuulustelija saa kysyä mitä haluaa, joten yksinkertaisinta on tietenkin kysyä asiaa haastateltavilta suoraan. Oletetaan esimerkin vuoksi, että B on mies ja C on nainen. Kuulustelijan ongelma on siinä, että B huijaa. Hänen tehtävänsä on siis saada kuulustelija luulemaan B:tä naiseksi. Henkilö C taas koittaa auttaa kuulustelijaa. Tilanne on siis sellainen, että kuulustelija keskustelee tekstiterminaalin välityksellä kahden henkilön kanssa, joista molemmat väittävät olevansa naisia, ja kuulustelijan tehtävä on selvittää kumpi valehtelee. Turing esitti, että alkuperäinen kysymys ”Voiko kone ajatella?”, voidaan korvata seuraavalla kysymyksellä: Mitä tapahtuu kun kone laitetaan pelaajan B tilalle? Voidaanko kone ohjelmoida siten, että kuulustelija arvioi yhtä usein oikein ja väärin kun imitaatiopeli järjestetään tällä tavalla? (Turing, 1950, s.433–434.)

Yleensä Turingin testi esitetään yksinkertaisemmin siten, että kuulustelija keskustelee koneen ja ihmisen kanssa, ja yrittää selvittää kumpi heistä on kumpi. Jos kuulustelija ei pysty luotettavasti erottamaan konetta ihmisestä, meidän tulisi pitää konetta ajattelevana. Testi on järjestetty siis siten, että kuulustelija ei näe eikä kuule kuulusteltavia, joten tunnistus on tehtävä pelkästään kielellisen ulosannin perusteella. Turing ei itse juurikaan käsittele alkuperäisen kysymyksen *voiko kone ajatella?* ja uuden kysymyksen *voiko kone pärjätä hyvin imitaatiopelissä?* välistä eroa. Itse asiassa hän toteaa, että alkuperäinen kysymys on niin epäselvä, ettei se vaadi tarkempaa käsittelyä (*ibid.*, s.442), mutta käyttää toisaalta suurimman osan artikkelistaan perustelemaan uuden kysymyksen mielekkyyttä alkuperäisen kysymyksen korvaajana. Vaikkei siis Turing suoraan tätä sanokaan, uusi kysymys on oleellisesti alkuperäisen kysymyksen operationalisoitu muoto. Mikäli tällainen kone onnistutaan rakentamaan, niin hänen mukaansa koneajattelukysymykseen saadaan yleinen myöntävä vastaus. Näin ollen Turing mitä ilmeisimmin on sitä mieltä, että ajattelun simulaatio on itseasiassa ajattelun implementaatio, ja onnistuneen implementaation kriteeri on, että koneen toimintaa ei voi erottaa ihmisen kielellisestä käyttäytymisestä.

Missään kohtaa vuoden 1950 artikkelia Turing ei väitä, että ihmiset olisivat digitaalisia koneita, tai tarkemmin sanoen, että ihmisen kognitiivinen koneisto olisi jonkinlainen vN-kone. Kuitenkin hän nähtävästi katsoo, että mikäli voidaan osoittaa, ettei ihmismieltä voida kuvata jonkinlaisena automaattina, tämä riittää näyttämään, etteivät koneet voi ajatella. (*ibid.*, s.452–453.) Vaikkei hän tätä sanokaan, niin edellinen seuraa UTM:n universaalisuudesta ja implementaation ehdoista, eli siitä, että mielivaltainen systeemi voidaan implementoida Turing-koneella jos ja vain jos se on jonkinlainen representaatioita käsittelevä systeemi. Näin ollen ajatteleva kone on mahdollinen jos ja vain jos ihminen

– tai tarkemmin siis ihmismieli – on symboleja käsittelevä automaatti. Tässä kuitenkin testin empiirinen luonne hieman unohtuu, sillä tarpeeksi monimutkainen kone voisi yksinkertaisesti hämätä ihmistä, vaikkei se kykenisi ajattelemaan inhimillisellä tavalla. Ilmeisesti Turingilla on kuitenkin mielessään, että jos ihminen on jossain mielessä konetta ylivertaisempi ajattelun saralla, niin tietyllä tavalla kuulustelemalla tämä näennäisen koneajattelun oletettu puute voitaisiin saada selville. Itse asiassa Turing ei uskonut, että mieli on kirjaimellisesti vN-kone – tai ylipäätään minkäänlainen diskreetteihin tiloihin perustuva kone – sillä perusteella, että aivotilat eivät ole diskreettejä, ja tällaiset jatkuviin tiloihin perustuvat koneet eivät ole vN-koneita (*ibid.*, s.451 ja 455). Hän kuitenkin uskoi, että testin läpäisevä kone on mahdollista rakentaa mahdollisesti jo seuraavan 50 vuoden sisällä, eli nyt vuoden 2010 perspektiivistä edellisen vajaan 60 vuoden aikana.

Nyt lukija saattaa pitää Turingin kantoja hieman ristiriitaisina. Millä perusteella kone on ajattelun implementaatio eikä simulaatio, jos kerran tietokoneet tosiasiassa toimivat laadullisesti eri tavalla kuin aivot? Turingin artikkeli ei ikävä kyllä ole käsitteellisesti aivan niin tarkka kuin saattaisi toivoa, mutta on syytä pitää mielessä, että Turingin testin kannalta oleellista on se, mitä kone tekee, ei miten se sen tekee.

Nyt kasassa alkaa olla kaikki palaset joista mielen käsittäminen eräänlaisena tietokoneena kumpuaa. Ensinnäkin taustalla on hyvin selvästi representationaalinen mielenteoria. Kielellisen käyttäytymisen tärkeys mentaalisuuden merkkinä johtuu siitä, että kieli on formaatti jolla mentaaliset representaatiot voidaan ilmaista. Representationaalisen mielenteorian keskeinen ajatushan on, että mielen sisällöt, eli mentaaliset representaatiot, ovat propositioita. Logiikka puolestaan tarjoaa syntaksin propositioille ja teorioille sekä kalkyylin päättelyille. Lisäksi predikaattilogiikan täydellisyystulos takaa, että kaikki ensimmäisen kertaluvun päättelyt on mahdollista suorittaa käyttämällä verrattain yksinkertaista formaalia systeemiä. Turing-täydelliset koneet voivat puolestaan implementoida kaikki mahdolliset formaalit systeemit, joten koneet voivat periaatteessa suorittaa mitä tahansa päättelyitä. Koneisiin voidaan ohjelmoimalla upottaa teorioita ja malleja, ja lisäksi ne voivat vastata mihin tahansa periaatteessa ratkeavaan kysymykseen, joista niillä on tarpeeksi tietoa. Täydellisyystulokseen sisältyvä luotettavuustulos myös takaa, että kone voidaan periaatteessa ohjelmoida päättelemään siten, että sen suorittama järkeily noudattaa kielen semanttisia periaatteita, eli käytännössä lauseiden totuusehtoja. Nyt jos ajattelu rinnastetaan päättelemiseen, ongelmanratkontaan, vastausten löytämiseen ja vastaavanlaiseen propositionaaliseen tietojenkäsittelyyn, niin tietokoneet pystyvät periaatteessa ruumiillistamaan kaiken tämän.

Ei myöskään ole syytä vähätellä varsinaisen fyysisten tietokoneiden heuristista arvoa komputationaaliselle mielenteorialle. Esimerkiksi Turing-koneesta ei erityisen voimakkaasti tule mieleen ihminen tai muu kognitiivinen agentti. Von Neumann -koneen muistilaite mahdollisti tietokoneiden teoreettisen kapasiteetin valjastamisen täysimittaiseen käyttöön, mutta vN-koneiden rakenteessa on myös muutama mielen mallin kannalta silmiinpistävä piirre. Muistin ja syöte–tuloste-laitteiston erottelu tuo mieleen aistin- ja motorisen järjestelmän erottelemisen kognitiivisesta järjestelmästä. Syötelaitteelle annetun ja muistissa sijaitsevan datan ero puolestaan muistuttaa jossain määrin ärsykkeiden ja omaksutun tiedon välistä eroa. Vastaavasti datan ja ohjelmakoodin erottaminen koneen muistissa

tuo mieleen tiedon ja kognitiivisten kykyjen välisen eron. Tämänlaiset enemmän tai vähemmän löyhät samankaltaisuudet eivät tietenkään ole hyviä syitä pitää koneita ajattelevina olioina, mutta koneet on helpompi mieltää kognitiivisina agentteina, kun ne kovasti näyttävät sellaisilta. Von Neumannin "First Draft" sattuukin olemaan hyvä esimerkki tietokoneiden antropomorfisoinnista. Raportissaan hän esimerkiksi kutsuu elimiksi (*organs*) järjestelmiä, jotka huolehtivat muisti- ja kontrollilaitteiden kommunikaatiosta syöte- ja tulostelaitteiston kanssa, ja vertaa koneiden releitä ja elektroniputkia neuroneihin (von Neumann, 1945, s.3–6).

Turing oli myös hyvinkin tietoinen von Neumannin koneenrakennussuunitelmista ja viittaakin vuonna 1946 raportissaan "Proposed Electronical Calculator", joka on Turingin oma muistilaitteellisen koneen suunnitelma, suoraan von Neumannin "First Draft" raporttiin: "The present report gives a fairly complete account of the proposed calculator. It is recommended however that it be read in conjunction with J. von Neumann's 'Report on the EDVAC'." (Turing, 1946, s.3.). Ajattelun rinnastaminen symboliseen laskentaan on komputationaalisen mielenteorian reduktionistinen perusidea, jonka on tarkoitus selittää mitä ajatteleminen on. Varsinaisten koneiden rooli tässä on osoittaa, että fyysinen, kohtuullisen yksinkertainen automaatti kykenee suorittamaan kaiken mahdollisen laskennan, tai yleisemmin symbolien tai representaatioiden käsittelyn, ja voi konkreettisesti sisältää teorioita ja tietoa. Tässä mielessä koneet toimivat eräänlaisena olemassaolotodistuksena representationaalisen mielenteorian ja metafyysisen fysikalismin yhteensopivuudesta.

Vaikkei Turing käyttänytkään perusteenaan tämänlaisiin tuloksiin vetoavaa argumenttia, kaikki yllä esitetyt näkökohdat ovat varmasti olleet hänen tiedossaan. Ylipäätään formaalien tieteiden kehityksen seuraaminen modernin logiikan synnystä EDVACiin tekee ajatuksen koneellisesta ajattelusta hyvin ilmeiseksi, vaikka ensinäkemältä se saattaa monista tuntua varsin luonnottomalta. Voinee turvallisesti väittää, että ellei Turing olisi kirjoittanut artikkeliaan koneajattelusta vuonna 1950, lähivuosina joku muu olisi varmasti tuottanut vastaavanlaisen tekstin. Tätä taustaa vasten ei myöskään ole kovin kummallista, että vakavasti otettava kirjoitelma koneajattelusta tuli nimenomaan matemaatikolta. Tietenkään syy ehdotuksen ottamiseen vakavasti ei ollut yksistään Turingin artikkelin ansioissa vaan siinä, että mainitut tulokset olivat laajalti tieteentekijöiden ja filosofien tiedossa. Näin ollen Turingin ajatuksen ilmeinen järkevyys oli kaikkien nähtävissä. Lisäksi on syytä mainita, että seitsemän vuotta aikaisemmin, eli vuonna 1943, neurotieteilijät Warren McCulloch ja Walter Pitts julkaisivat artikkelin "A logical calculus of the ideas immanent in nervous activity", jossa pyritään osoittamaan, että aivotoimintaa voi tulkita loogisena kalkyylina, ja väitetään, että mielen toiminta voidaan selittää heidän teoriallaan (McCulloch & Pitts, 1943, s.38–39). Huomion arvoista tässä on, etteivät he väittäneet mielen toiminnan selittyvän aivotoiminnalla sinänsä, mikä on varsin yksinkertainen materialistinen hypoteesi, vaan sillä, että aivot implementoivat jonkinlaisen loogisen systeemin.

Itse asiassa on hieman kummallista, ettei Turing kommentoi mainittua työtä millään tavalla. McCullochin ja Pittsin artikkeli nimittäin vaikutti esimerkiksi von Neumanniin ja hän viittaakin siihen "First Draft" -raportissaan (von Neumann, 1945, s.37), jonka Turing kyllä tunsi hyvin. Toisaalta McCulloch ja Pitts koittavat perustella hieman eri asiaa kuin Turing. He pyrkivät osoittamaan, että mieltä voi pitää aivojen implementoimana

komputationaalisena systeeminä, kun taas Turing pyrki osoittamaan, että sopivanlaista komputationaalista systeemiä voi pitää mielenä. Vaikuttaa hieman siltä, että McCulloch ja Pitts pitivät jo melko selvänä, että sopivanlaisen komputationaalisen systeemin toteuttaminen riittää mielen implementoimiseksi, ja he pyrkivät osoittamaan miten aivot tämän tekevät. Yleensä Turingia pidetään komputationaalisen mielenteorian arkkitehtina, mutta tämä titteli ehkä kuuluisi kuitenkin McCullochille ja Pittsille. He sentään muodostivat varsinaisen empiirisesti motivoidun matemaattisen teorian mielen toiminnasta, siinä missä Turing toimi lähinnä jonkinlaisena ideologina (Piccinini, 2004, s.175–177). Olkoonkin niin, ettei heidän teoriansa ole koskaan herättänyt kovin laajaa mielenkiintoa.

Materialistinen ja mekanistinen käsitys ihmisestä ja mielestä ei tietenkään ollut mitenkään mullistava näkemys 60 vuotta sitten, vaan ajatusta on kyllä viljelty pitkin filosofian historiaa. Lyhin ja ehkä tyypillisin materialistinen argumentti kulkee jotakuinkin niin, että maailmankaikkeudessa kaikki koostuu luonnonlakien alamaisena toimivasta materiasta ja ihminen on osa tätä luonnonjärjestystä, siispä myös ihminen on eräänlainen materiaalinen kone. Tämänlaista ajattelua löytää jo vaikkapa Demokritokselta yli kahdentuhannen vuoden takaa (Barnes, 2001, s.217–222). Myös kaikenlaisia materialistisia konemetaforia on esitelty ainakin uudella ajalla. Nämä ovat muotoa, että aivot, joten siis myös mieli, on vain monimutkainen jousista, hammasrattaista tai väkipyöristä koostuva kone. Kuuluisia esimerkkejä tästä esiinyy muun muassa teoksissa La Mettrien *Ihmiskone* (1747) ja Hobbesin *Leviathan* (1651), jonka esipuheesta löytyy tällainen materialistinen ihmiskuva, ja lisäksi osan ”I. Of Man” luvusta ”V. Of Reason, and Science” yllättävän paljon komputationaalista mielenteoriaa muistuttava käsitys mielen toiminnasta (Hobbes, 1651, s.34–42). Kuitenkin kun Hobbes, kauan ennen tietokoneiden syntyä, sanoi järkeilyn olevan eräänlaista laskentaa, oli tämä lähinnä kuvaus mielen toiminnasta, ei sen selitys.

Pelkästään sen sanominen, että ihminen on materiaalinen kone ja ajattelu mekaanista, ei auta ymmärtämään ihmistä tai ajattelua millään tavalla, vaan yleensä on enemmänkin itsestään selvä ja melko mielenkiinnoton seuraus metafyysisestä materialismista, jonka puitteissa on selvää, että mieltä verrataan jonkinlaiseen koneeseen. On kuitenkin syytä painottaa, että mielen vertaaminen tietokoneeseen ei ole täysin tässä hengessä muotoiltu materialistinen oppi. Tosiasialliset komputationaaliset systeemit ovat tietenkin materiaalisia koneita, mutta tällaisten koneiden implementoimat virtuaalikoneet puolestaan eivät. Tietysti virtuaalikonekuvaus tietokoneesta on vain eräänlainen abstrakti perspektiivi fyysiseen laitteeseen, mutta on kuitenkin syytä pitää mielessä, että siinä missä fyysisen koneen kuvaus koostuu esimerkiksi siitä, miten sähkövirta kulkee virtapiireissä, mistä materiaalista muistipiirit koostuvat ja niin edelleen, virtuaalikoneen kuvaus koostuu koneen tiloista, muistin ja syötteen informaatiosisällöstä sekä säännöistä, jotka kuvaavat miten kone manipuloi symboleja. Luonnollisesti virtuaalikone edellyttää fysikaalisen implementaation ollakseen mentaalinen systeemi. Esimerkiksi paperille laadittu tietokone ei varsinaisesti käsittele symboleita tai ylipäätään tee yhtään mitään, joten se ei varsinaisesti muodosta mitään kausaalista systeemiä. Elektroninen tietokone on ehkä vain hienostunut versio jousista ja hammasrattaista rakennetusta mekanismista, mutta fysikaalisesti implementoitu virtuaalikone on ennenkaikkea kausaalinen symboleita käsittelevä systeemi, jonka fysikaalinen koostumus on periaatteessa samantekevää. Jos virtuaalikoneen käsittele-

mät symbolit yhdistetään mentaalisiin representaatioihin ja itse käsittely ajatteluun, mieli rinnastetaan tällöin nimenomaan *virtuaali*koneeseen.

Esimerkiksi Turingille tämä erottelu tuntui olevan hieman epäselvä, koska hän totesi, ettei mieli ole digitaalinen kone koska aivot eivät ole (Turing, 1950, s.455). Kuitenkin hän itsekin huomautti samassa artikkelissaan, että todelliset fyysiset laitteet ovat vastaavasti epädigitaalisia, tai tarkemmin sanottuna epädiskreettejä (*ibid.*, s.439). Koneiden fysikaaliset tilat ovat tosiasiassa jatkuvia, kuten on asiainlaita esimerkiksi valokatkaisimen tapauksessa: toiminnallisesti tarkasteltuna katkaisimella on kaksi tilaa, *päällä* tai *kiinni*, mutta todellinen fyysinen vipu tai nappi liikkuu jatkuvaisluonteisesti näiden tilojen välillä. Vastaavasti siitä, että aivot eivät ole diskreetteihin tiloihin perustuva kone, ei seuraa ettei aivojen mahdollisesti implementoima virtuaalikone voisi olla luonteeltaan digitaalinen. Ylipäätään tietokonemetaforan ydin on, että ajattelu redusoituu virtuaalikoneisiin, joten mielen vertaaminen fyysiseen koneeseen ei ole vain epäoleellista vaan kategoriavirhe.

Näin siis ajattelevan koneen idea nousee jokseenkin luontevasti kun laskettavuutta ja formaalikieliä koskevat teoriat kohtaavat ohjelmoitavat tietokoneet, olettaen että taustalla ajattelusta on representationaalinen mielenteoria. On ainakin ajateltavissa, että toiminnaltaan ihmismieltä vastaava kone voidaan valmistaa, ja esimerkiksi Turing uskoi, että 1900-luvun loppuun mennessä ainakin melkoisen hyvä yritys testin läpäiseväksi koneeksi kyetään rakentamaan, tai oikeastaan ohjelmoimaan. Nimittäin vaikka testissä hyvin suoriutuvalta koneelta vaaditaan oletettavasti sekä aikalailla laskentatehoa että erityisesti suurta muistikapasiteettia, digitaalikoneen universaalisuuden takia tällaisen koneen tekeminen on kuitenkin pohjimmiltaan ohjelmointiongelma. Turing itse esitti, että paras strategia koneen ohjelmoimiseksi on kumulatiivinen: koneen sisältämää ohjelmaa ja tietomäärää lisätään sekä muutellaan siten, että koneen suoriutuminen testissä hiljalleen paranee, eli koneen toiminta lähenee ihmisen psykologista käyttäytymistä. Tähän hän tarjosi kahta lähestymistapaa. Joko voidaan lähteä liikkeelle verrattain abstrakteista älyllisistä haasteista, kuten shakin pelaamisesta, ja ohjelmoida kone aluksi suoriutumaan tämänlaisista tehtävistä, tai sitten voidaan koittaa asentaa koneeseen mahdollisimman hyvä aistinjärjestelmä ja ohjelmisto, jonka avulla sitä voidaan ryhtyä opettamaan samaan tapaan kuin vaikkapa lapsia. Ajatus kummassakin lähestymistavassa on, että kun ohjelmisto kasvaa tarpeeksi monimutkaiseksi, kone ei enää käyttäydy ilmiselvän automaatin tapaan, vaan se alkaa muodostamaan yksinkertaisista tiedonpalasista monimutkaisempia ideoita ja teorioita (Turing, 1950, s.454–560). Optimismistaan huolimatta Turing totesi, että koneajattelukysymys loppujen lopuksi ratkeaa vain empiirisesti ja aika näyttää onnistutaanko tällainen laite koskaan laatimaan. ”Computing Machinery and Intelligence” oli siis ennen kaikkea ohjelmanjulistus eikä suunnitelma ajattelevan koneen laatimiseksi.

Ajattelevan koneen mahdollisuus on monella tavalla mielenkiintoinen idea. Mikäli tällainen kone on periaatteessa rakennettavissa, tarkoittaa tämä muun muassa, ettei ajattelu edellytä aivoja muttei toisaalta myöskään erillistä epämateriaalista substansissa tai mystistä epämekaanista kykyä. Ei kuitenkaan ole selvää, mitä seurauksia tästä muuten on mentaalisuuden ymmärtämiselle. Ehkä laskennan avulla voidaan toteuttaa mentaalinen systeemi, mutta tästä ei ainakaan suoraan seuraa, että esimerkiksi meidän mentaalisuutemme olisi laskentaa tai perustuisi siihen. Toiseksi auttaako komputationalismi

ymmärtämään meitä tuntevina ja haluavina olioina, vai rajoittuuko teorian mahdollinen selitysvoima vain johonkin mentaalisen toiminnan siivuun, kuten päättelyyn tai muuhun niin sanotusti kylmän rationaaliseen toimintaan? Perehdytään seuraavaksi funktionalismiin, joka on paljolti komputationalismista ponnistava teoria, jonka on tarkoitus selittää reduktiivisesti mentalistiset käsitteet.

3 Funktionalismi ja mielentilojen luonne

Tässä luvussa käsitellään pääsääntöisesti funktionalismia, joka on keskeisin komputationalismiin ja kognitiotieteisiin liittyvä teoria mielentiloista. Funktionalismi viittaa oikeastaan kokoelmaan varsin erityyppisiä teorioita, joita yhdistää käsitys, että psykologisten tilojen identiteetti tai olemus määräytyy niiden kausaalisen aseman perusteella tarkasteltavan organismin kokonaispsykologiassa. Sinänsä funktionalistiset mielenteoriat eivät ole välttämättä komputationalistisia, mutta viimeisen noin puolen vuosisadan ajan merkittävin työ tieteellisen mielenteorian parissa on keskittynyt muotoilemaan komputationaalisin termein funktionalistinen teoria mielestä, tai ainakin lähtenyt siitä oletuksesta, että tyydyttävä tällainen teoria on mahdollista muotoilla. Yritän myös selventää mikä komputationalismin ja funktionalismin suhde oikeastaan on, ja miksi ne pohjimmiltaan ovat eri teoriat, vaikka usein ne osin historiallisista ja osin sisällöllisistä syistä sekoitetaan toisiinsa.

Eräs funktionalismiin liittyvä filosofisesti mielenkiintoinen piirre on sen tarjoama mahdollisuus ratkaista kysymys olioiden mentaalisuudesta, tekoälysysteemit mukaan lukien. *Heikko tekoälyteesi* on väite, jonka mukaan tavalla tai toisella on periaatteessa mahdollista ohjelmoida kone, jonka toiminta on erottamatonta ihmisen psykologisesta käyttäytymisestä. Tässä kriteerinä erottamattomuudelle voi olla esimerkiksi Turingin testi tai muu sellainen. *Vahva tekoälyteesi* puolestaan väittää, että tällä tavalla toimiva kone todella on mentaalinen siinä missä ihminenkin. Nämä väitteet ovat toisistaan loogisesti riippumattomia, vaikka on helppoa ajatella, että vahva teesi automaattisesti sisältää heikon. Heikko teesi kuitenkin on enemmänkin empiirinen väite siinä missä vahva teesi lähinnä käsitteellinen tai metafyysinen. On mahdollista hyväksyä mielen analyysi, jonka mukaan jollain oleellisella tavalla inhimillisesti toimiva systeemi on mentaalinen, ja kuitenkin pitää mahdottomana tällaisen koneen rakentamista. Esimerkiksi Turing selvästi uskoi vahvaan teesiin, tai ainakin ajatteli, ettei koneen ajattelukyky ole sen älykkäästä toiminnasta erillinen ongelma, ja piti heikkoa tekoälyteesiä uskottavana, joskin loppujen lopuksi avoimena empiirisenä kysymyksenä. Luonnollisesti heikko teesi osoittautuu todeksi, jos kiistatta älykkäästi toimiva kone saadaan rakennettua. Vahva teesi puolestaan edellyttää mielen tiloista ja prosesseista jonkinlaista teoriaa, jonka perusteella konetta voi pitää kirjaimellisesti mentaalisena.

Viime luvussa esiteltiin miten materiaalinen olio voi käsitellä representaatioita tavalla, joka noudattaa niiden semanttisia ominaisuuksia. Tämä puoli kognitivistista teoriaa pyrkii vastaaman miten materiaalinen olio voi olla mentaalinen, tai tarkemmin sanoen miten mentaalisen kausaation ominaiset piirteet ovat ylipäätään mahdollista toteuttaa materiaalisin mekanismein. Toinen puoli teoriaa taas on selonteko siitä, miten mentaaliset tilat voidaan määritellä ei-mentaalisesti, ja sikäli reduktiivisesti. Eräs tällainen teoria on funktionalismi, jonka syntyhistoriaan ja keskeisiin teeseihin seuraavaksi syvennytään. Molemmat teoriat siis omalla tavallaan koittavat selventää, miten olio voidaan nähdä samalla mentaalisena ja intentionaalisena sekä samalla toisaalta materiaalisena ja mekaanisena. Teoriat eivät kuitenkaan ole kilpailevia, vaan toisiaan täydentäviä.

3.1 Psykoneuraalinen identiteettiteoria ja konefunktionalismi

Lyhyesti luonnehdittuna funktionaalien analyysi pyrkii määrittelemään jonkin tilan tai ilmiön sen kausaalisten ominaisuuksien perusteella jossain suuremmassa kokonaisuudessa. Vaihtoehtoinen lähestymistapa on määritellä ilmiöt esimerkiksi niiden sisäisen rakenteen tai ei-relationaalisten ominaisuuksien perusteella. Esimerkiksi fysikaaliset oliot ja ilmiöt oletettavasti ovat mitä ovat ympäristöstään riippumatta, ja monimutkaisten fysikaalisten systeemien tilojen luonne voidaan analysoida systeemin rakenneosien ominaisuuksien avulla. Toisin sanoen fysikaalisesta perspektiivistä katsoen monimutkaisten systeemien ominaisuudet palautuvat niiden rakenneosien fysikaalisiin ominaisuuksiin. Jos vaikkapa hehkulamppu on päällä, tarkoittaa tämä, että lampun hehkulangassa kulkee sähkövirta. Sähkövirta puolestaan on elektronien virtausta johtimessa, ja elektronit ovat hiukkasia, joilla on noin $1,6 \times 10^{-19}$ coulombin vahvuinen negatiivinen varaus, 9×10^{-31} kilogramman massa ja niin edelleen. Vaikkakin fysikaalisten suureiden, kuten coulombin, vertailuarvot joudutaan pohjimmiltaan määrittelemään kappaleiden kausaalisten vuorovaikutusten avulla, niin yleensä ajatellaan, että tarkasteltavan ilmiön rakenneosilla on tietyt luontaiset ominaisuudet, joihin ilmiö tai sen ominaisuudet voidaan palauttaa. Tämänlaista ilmiön määrittelemistä rakenneosiin purkamalla kutsutaan *mikroreduktioksi* (Oppenheim & Putnam, 1958, s.5–7). Funktionalismi puolestaan on systeemiteoreettinen lähestymistapa, jonka puitteissa tarkasteltava ilmiö puretaan osiin, mutta rakenneosien ominaisuudet eivät palaudu niiden sisäiseen koostumukseen, vaan siihen, miten ne mahdollistavat systeemin kokonaistoiminnan.

Polttomoottorin sytytystulppa tarjoaa klassisen esimerkin rakenneosan funktionaalisesta määritelmästä. Mikroreduktionistisen analyysin perspektiivistä sytytystulppa on laite, joka kipinöi, kun siihen johdetaan sähkövirta. Varsinaisen analyysin tehtävä on selvittää, miten jokin tietyn tyyppinen tulppa tuottaa kipinän. Funktionalismin perspektiivistä puolestaan sytytystulppa on laite, joka sytyttää polttoaineen moottorin sylinterissä. Funktionalistisen analyysin tehtävä on selvittää, mikä tulpan tarkoitus moottorin toiminnassa on. Funktionaalisesti ajatellen polttomoottori on laite, joka muuttaa kemiallista energiaa liike-energiaksi. Tämä tapahtuu polttamalla polttoainetta sylinterissä, mikä johtaa sylinterin paineen kasvamiseen, jonka johdosta sylinterin mäntä liikkuu ja niin edelleen. Sytytystulppa taas on mikä tahansa laite, joka saa polttoaineen syttymään. Lyhyesti siis sytytystulppa on olio, joka mikroreduktionismin perspektiivistä on kipinän synnyttäjä ja funktionalismin perspektiivistä polttoaineen sytyttäjä, jonka sisäisellä rakenteella ei ole merkitystä, kunhan se hoitaa hommansa moottorin toiminnassa. Kaksi eri tavalla toimivaa sytytystulppa eivät ole identtisiä mikroreduktionistisesta mutta kylläkin funktionalistisesta perspektiivistä, jos ne voidaan vaihtaa keskenään vaikuttamatta moottorin toimintaan.

Mielenteorian yhteydessä funktionaalisen analyysin kohteena on luonnollisesti mieli ja sen toiminta. Tässä mielen ajatellaan olevan tilasta toiseen siirtyvä kausaalinen systeemi. Mentaaliset prosessit ovat mielentilojen välisiä kausaalisia siirtymiä, jotka muun muassa välittävät ärsykkeiden ja reaktioiden välisiä suhteita. Karkeasti ottaen psykologian katsotaan tutkivan sitä, mitkä aistiärsykkeet aiheuttavat mitäkin mielentiloja, miten mielen-

tilat ovat kausaalisesti riippuvaisia toisistaan ja mitkä mielentilat aiheuttavat minkäkinlaista käyttäytymistä. Perusajatus on, että mielen toiminnasta voidaan muotoilla systeemiteoreettinen kausaalinen malli, jossa ei ole tarpeen käyttää mentalistisia termejä, vaan mikä tahansa mielentila voidaan kuvata täydellisesti kertomalla sen asema tilojen muodostamassa kausaalisessa verkostossa. Mielen funktionaalinen analyysi voidaan muuntaa psykologiseksi teoriaksi tarjoamalla siinä esiintyville tiloille tulkinta mentalistisin termein.

Huomautettakoon, että jos "ärsyke" ja "käyttäytyminen" korvataan termeillä "syöte" ja "tuloste", sekä mielentilojen väliset kausaaliset suhteet ajatellaan systeemin tilojen välisinä siirtyminä, funktionaalinen analyysi muistuttaa jonkin verran mielen kuvaamista Turing-koneena. Ottamatta vielä kantaa, voiko mieltä mielekkäästi kuvata tällä tavalla tai onko tämä edes oleellista funktionalismin kannalta, niin jos tässä vaiheessa "kausaalinen kuvaus mielestä tilojen välisinä siirtyminä" tuo mieleesi Turing-koneet ja "mielentilojen funktionaalinen määrittely" sen, että koneen tiloihin liimataan mentalistisia nimilappuja, kuten esimerkiksi *tila* q_{134} *on "pelko"*, niin olet oikeilla jäljillä.

Kipua on usein käytetty esimerkkinä mielentilojen funktionaalisesta määrittelystä. Ajatellaanpa, että organismi ajautuu aina esimerkiksi kudosvaurion seurauksena tiettyyn mielentilaan q, joka puolestaan systemaattisesti aiheuttaa ahdistusta, vetäytymisrefleksin, kiroilua, huutamista ja sen sellaista. Tällöin sanoisimme, että tila q vastaa kipua, eli tilassa q oleminen on kivuissaan olemista. Kipua voi ajatella omana enemmän tai vähemmän hyvin määriteltynä mielentilanaan, mutta funktionalismin keskeinen ajatus on, ettei mitään mielentilaa voida eristää mentaalisesta koneistosta ja tarkastella sitä sellaisenaan. Jos tällainen eristäminen edes käsitteellisesti on mahdollista, mentaalisesta systeemistä irrotettuna tila menettäisi ne kausaaliset ominaisuutensa, jotka tekevät siitä tietyn mielentilan, hieman vastaavasti kuin sytytystulppa lakkaa toimimasta polttoaineen sytyttimenä, kun se irrotetaan moottorista. Ehkä valaisevampi vertailukohde saadaan kuitenkin Turing-koneiden tiloista. Koneessa M tila q määrää yhdessä nauhan sisällön kanssa sen, mitä kone tekee. Mikäli tilaa q tarkastellaan erikseen eikä koneen M rakenneosana, tilasta ei voida sanoa yhtään mitään. Koneesta irrotettuna, siis ilman viittausta siirtymäfunktioon ja muihin konetiloihin, ei tilalla q ole yhtään mitään ominaisuuksia. Koko konetilan käsitteellä ei edes ole mitään mielekästä käyttötarkoitusta muuten kuin Turing-koneiden yhteydessä. Vastaavasti funktionaalisen analyysin mukaan ilmiötä voi tarkastella mielentilana vain, kun se esiintyy osana kokonaista mieltä.

Yllä oleva kivun pika-analyysi on tietenkin hyvin yksinkertaistava. Kipu lienee tyypillisesti käytetty esimerkki, koska äkkiseltään se vaikuttaa olevan perin yksinkertainen, suorastaan atominen, psykologinen tila, jolla on hyvin selkeäpiirteiset tuntomerkit ja syyt. Toisaalta on hyvin selvää, että kipuja on kokemuksellisesti erilaisia, eri syistä johtuvia ja erilaisilla tavoilla vaikuttavia, eikä edes tyypillinen kudosvauriosta johtuva kipu ole psykologisesti atominen tila. (Hardcastle, 2001, s.298–303) Tästä huolimatta esimerkki alustavasti paljastanee riittävän hyvin funktionalistisen analyysin luonteen.

Funktionalismi on tietenkin luonteva, ellei jopa välttämätön, kumppani komputationaaliselle mielenteorialle. Samat algoritmit voidaan toteuttaa fysikaalisesti ja toiminnallisesti hyvin erilaisilla koneilla. Vaikka nykyiset koneet pohjautuvat von Neumann-

arkkitehtuuriin, ei esimerkiksi EDVAC ja uudet kämmentietokoneet jaa fyysisen toteutuksensa puolesta juuri mitään yhteistä. Koska molemman tyyppiset koneet ovat Turing-täydellisiä, ne kuitenkin voivat suorittaa täsmälleen samat algoritmit ja periaatteessa sopivien tulkkien avulla ajaa samaa ohjelmakoodia. Ylipäätään Turing-täydellisiä koneita voidaan rakentaa äärettömän monilla tavoilla, eikä niillä tarvitse olla fyysisesti tarkasteltuna oikeastaan mitään yhteistä. Komputationaaliset prosessit ovat siis paraatiesimerkki monitoteutuvasta ilmiöstä. Erityisesti jos vahva tekoälyteesi otetaan vakavasti, psykologiset tilat eivät tietenkään voi palautua ainakaan aivotiloihin, sillä teesi edellyttää, että mentaalisia prosesseja voi tapahtua kirjaimellisesti aivottomissa koneissa. Tekoälyn ja komputationalismin keskeinen ajatushan on, että ollakseen mentaalisia olioita, koneiden ei tarvitse olla aivoja eikä aivojen elektronisia kopioita. Näin ollen komputationalismi edellyttää mielentilojen analyysiä, joka on neutraali mielentilojen fysikaalisen implementaation suhteen.

Mikäli mielen toiminta on eräänlaista laskentaa, laskennan monitoteutuvuudesta seuraa suoraan myös mentaalisuuden monitoteutuvuus. Kaikki algoritmiset prosessit tapahtuvat jonkinlaisessa fyysisessä koneessa, ja näin ollen jokaisesta tosiasiallisesti tapahtuvasta laskennasta on olemassa täydellinen fysikaalinen kuvaus. Komputationalistien mielestä tämä tietysti pätee erityisesti myös psykologisille prosesseille. Kuitenkin on erittäin epätodennäköistä, että fyysisillä systeemeillä, jotka kykenevät implementoimaan psykologiset prosessit, on mitään muuta fysikaalisesti yhteistä, kuin se, että ne ovat fysikaalisesti kykeneviä suorittamaan samoja algoritmeja. Mikäli näin on, psyykkisen toiminnan mahdollisilla fysikaalisilla implementaatioilla ei varmaankaan ole mitään sellaista mielenkiintoista yhteyttä, joka olisi ilmaistavissa fysikaalisin termein, vaan näiden systeemien yhteys löytyy niiden psykologisesti tulkitusta komputationaalisesta kuvauksesta. Mentaaliset systeemit katsotaan siis monitoteutuviksi täsmälleen samaan tapaan kuin komputationaaliset. Huomattakoon, että kyse on nimenomaan kaikista *mahdollisista* mentaalisista systeemeistä. Saattaa olla empiirinen tosiasia, että kaikki olemassa olevat mielet ovat riippuvaisia tietyllä tavalla organisoituneista aivoista, mutta vaikka näin olisikin, niin funktionalistit kiistävät, että mielten riippuvuus aivoista olisi nomologinen välttämättömyys. Mentaalisuuden monitoteutuvuudesta edelleen seuraa, että psykofyysinen teoria, joka edellyttäisi mentaalisten olioiden olevan aivollisia, olisi virheellinen.

Mentaalisuuden monitoteutuvuus ei liene itsestään selvää, joten rinnan funktionalismin kanssa on ehkä syytä aluksi tarkastella mitä ongelmia liittyy monitoteutuvuuden hylkääviin mielenteorioihin. Behaviorismin ja funktionalismin välillä ehti hetken aikaa elää *tyyppi-identiteettiteoriana* tunnettu materialistinen mielenteoria. Kyseessä on oikeastaan eräänlainen hypoteesi, jonka mukaan mielentilojen luokat, tai psykologiset tyypit, kuten esimerkiksi *kipu*, *ilo* ja *toivomus, että sataa*, ovat identtisiä tiettyjen aivotilojen kanssa. Teoriaa kutsutaan usein *psykoneuraaliseksi identiteettiteoriaksi*, mitä katson parhaaksi käyttää nimen kuvaavuuden takia. Teorian perusajatus on, että kirjaimellisesti mieli on aivot, tai mielentilat ovat aivotiloja, samassa mielessä, kuin esimerkiksi vesi on divetymonoksidia. (Kim, 2006, s.86–106.)

Tarkalleen ottaen identiteettiteoria ei ajallisesti juurikaan edeltänyt funktionalismia, vaan teoriat tulivat markkinoille suunnilleen samoihin aikoihin. Identiteettiteorian pioneeriteks-

tejä ovat Ullin Placen ”Is Consciousness a Brain Process?” (1956) ja James Smartin ”Sensations and Brain Processes” (1959), kun taas funktionalismin lähtölaukauksena yleensä pidetään Hilary Putnamin artikkelia ”Minds and Machines” (1960). Loogisen behaviorismin alkaessa menettää vetovoimaansa 50-luvun lopulla, avautui naturalistisesti suuntautuneille filosofeille taas työmaata sen selvittämisessä, mitä mielentilat sitten voisivat olla, jos ne eivät ole käyttäytymistaipumuksia. Behaviorismin ongelmaksi osoittautui pitkälti se, että käyttäytymisen ja ärsykkeen välistä suhdetta ei voida selittää viittaamatta jonkinlaiseen organismin sisäiseen toimintaan. Lisäksi käyttäytymisen oikeanlaisen kuvaamisen kannalta oleellista vaikuttaa olevan intentionaalisten kuvausten käyttö, siis puhuminen uskomuksista, haluista ja muista propositionaalisista asenteista. Koska aivojen ja mielen läheinen yhteys on tunnettu jo kauan, hyvin luontevan oloinen materialistinen mielenteoria, joka viittaa organismin sisäisiin tiloihin, saadaan samaistamalla aivotilat mielentilojen kanssa ja aivojen toiminta mentaalisten prosessien kanssa.

Identiteettiteorian heti silmiinpistävä ongelma on, että aivotilat ja mielentilat näyttäisivät olevan olemukseltaan varsin erilaisia. Tietyssä tilassa olevat aivot täyttävät tietyn alueen avaruudesta. Niiden painon voi periaatteessa punnita, ja voidaan sanoa, että aivot ovat tietyn värisiä ja näköisiä ja melko rasvaisia. Mentaalisia prosesseja puolestaan voi kuvata vaikkapa kiihkeiksi tai rationaalisiksi, mutta samaa ei voine sanoa aivoprosesseista, paitsi ehkä erittäin metaforallisessa mielessä. Vaikka aivotilat tunnetusti korreloivat mielentilojen kanssa, niin äkkiseltään ajateltuna näiden pitäminen samana asiana on karkea kategoriavirhe. Placen artikkeli on varsin lyhyt ja hahmotelman oloinen keskustelunavaus, jossa hän esitti, että huolimatta edellisen kaltaisista käsitteellisistä tarkasteluista, aivojen ja mielen samuus on kuitenkin samalla tavalla mielekäs tieteellinen kysymys, kuin esimerkiksi yleisesti hyväksytty salaman ja tietynlaisen elektronivirran välinen samuus. Placen mukaan mieli–aivot-identiteetti saattaisi empiirisesti osoittautua todeksi, vaikka nojatuolista käsin tätä samuutta on vaikea nähdä. (Place, 1956, s.34–46) Hänen argumentaationsa on perin varovaista, ja oikeastaan koko tekstin idea on, että vaikkei mielen ja aivojen samuus päde loogisesti tai käsitteellisesti, ei se myöskään ole käsitteellinen mahdottomuus.

Edellä mainittua Smartin artikkelia voinee pitää eräänlaisena Placen kirjoitelman laajennuksena. Hän käsittelee useita melko ilmeisiä vasta-argumentteja identiteettiteorioille. Näistä kolme ansaitsee erityistä huomiota. Smart tarttuu ongelmaan, että mielentiloilla ja aivoitiloilla vaikuttaisi olevan hyvin eri tyyppisiä ominaisuuksia. Toiseksi erityistä huomiota saavat näiden tilojen erilaiset epistemologiset ominaisuudet. (Smart, 1959, s.146–152,158–159) Nämä mielentilojen erityislaatuiset tiedolliset ominaisuudet ovat periaatteeltaan samoja, kuin mitkä aikoinaan huolettivat behavioristeja. Aivotilat ovat epistemologisesti ottaen julkisia, eli kenen tahansa tarkasteltavissa, kun taas mielentilat vaikuttaisivat olevan jollain tavalla yksityisiä siinä mielessä, että subjektilla on erityinen espistemologinen asema niiden suhteen. Hän on ainoa, joka varsinaisesti voi havaita tai kokea mielentilansa suoraan, kun taas kaikki muut voivat saada niistä tietoa vain epäsuorasti päättelemällä tai luottamalla kyseisen henkilön niitä koskeviin raportteihin. Saman tyyppinen epistemologinen epäsymmetria vallitsee subjektin aivo- ja mielentilojen välillä hänelle itselleen. Jos tarkastelemme omia aivojamme, esimerkiksi jonkin laitteen avulla, niin tietomme aivotiloistamme perustuu empiirisiin havaintoihin, joiden suhteen on mah-

dollista erehtyä, kun taas mielentilamme ovat meille epistemologisesti suoraan annettuja, emmekä voi erehtyä olemmeko vaikkapa kivuissamme tai ajattelemmeko kissoja vaiko koiria.[33] Kolmanneksi hän tarttuu mentaalisuuden monitoteutuvuuteen, josta kumpuavan ongelman hän muotoilee siten, että ajattelu ei voi ainakaan käsitteellisessä mielessä edellyttää aivoja, koska on kuviteltavissa, että esimerkiksi kivestä muodostuvat oliot voisivat ajatella (Smart, 1959, s.152–153).

Mihinkään näistä ongelmista Smart ei mielestäni kykene täysin tyhjentävästi vastaamaan, joskin hän ehkä onnistuu osoittamaan, että näiden vasta-argumenttien pätevyys ei ole aivan ilmeistä. Ensimmäistä hän pitää haasteista vaikeimpana, mutta toteaa, ettei ole mikään ongelma, että samoja ilmiöitä voidaan kuvata erilaisilla käsitteillä. Tässä hän on tietenkin oikeassa. Smartin mukaan ongelman ydin on, että puhuttaessa mentaalisista tiloista usein sekoitetaan itse tila ja sen sisältö. Esimerkiksi kun näen keltaisen välähdyksen, havaintosisältöäni voi kuvata keltaiseksi, mutta itse havaintotapahtumaa ei. Toisin sanoen se tapahtuma, kun koen keltaista, ei itse ole keltainen. Kun psykoneuraaliset identiteetit muodostetaan aivotilojen ja psykologisten tapahtumien, eikä mielensisältöjen, välille, ei ole mitenkään ilmeistä, että mielen- ja aivotilojen laadulliset eroavaisuudet muodostaisivat todellista ongelmaa. Siis ei ole mitenkään ilmeistä, ettei aivotiloilla voisi olla täsmälleen samat ominaisuudet kuin mentaalisilla tapahtumilla. (*ibid.*, s.148–152) Mitä tulee mielentilojen epistemologiseen yksityisyyteen, Smart vain toteaa meidän olevan tällä hetkellä sellaisessa tietämättömyyden tilassa, ettei meillä ole perusteita väittää voivamme havaita muiden mielentiloja, mutta tulevaisuuden neurotieteen avulla tämä tilanne voi muuttua (*ibid.*, s.152). Aivottomien, esimerkiksi kivestä tehtyjen, olioiden ajattelun mahdollisuudesta hän toteaa, että on myös ajateltavissa, ettei salama ole elektronivirtaus, mutta sitä se kuitenkin on. Aivan vastaavasti väitettä, että ajattelu on aivotoimintaa eikä mitään muuta, voidaan hänen mukaansa kaiken asiasta tiedetyn valossa pitää mielekkäänä tieteellisenä hypoteesina (*ibid.*, s.152–153).

Mainitut Placen ja Smartin artikkelit sekä Herbert Feiglin edellisiä perinpohjaisempi tutkielma ”The ’Mental’ and the ’Physical’” (1958) muodostavat kolmikon, joka lunasti identiteettiteorialle paikan mielenfilosofian historiassa. Mainitut filosofit painottivat, että vaikka mielentilat eivät tarkoita samaa asiaa kuin aivotilat, voivat ne silti viitata samaan asiaan. Toisin sanoen vaikka termeillä ”mielentila x” ja ”aivotila y” on eri intensio, niillä voi hyvinkin olla sama ekstensio. Oletettu metodologia psykoneuraalisten identiteettiväitteiden muotoilemiseksi on siis ilmeisesti etsiä näiden erityyppisten tilojen välisiä korrelaatioi-

[33]Tämä ainakin on hyvin yleinen filosofinen näkemys, jossa on intuitiivisesti jotain oikeaa, mutta myös jotain hyvin hämärää. En ryhdy käsittelemään asiaa enempää, mutta katson parhaaksi mainita eräitä kriittisiä puheenvuoroja. Jotkut vahvan naturalistisesti suuntautuneet filosofit ovat pyrkineet kiistämään, että toisten ja omien mielentilojen tietämisen välillä olisi mitään laadullista eroa, esim. (Ryle, 1949, s.160–173) ja Churchland useissa teksteissä, ks. erityisesti (1985) ja (1989, s.67–76). Lisäksi esimerkiksi Daniel Dennett ja Stephen Stich ovat esittäneet, että ihmisten käsitykset heidän omista mielentiloistaan ovat teoriapitoisia ja tästä syystä virheille alttiita (Stich 1983, s.228–231; Dennett 1991, s.67–68,303–309). Tällainen näkemys muistuttaa oleellisesti myös Sellarsin käsitystä mielentiloihin viittaavan kielen luonteesta. Lisäksi on empiiristä, joskin hieman hankalasti tulkittavaa, näyttöä, jonka mukaan ihmiset todella ovat erehtyväisiä mielentilojensa suhteen, esim. (Nisbett & Wilson, 1977), (Wilson, 2002, s.162–169) ja (Schwitzgebel, 2002), joista Nisbettin ja Wilsonin tulokset näyttäisivät puoltavan Dennetin ja Stichin kantaa.

ta, ja pyrkiä löytämään jonkinlaisia aivotilojen luonnollisia luokkia, jotka systemaattisesti korreloivat tiettyjen mielentilojen kanssa.[34] Vaikka identiteettiteorian läpimurto tapahtui 50-luvun lopulla, teorian juuret löytyvät jo ainakin 30-luvulta. Kirjassaan *The Physical Dimensions of Consciousness* (1933) psykologi Edwin Boring kirjoittaa:

> To the author a perfect correlation is identity. Two events that always occur together at the same time in the same place, without any temporal or spatial differentiation at all, are not two events but the same event. The mind-body correlations as formulated at present, do not admit of consideration as spatial correlation, so they reduce to matters of simple correlation in time. The need for identification is no less urgent in this case." (Boring, 1933, s.16)

Luultavasti kaikki ovat yhtä mieltä siitä, että pelkkä korrelaatio ei ole identiteetin riittävä ehto. Edes kahden termin sama ekstensio ei vielä tarkoita identiteettiä, mutta *täydellinen* korrelaatio voi tätä hyvinkin tarkoittaa. Erityisesti, jos täydellisyys tarkoittaa korrelaation nomologista välttämättömyyttä. Ainakin Feigl piti Boringin teosta merkittävänä lähteenä (Feigl, 1958, s.478), ja lainattu kohta tuntuu kuvaavan hyvin identiteettiteoreetikkojen yleistä käsitystä mentaalisen ja fysikaalisen ilmiöiden suhteesta. Mielen ja aivojen tapauksessa korrelaatiota voi kuitenkin olla huomattavan hankala tutkia, koska mielentiloja ei ilmeisesti voida mitata tai ylipäätään havaita suoraan. Voimme esimerkiksi havaita varpusen aivotiloja, mutta miten niiden korrelaatio tiettyjen mielentilojen kanssa pitäisi empiirisesti perustella? Lisäksi identiteettiteorian pitäisi jotenkin selittää, miksi aivottomilla olioilla, esimerkiksi roboteilla, ei voi olla mentaalisia tiloja. Näin ollen identiteettiteorian ongelmat ovat ensisijaisesti käsitteellisiä eivätkä empiirisiä.

Mitä identiteettiteoria tarvitsisi on mikroreduktionistinen analyysi mielentiloista. Esimerkiksi Ullin Placen mukaan väite "mielentila x on aivotila y" tulee tulkita siten, että mielentila x koostuu siitä, mistä aivotila y, eikä mistään muusta (Place, 1956, s.45,47). Aivotilat puolestaan koostunevat kokoelmasta neuronien tiloja. Psykoneuraalisten identiteettien osoittaminen edellyttäisi siis analyysiä, joka näyttäisi, miten mielentilat ovat purettavissa neurologisiksi, ja vain neurologisiksi, palasiksi. Tällainen teoria siis selittäisi myös, miksi aivottomilla olioilla ei voi olla mentaalisia tiloja. Voisi ajatella, että tämän jälkeen korrelaatioiden löytäminen olisi erillinen empiirinen ongelma, mutta tarkalleen ottaen työ olisi jo tehty. Tällöin mielentiloilla nimittäin jo olisi neurologiset määritelmät, eikä mitään mentaalista korrelaattia tarvitsisi erikseen etsiä. Tähän päivään mennessä uskottavaa – jos minkäänlaista – mikroreduktionistista teoriaa mielentiloista ei ole esitetty. Sen sijaan on kylläkin tarjottu hyviä syitä olettaa, ettei tämä ole alkuunkaan oikea tapa analysoida mentaalisia ilmiöitä. Siirrytään seuraavaksi funktionalismin alkulähteille tarkastelemaan näitä argumentteja.

Hilary Putnamin artikkelia "Minds and Machines" (1960) pidetään yleisesti funktionalismin pioneeritekstinä, ja sen voi katsoa olevan eräänlaista jatkoa Turingin "Computing Machinery and Intelligencelle". Putnam esitti, että tietyt mielen ja ruumiin väliseen

[34] Jerry Fodor on esittänyt kohtuullisen vakuuttavan argumentin, jonka perusteella psykoneuraalisia identiteettejä tuskin voi rakentaa kirjaimellisesti suoraan aivo- ja mielentilojen välille, vaan teoria on mielekäs vain, jos mielentilat samaistetaan aivotilojen luokkien kanssa. (Fodor, 1968a, s.117–119.)

suhteeseen liittyvät ongelmalliset kysymykset ovat hyvin samankaltaisia, kuin tietyt laskentaan ja Turing-koneisiin liittyvät ilmeisen ongelmattomat seikat. Ensimmäinen näistä koskee mielentilojen yksityisyyttä ja aivotilojen julkisuutta, toinen taas mielentilojen ja aivotilojen välistä yhteyttä. Hän tarkasteli kuvitteellista laajennettua Turing-konetta, joka voi tutkia itseään ja muita koneita jonkinlaisen aistinjärjestelmän avulla. Erityisesti se voi havainnoida omia ja muiden koneiden syötteitä, tulosteita ja fysikaalisia tiloja. Lisäksi hän oletti, että tämä kone voisi muodostaa jonkinlaisia teorioita tekemistään havainnoista. Tässä ei tarvitse olettaa, että kone olisi kovinkaan pätevä tai luova tietotyöläinen. Riittää ajatella, että kone voi tuottaa jonkinlaisia induktiivisia yleistyksiä sekä teorioita ainakin jossain löyhässä, esimerkiksi sivulla 31 esitetyssä, teorian merkityksessä. (Putnam, 1960, s.148–149) Riippumatta tekoälyteesien pätevyydestä, tämänlainen kone lienee periaatteessa rakennettavissa.

Putnam huomautti, että tällaisen koneen konetilojen suhde sen fyysisiin tiloihin on epistemologisesti varsin samankaltainen, kuin psykologisten ja aivotilojen välinen suhde mentaalisilla olioilla. Oletetaan, että kone on esimerkiksi konetilassa q tasan silloin, kun sen fyysinen kytkin k on päällä, ja se kykenee muodostamaan itseään koskevan arvostelman ”Olen tilassa q, jos ja vain jos kytkin k on päällä.” Tämä on tietenkin aidosti empiirinen tiedonpalanen, koska kone voisi olla rakennettu monella sellaisella tavalla, että kytkin k vastaisi jotain muuta konetilaa, tai esimerkiksi siten, että kyseistä kytkintä ei olisi olemassakaan. Lisäksi kone voisi olla tilassa q ja havaita, että k onkin pois päältä, jolloin se joko voisi hylätä muodostamansa yleistyksen tai olettaa tehneensä virheellisen havainnon. Ihmisen tapauksessa vastaavanlainen havainto voisi esimerkiksi olla ”Olen kivuissani, jos ja vain jos C-ryhmän aksonini ovat aktivoituneita” – klassinen esimerkki identiteettiteorioiden yhteydessä. (*ibid.*, s.149–150.) Yleisesti ottaen siis koneen muodostamat arvostelmat muotoa ”olen konetilassa q, jos ja vain jos kytkimeni ovat tilassa k” oleellisesti vastaavat empiiristä havaintoa ”olen mielentilassa x, jos ja vain jos aivoni ovat tilassa y”, edellyttäen, että koneella on jollain tavalla samanlainen epistemologinen suhde omiin konetiloihinsa kuin ihmisellä mielentiloihinsa.

Miten kone tietää missä tilassa se on? Jos ajatellaan, että koneen tulee raportoida tilansa, jotta on mielekästä sanoa, että se on varmistanut asian, kone voisi tulostaa nauhalle vaikkapa symbolin # aina ollessaan tilassa q. Mutta vastaavasti kuin ihmisen ei tarvitse suorittaa mitään erityistä introspektion aktia tietääkseen olevansa tuskissaan, ei koneenkaan tarvitse mitenkään luodata itseään varmistuakseen omasta tilastaan. Symbolin # tulostaminen seuraa suoraan tilasta q samalla tavalla, kuin Karin itku ja hampaiden kiristys hänen tuskistaan. Molemmissa tapauksissa ylimääräinen empiirinen tai introspektiivinen tarkistus on tarpeeton ja tavallaan jopa mahdoton. Mikään tutkimus ei kummassakaan tapauksessa voi tuoda oliolle itselleen mitään lisätietoa asiasta. Tässä mielessä väitteet ”Kone M tietää, että se on tilassa q” ja ”Kari tietää olevansa tuskissaan” ovat analogisia, tai noudattavat samaa logiikkaa, kuten 60-luvulla vielä oli tapana sanoa. (*ibid.*, s.154–157.) Jonkinlaisen häiriön sattuessa kone voi tietenkin raportoida tilansa väärin, mutta samoin Kari voi jossain henkisessä tai ruumiillisessa erikoistilanteessa käyttäytyä miten sattuu. Joka tapauksessa tietyssä mielessä kone ei voi erehtyä omasta konetilastaan sen enempää kuin me omista mielentiloistamme.

Toisaalta jos haluamme edellä mainituista raporteista riippumatta tietää, onko kone tilassa q tai onko Kari tuskissaan, niin asian selvittäminen voi olla hyvinkin haastavaa. Periaatteessa jos tiedämme mitä kone missäkin tilassa tulostaa ja miten Kari missäkin tilassa käyttäytyy, niin asian selvittäminen olisi yksinkertaista. Jos kuitenkin lähtisimme selvittämään tilannetta niin sanotusti puhtaalta pöydältä, joutuisimme selvittämään koneen ohjelman rakenteen, jotta voisimme tietää sen tulostavan symbolin # vain, jos se on tilassa q, ja että se päätyy tilaan q vain syötteillä $\alpha, \beta, \ldots$ ja niin edelleen. Vastaavasti psykologisen tilan selventäminen vaatisi tietoa siitä, mikä Karin ajaa tuskiinsa ja miten hän tällöin oikeastaan käyttäytyy. Toisaalta koneelle sen oman fysikaalisen rakenteen selvittäminen voisi vaatia vastaavanlaista empiiristä tutkimusta. Jos kone haluaisi selvittää esimerkiksi onko sen transistori y kärähtänyt, tämä saattaisi vaatia mekaanikolla käyntiä samalla tavalla, kuin Kari ehkä joutuisi käymään lääkärissä tarkistaakseen onko hänellä sappikiviä. Siispä muiden olioiden kone- ja mielentilat ovat sellaisia, mistä me voimme erehtyä, ja toisaalta koneen ja Karin fysikaaliset tilat ovat sellaisia, joista ne voivat omalla kohdallaan erehtyä. (Putnam, 1960, s.157–162.) Näin ollen mielen ja ruumiin tilojen epistemologiset ominaisuudet ovat kaikille osapuolille oleellisesti samanlaiset, kuin koneen ohjelman ja sen fysikaalisten tilojen suhde.

Mainitussa artikkelissa Putnam ei väittänyt, että mentaaliset oliot ovat Turing-koneita tai muita komputationaalisia systeemeitä, vaan päinvastoin kiisti tämän (*ibid.*, s.161). Hänen pyrkimyksenään oli vain osoittaa, että mieli–ruumis-ongelmassa ja ohjelma–kone-epäongelmassa kyse näyttäisi olevan loogisesti samanlaisesta asiasta. Mikäli näin on, mentaalisten tilojen luonne on monella tapaa samankaltainen kuin konetilojen. Tämä on niin sanotun *tietokonemetaforan* tai *-analogian* ydinajatus: olipa ihminen kirjaimellisesti jonkinlainen tietokone tai ei, ihmismielen saloja voidaan ymmärtää vertaamalla mieltä virtuaalikoneisiin. Minkälainen sitten on konetilojen luonne? Itsessään ne eivät oikeastaan ole yhtään mitään. Putnam huomautti, että konetiloista voi mielekkäästi sanoa vain sen, miten siirtymäfunktio liittää ne toisiinsa, ja mikä kausaalinen rooli milläkin tilalla on konetta kokonaisuutena tarkastellen. Erityisesti konetilat eivät ole koneiden fyysisiä tiloja. Esimerkiksi kytkimen x ja konetilan q välinen yhteys koskee tietyn Turing-koneen M tietynlaista fysikaalista implementaatiota, eikä millään tavalla liity koneen M tai tilan q olemukseen sinänsä. Turing-koneen täydellinen kuvaus ei pidä sisällään minkäänlaista viittausta sen mahdolliseen fysikaaliseen implementaatioon, ja tila q on täysin määritelty koneen kuvauksessa. (*ibid.*, s.148–162) Mikäli Putnam on tietokoneanalogiassaan oikeassa, sama pätee siis mielentilojen ja aivotilojen suhteeseen, jolloin mielentilat eivät ole aivotiloja eivätkä mentaaliset ominaisuudet neurofysiologisia ominaisuuksia.

Vuonna 1967 Putnam lopulta esitti empiirisenä hypoteesina, että ihmiset – tai mentaaliset oliot ylipäätään – ovat eräänlaisia Turing-koneiden kaltaisia automaatteja ja psykologiset tilat konetiloja (Putnam 1967a, s.162–163; 1967b, s.424). Sinänsä ei ole kovin mielenkiintoista, että ihmiset ovat jossain mielessä Turing-koneita. Mikä tahansa fyysinen systeemi voidaan kuvata jonkinlaisena automaattina, yleensä monellakin tavalla. Esimerkiksi tiiliskiven voi kuvata yksitilaisena automaattina tai nauhattomana Turing-koneena, jonka siirtymäfunktio, nauha-aakkosto ja hyväksymistilojen joukko on tyhjä. Tällainen automaatti ei siis reagoi mihinkään syötteeseen, ja se tunnistaa tyhjän kielen. Kun fysikaalista systeemiä tarkastellaan automaattina, tai Turing-koneena, niin oleellista on,

että systeemistä löytyy jonkinlainen vakaa kausaalinen rakenne, joka implementoi koneen tilasiirtymäfunktion. Lisäksi koneen käsittelemät symbolit tulee olla tavalla tai toisella fysikaalisesti toteutettu systeemissä. Teknisesti ottaen siis fysikaalisen systeemin F tulkinta Turing-koneena M on surjektiivinen homomorfismi $h : F \to M$, ja F implementoi koneen M, jos ja vain jos tulkinta h on olemassa.[35] Näin ollen mikä tahansa fysikaalinen systeemi implementoi ainakin jonkinlaisen Turing-koneen, mutta ei kuitenkaan mitä tahansa konetta. Esimerkiksi tiiliskiveä tuskin voi pitää kovinkaan monimutkaisena automaattina. Jotta Putnamin ehdotus olisi erityisen mielenkiintoinen, edellytääisi tämä selontekoa, miten ihminen voidaan kuvata Turing-koneena tavalla, jolla on jotain tekemistä ihmisen psykologian kanssa.

Artikkelissaan "Psychological Predicates" Putnam esitti varsin luontevan ajatuksen, että koneen syöte yhdistetään organismin aistijärjestelmän tuottamaan informaatioon ja tuloste puolestaan motoriseen vasteeseen. Organismit luonnollisesti kulkevat jatkuvasti läpi erilaisten fysiologisten tilojen $t_1, \ldots, t_n$, jotka ovat jossain kausaalisessa suhteessa sekä toisiinsa että aisti-informaatioon ja motoriseen vasteeseen. Systeemin *funktionaaliseksi organisaatioksi* Putnam kutsui kuvausta, joka kertoo millä todennäköisyydellä organismi siirtyy tilasta t tilaan t' aistiärsykkeen α saatuaan, ja mitä motorista vastetta β se tällöin tuottaa. Hän jätti tarkoituksella avoimeksi, onko tilojen t tarkoitus olla esimerkiksi aivotiloja. Nyt organismin Turing-konekuvaus saadaan edellisestä suoraviivaisesti yhdistämällä aistijärjestelmän tuottama informaatio koneen syötteeseen, motorinen vaste tulosteeseen, tilat $t_1, \ldots, t_n$ konetiloihin ja funktionaalinen organisaatio siirtymäfunktioon. Tässä siis Turing-koneen määritelmää on muutettu sen verran, että nauha on korvattu sensorimotorisella systeemillä. Lisäksi Putnam katsoi parhaaksi pitää siirtymäfunktiota probabilistisena.[36] On tietysti selvää, että kun "fysiologiset tilat" jätetään sen tarkemmin määrittelemättä, organismin funktionaalinen organisaatio ei ole yksiselitteinen. Joka tapauksessa Putnam esitti, että organismin täytyy implementoida tietynlainen Turing-kone, jotta sillä voisi olla psykologisia tiloja, kuten kipuja, iloja tai pelkoja, ja tietyssä psykologisessa tilassa oleminen on tietyssä konetilassa olemista. (Putnam, 1967a, s.162–163.) Malli toimii suurinpiirtein siten, että esimerkiksi jos oliot A ja B pelkäävät hämähäkkejä, niin riippumatta siitä, miten erilaisia A ja B keskenään muuten ovat, molemmat voidaan kuvata samanlaisena koneena M, joka päätyy aina hämähäkkejä kohdatessaan tiettyyn konetilaan q, joka aiheuttaa tiettyä pelkoon liittyvää luonteenomaista käyttäytymistä.

Vastoin kuin behaviorismi, konefunktionalismi selvästikin pitää mentaalisia tiloja organismille sisäisinä. Konefunktionalismia voi kuitenkin pitää jonkinlaisena laajennettuna behaviorismina. Molemmat sisältävät ajatuksen, että jotkin fysiologiset ilmiöt välittävät aistiärsykkeen ja käyttäytymisen välistä suhdetta. Nämä ovat jonkinlaisia refleksejä behaviorismin yhteydessä ja konetilojen siirtymiä funktionalismin tapauksessa. Kuitenkin konetilafunktionalismi sisältää mallin organismin sisäisten tilojen mekaniikasta, joka joh-

[35] Ks. edellinen luku, s.44 ja Dennett (1978c, s.256–266). (Piccinini, 2007) käsittelee tätä kysymystä ja sen merkitystä mielenfilosofialle yleisemmin.

[36] Esimerkiksi jos organismi siirtyy tilasta t tilaan t' todennäköisyydellä 0.4, ja tilaan t'' todennäköisyydellä 0.6 saadessaan aistiärsykkeen α, niin probabilistinen siirtymäfunktio δ on sellainen, että $\delta(q, \alpha) = (q', \beta)$ todennäköisyydellä 0.4 ja $\delta(q, \alpha) = (q'', \beta^*)$ todennäköisyydellä 0.6, missä β ja β^* ovat joitain motorisia vasteita, q vastaa tilaa t ja niin edelleen. Normaali Turing-kone on probabilistisen koneen eritystapaus, missä siirtymätodennäköisyys funktion δ määräämään yksikäsitteiseen tilaan on 1.

taa hyvin erilaiseen analyysin mielentiloista sekä psykologian teorianmuodostuksesta ja tutkimuskohteesta. Siinä missä behaviorismin mukaan psykologian tutkimuskohteena on ainoastaan ärsykkeen ja reaktion välinen suhde, organismin sisäisen rakenteen jäädessä psykologian ulkopuolelle, konefunktionalismissa psykologiset tilat vastaavat konetiloja, ja psykologiset prosessit ovat komputationaalisia mekanismeja, jotka liittävät käyttäytymisen ärsykkeeseen. Huomion arvoista on, että konetilafunktionalismissa – kuten myös muissa vastaavissa komputationalistisissa teorioissa – psykologiset tilat eivät ole yksinkertaisia fysiologisia refleksejä, vaan sisäiset tilat voivat muodostaa hyvin monimutkaisia ketjuja. Tällainen sisäisen mekaniikan huomioiminen mahdollistaa muun muassa sellaisen mentaalisen toiminnan ymmärtämisen, joka ei näy ulkoisena käyttäytymisenä, sekä sen selittämisen, miksi sama ärsyke ei aina johda samaan reaktioon. Turing-koneessa, kuten komputationaalisissa systeemeissä yleensä, tuloste ei nimittäin määräydy pelkästään syötteen vaan myös sisäisen tilan perusteella.

Keskeisin syy, miksi identiteettiteoria jäi äkkiä funktionalismin jalkoihin, on jälkimmäisen tarjoama mahdollisuus mentaalisuuden monitoteutuvuuteen. Miksi tämä seikka sitten on niin kriittinen? Vakioargumentaatio etenee suurinpiirtein seuraavasti:[37] Eri nisäkäslajeilla on pitkälti samanlaiset aivot, joten muiltakin nisäkäslajeilta kuin ihmisiltä löytynee aivotiloja, jotka identiteettiteorian mukaan laskettaisiin mielentiloiksi. Joskin tämä luonnollisesti riippuu siitä, miten tiukasti psykoneuraaliset identiteettit määritellään. Toisaalta taas esimerkiksi mustekalojen hermosto poikkeaa nisäkkäiden vastaavasta siinä määrin, että ihmiset ja meritursaat tuskin jakavat ainoatakaan aivotilaa keskenään. Mikäli näin on, identiteettiteoriasta seuraa suoraan, että tursailla ja ihmisillä ei voi olla ainoatakaan yhteistä psykologista tilaa. Jos me ja meritursaat kuitenkin voimme olla jossain samassa psykologisessa tilassa, kuten esimerkiksi nälissämme, niin psykoneuraalinen identiteettiteoria on ylittämättömissä ongelmissa. Nimittäin jos ihmisten tapauksessa onnistutaan löytämään identiteetti *nälkä* = *aivotila* x, pitäisi mustekalojen aivoista löytyä vastaava tila x, joka identifioidaan puhtaasti neurofysiologisten kriteereiden perusteella. Ilmeisesti tämä on kuitenkin mahdotonta. Identiteettiteoreetikko voi tietenkin puolustautua väittämällä, että meritursailla tuskin on tosiasiassa yhteisiä psykologisia tiloja ihmisten kanssa, vaikka pinnallisesti siltä saattaa näyttää, ja on katteetonta antropomorfismia nähdä samanlaisuutta pinnallisessa samankaltaisuudessa. Tätä väistöliikettä varten joudutaan turvautumaan hieman mielikuvituksellisempaan argumenttiin.

Ajatellaanpa, että joku päivä maapallolle saapuu yllättäen Marsin pinnan alla elelevän sivilisaation lähetystö. Alkuhankaluuksien jälkeen kommunikaatio saadaan toimimaan ja marsilaiset osoittautuvat yllättävän miellyttäviksi ja ihmisenkaltaisiksi kavereiksi. Syntyy kauppa- ja ystävyyssuhteita, vierailijat keskustelevat kanssamme melko sujuvasti aiheesta kuin aiheesta, käyvät elokuvissa, kirjoittelevat kirjoja ja ylipäätään vaikuttavat olevan mentaalisilta kyvyiltään ihmisten kanssa täysin tasaveroisia. Eräänä päivä Marsin suurlähettiläs menehtyy epäilyttävissä olosuhteissa. Seuraa tutkinta ja ruumiinavaus, jossa paljastuu, että marsilaisilla ei ole aivoja ollenkaan, vaan pääkopasta – tai vastaavasta ruumiinosasta – löytyykin jonkinlaista nanomittakaavan supertietokonetta muistuttava koneisto, jolla sattuu olemaan täsmälleen ihmisaivoja vastaava funktionaalinen organisaatio. Mikä olisi tämän löydön merkitys? Identiteettiteoreetikolle löytö paljastaa, ettei marsilai-

[37] Ks. esim. (Putnam, 1967a, s.164–165).

silla tosiasiassa ollutkaan minkäänlaisia mielentiloja, vastoin valtavaa anekdotaalista ja behavioraalista todistusaineistoa. Toisaalta marsilaisten funktionaalinen organisaatio sattuu olemaan samanlainen kuin meillä, joten mikä tahansa ihmisiä koskeva psykologinen teoria soveltuisi marsilaisten käyttäytymisen kuvaamiseen ja ennustamiseen yhtä hyvin kuin ihmisten. Kuitenkin identiteettiteorian mukaan psykologisten teorioiden soveltaminen näihin otuksiin olisi käsitteellisistä syistä virheellistä. Tämä johtopäätös vaikuttaa järjen vastaiselta. Jos psykologiset teoriat eivät erottele marsilaisia ihmisistä mutta psykoneuraalinen teoria erottelee, niin tällöin mielentilat joko eivät ole aivotiloja tai sitten ne eivät ole psykologisia tiloja. Valinta näiden välillä ei liene vaikea.

Psykoneuraalinen identiteettiteoria vaikuttaa nousseen lähinnä behaviorismin hylkäämistä seuraavasta huolesta, että mielenteoria saattaa luisua takaisin dualismiin, jollei uskottavaa materialistista vaihtoehtoa ole tarjolla.[38] Se ei syntynyt mistään uudesta empiirisestä löydöstä tai teoreettisesta oivalluksesta, ja tästä syystä se ei kyennyt tarjoamaan uutta mielenkiintoista tapaa ymmärtää mieltä tai tehdä psykologiaa. Funktionalismin noste puolestaan tuli lähinnä komputationaalisesta mielenteoriasta, joka oli uusi ja mielenkiintoinen teoreettinen oivallus. Lisäksi, kuten aiemmin on tullut mainittua, psykologia ja kielitiede oli 60-luvulle tultaessa siirtynyt vaiheeseen, jossa mielen käsittäminen tietojenkäsittelyjärjestelmänä nähtiin tarpeelliseksi, ellei jopa välttämättömäksi. Tällaisen tieteen filosofiaksi funktionalismi sopii erinomaisesti. Lisäksi konetilafunktionalismi oli yhteensopiva materialismin kanssa, mutta vältti yksioikoisen fysikalismin ongelmat tarjoamalla mielentiloille abstraktimman ja monessa mielessä luontaisemman analyysin. Konefunktionalismi oli kuitenkin vain yksi ja lyhytikäinen funktionalismin muoto, joka aloitti sarjan teoreettisia kehitelmiä, jotka myös ottivat jonkin verran etäisyyttä näin suoraviivaiseen komputationaaliseen reduktionismiin.

3.2 Yleistetty funktionaalinen anayysi

Putnamilainen konetilafunktionalismi osoittautui äkkiä ongelmalliseksi, mutta sen myötä syntyneet funktionalismin ydinajatukset jäivät elämään. Ennen siirtymistä funktionalismin jatkokehittelyyn, vedetään yhteen nämä 1900-luvun loppupuoliskon kognitiotiedettä ja mielenfilosofiaa hallinneet teesit.[39] 1° Mentaalisia olioita yhdistää se, että ne toteuttavat ainakin osittain saman funktion ψ, joka kuvaa organismin funktionaalisen organisaation tilan ja (aisti)syötteen uudeksi tilaksi ja (motoriseksi) vasteeksi. Funktion ψ avulla voidaan määritellä mentaalisten olioiden luonnollinen luokka. 2° Psykologiaa ei voi palauttaa neurotieteisiin, koska mikä tahansa funktio, ja tällöin siis erityisesti funktio ψ, voidaan laskea äärettömän monella tavalla, eikä laskennan fysikaalisella implementaatiolla ole periaatteessa merkitystä. 3° Vaikka aivojen ja mielen välisen suhteen tutkiminen on mielekästä ja mielenkiintoista, aivojen monimutkaisuuden takia niiden tutkiminen on käytännössä huono lähestymistapa funktion ψ selvittämiseksi. Joka tapauksessa, edellä mainituista syistä, aivojen tutkiminen ei ole mielen tutkimisessa välttämätöntä tai edes oleellista. 4° Kognitiotieteen tehtävänä on määrittää funktio ψ, ja edellä sanotusta seuraa

[38] Ks. (Place, 1956, s.44), (Feigl, 1958, s.446) ja (Smart, 1959, s.156).

[39] Tämä lista on pienin muutoksin lainattu Paul Churchlandin kriittisesti funktionalismia tarkastelevasta artikkelista "Functionalism at Forty" vuodelta 2005 (s.19–20).

kognitiotieteen autonomia suhteessa fysiikkaan, biologiaan ja neurotieteisiin. 5° Tutkimuksen kulmakivet ovat kognitiivinen psykologia, joka selvittää asiaa niin sanotun takaisinmallinnuksen (engl. *reverse engineering*) avulla tarkastelemalla olion reaktioita tietyissä olosuhteissa, sekä tekoälytutkimus, joka pyrkii luomaan ainakin osittaisia implementaatioita tutkittavasta funktiosta. Lisäksi usein oletetaan, että 6° arkipsykologia on oikean suuntainen, joskin hyvin karkea, kuvaus mielen toiminnasta, eli se kuvaa funktion ψ ainakin osittain oikein. Käytännössä tämä tarkoittaa, että funktion määrittelyssä esiintyvät tilat $t_1, t_2, t_3, \ldots$ ovat propositionaalisia asenteita tai jotain vastaavia. Näin ollen kognitiotieteen oletetaan tuottavan uskomus–halu-psykologialle reduktiivisen komputationaalisen teorian.

Tässä vaiheessa yllä olevien teesien pitäisi vaikuttaa hieman epämääräisiltä, mutta toivottavasti lukija kykenee nyt suurinpiirtein hahmottamaan keskeisimpien väitteiden 1° ja 2° merkityksen ja perustelut. Kohta 1° on oikeastaan Turingin idea, että oliosta ei tee ajattelevaa se, mikä se on vaan mitä se tekee. Toisaalta siinä missä hän katsoi tarkasteltavien syötteiden ja tulosteiden olevan kielellisiä, niin ainakin Putnamista lähtien niitä on yleensä ajateltu sensorimotorisina. Syy tähän lienee, että useimmat teoreetikot pitävät mieltä perimmiltään käyttäytymisen eikä ajattelun instrumenttina, ja kognitiivisen teorian pääasiallisena tavoitteena käyttäytymisen etiologian selvittämistä. Ajattelua ja kielellistä käyttäytymistä puolestaan yleensä pidetään käyttäytymisen erityistapauksina, joten kyseessä on siis oikeastaan Turingin näkökannan yleistys.

Mutta mitä tarkoittaa, että mentaaliset oliot toteuttavat jonkin saman funktion? Onko funktiossa ψ kyse organismien funktionaalisesta organisaatiosta vai esimerkiksi jostain psykologista teoriaa vielä yleisemmästä kuvauksesta, jonka erilaiset mentaaliset oliot toteuttavat omalla tavallaan? Esimerkiksi mustekaloilla ja ihmisillä ei varmasti ole kokonaisuudessaan samanlaista psykologiaa, joten jos niiden välinen psykologinen samankaltaisuus on jonkinnäköinen argumentti identiteettiteoriaa vastaan, pitäisi funktionalistien pystyä selvittämään, mitä tämä samankaltaisuus oikeastaan tarkoittaa. Ei varmasti pidä paikkaansa, että edes eri ihmiset yleisesti ovat psykologisesti identtisiä. Mikä siis oikeastaan on kognitiotieteen tutkimuskohde, millä kuvauksen tasolla mentaaliset oliot ovat identtisiä ja miltä funktionalistisen mielenteorian ylipäätään tulisi näyttää? Konetilafunktionalismi oli eräs vastaus näihin kysymyksiin, joka ei kuitenkaan kantanut kovin pitkälle.

Vaikka äkkiseltään konefunktionalismi näyttäisi välttävän psykoneuraalisen identiteettiteorian ongelmat, lähempi tarkastelu osoittaa, että se itse kaatuu lähinnä monitoteutuvuuden vaatimukseen. Ongelma syntyy siitä, miten Turing-koneet ja niiden tilat määritellään. Turing-koneiden tiloilla ei ole mitään itsenäisiä ei-relationaalisia ominaisuuksia eikä sikäli mitään itsenäistä identiteettiä. Kaikki mitä konetilasta q voidaan sanoa on, että mistä tiloista $q_1, \ldots, q_n$ milläkin syötteellä $\alpha_1, \ldots, \alpha_h$ tilaan voi päätyä, ja edelleen mihin tilaan $q'_1, \ldots, q'_m$ kone syöteillä $\beta_1, \ldots, \beta_m$ tilasta q päätyy. Mutta edes nämä relationaaliset ominaisuudet eivät täsmällisesti määrittele tilan q identiteettinä. Kahdesta eri Turing-koneesta voi molemmista löytyä tila q siten, että sillä on koneissa täsmälleen samat relationaaliset ominaisuudet, vaikka koneet kokonaisuudessaan ovat erilaisia ja laskevat jotain aivan eri funktiota. Tällöin näiden tilojen kausaalinen rooli syötteen ja tulosteen

välittämisessä voi olla täysin erilainen. Tilojen relationaalinen identifiointi edellyttää, että tilan q identifikaatiokriteereihin kuuluu myös selonteko siitä, miten siihen kytkeytyneet tilat $q_1, \ldots, q_n$ puolestaan kytkeytyvät muihin tiloihin $q'_1, \ldots, q'_m$, mikä rooli niillä koneen toiminnassa on ja niin edelleen. Toisin sanoen konetilan ainoa mielekäs identiteettikriteeri on täysin holistinen, eli mikä asema sillä on kokonaisessa Turing-koneessa. Lisäksi myös kaksi hyvin erilaista Turing-konetta voivat laskea täsmälleen samaa funktiota, joten jopa riippumatta siitä, tulostavatko koneet aina saman vasteen samalla syötteellä, koneet eivät välttämättä sisällä identtisiä tiloja. Näin ollen kaksi eri Turing-konetta ei jaa ainoatakaan yhteistä, tai missään mielekkäässä mielessä samaa tilaa, elleivät ne ole kokonaisuudessaan täysin samanlaisia.

Näin ollen konetilafunktionalismista seuraa, että kahdella oliolla voi olla samoja mielentiloja, vain jos niillä on psykologian suhteen identtinen funktionaalinen organisaatio. Koska esimerkiksi mustekaloilla on varmasti erilainen kokonaispsykologia kuin ihmisillä, täytyy mainituilla olioilla funktionaalisten organisaatioiden jonkin verran poiketa toisistaan. Konefunktionalismin perusteella ne eivät näin ollen jaa mitään yhteisiä psykologisia tiloja. Näin ollen teoria kaatuu täsmälleen samaan ongelmaan, josta se syytti psykoneuraalista identiteettiteoriaa. Tilanne pahenee edelleen, jos tarkastellaan esimerkiksi aivovaurion saaneita ihmisiä, esimerkiksi wernicken afaatikkoja, jotka menettävät kyvyn ymmärtää luonnollista kieltä. Koska afaatikkojen funktionaalinen organisaatio muuttuu aivovaurion seurauksena, konetilafunktionalismista seuraa se absurdi johtopäätös, että he menettäisivät kaikki psykologiset tilansa kerralla.

Yllä olevan vasta-argumentin alkujaan esittivät Ned Block ja Jerry Fodor tunnetussa konetilafunktionalismia kritisoivassa artikkelissaan "What Psychological States are Not." Kaikkinensa he esittivät kuusi varsin perusteellista argumenttia teoriaa vastaan. Myös Putnam lopulta sanoutui irti teoriastaan, ja jatkoi itse vasta-argumenttien listaa vuonna 1975 esseessään "Philosophy and Our Mental Life" (s.298–299). Nostan esiin tässä vielä lyhyesti muutaman Blockin ja Fodorin esittämän huomion, jotka ovat merkittäviä funktionalismin jatkokehittelyn kannalta.

Ensinnäkin konetilafunktionalismissa ei näyttäisi olevan mahdollista erottaa dispositionaalisia mielentiloja ei-dispositionaalisista, joita voitaisiin kutsua episodisiksi mielentiloiksi. Vaikka tämä erottelu on varsin oleellinen, usein mielentiloista puhuttaessa ei sitä huomata tehdä. Tarkastellaan lausetta "x pelkää y:tä", joka siis ilmaisee x:lle kuuluvan propositionaalisen asenteen *pelko* jotakin y:tä kohtaan. Selvästikään mainittu lause ei tarkoita x:n olevan jatkuvasti pelkotilassa y:n takia, vaan lause kertoo x:n erään psykologisen taipumuksen. Tässä mielessä pelko on dispositionaalinen mielentila. Jos x pelkää hämähäkkejä, niin hän on tyypillisesti peloissaan hämähäkkejä kohdatessaan, ja tällaisina hetkinä pelko on mielentila, jossa x kirjaimellisesti on. Näitä jälkimmäisiä mielentiloja kutsutaan episodisiksi. Konetilafunktionalismi antanee esimerkiksi pelolle varsin ongelmattoman analyysin: pelko on eräs konetila, johon organismi ajoittain siirtyy. Ongelma syntyy täysin dispositionaalisista mielentiloista, kuten uskomisesta. On tietenkin hetkiä, jolloin ihmiset ilmaisevat tai hyväksyvät uskomuksiaan, mutta tuskin sellaisia mentaalisia episodeja, jolloin ihmiset ovat minkäänlaisissa erityisissä uskomustiloissa. Ehkä konefunktionalismin puitteissa puhtaasti dispositionaalisille mielentiloille voisi olla jonkinlai-

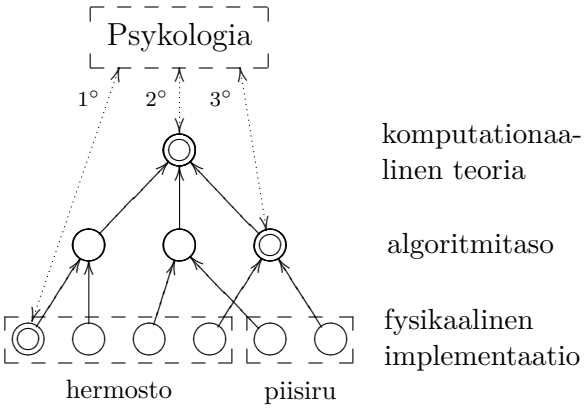
nen analyysi, mutta teorian väite, että mielentilat vastaavat yksiselitteisesti konetiloja, on tarkemmin katsottuna hyvin epäuskottava. (Block & Fodor, 1972, s.168–170)

Toiseksi konetilafunktionalismi ei ilmeisesti mahdollista kompleksisten episodisten mielentilojen olemassaoloa, koska automaatit ovat yhdessä määrätyssä tilassa kerrallaan. Näin ollen malli kuvaa mielentilojen välisen kausaation etenevän aina yhdestä tietystä tilasta toiseen. Tämä on teorian etu verrattuna behaviorismiin, jossa mielentilojen ketjujen selittäminen tuottaa hieman hankaluuksia. Toisaalta käyttäytymisen selittämisen kannalta on usein oleellista, mitä agentti milläkin hetkellä esimerkiksi haluaa, pelkää, havaitsee ja niin edelleen. Mikäli käyttäytyminen joskus riippuu useista samanaikaisista episodisista mielentiloista, niin tämän selittäminen edellyttää mallia, joka kykenee selventämään, miten olio voi olla useissa tiloissa samaan aikaan. Konetilafunktionalismi huomioi mielentilojen välisen kausaation, mutta jos teoria ei kykene kuvaamaan miten käyttäytyminen riippuu nimenomaan mielentilojen komplekseista, teoria on vaarassa kaatua samoihin ongelmiin kuin behaviorismi. (Block & Fodor, 1972, s.170–172)

Kolmanneksi teoriassa ei näyttäisi myöskään olevan sijaa mielentiloille, joissa kompleksinen mentaalinen representaatio on rakennetekijänä. Representationalistisen mielenteorian varmaankin pitäisi pystyä selittämään, miten *x uskoo, että A ja B* on systemaattisesti riippuvainen siitä, että *x uskoo, että A*, mutta on hankala hahmottaa, miten konefunktionalismi tähän kynenisi. Tietysti voidaan ajatella, että kompleksiset mielentilat eivät ole erillisiä tiloja, vaan *x uskoo, että A ja B* tarkoittaa, että *x uskoo, että A* ja *x uskoo, että B*, mutta on hankala nähdä miten tämä taas toimisi disjunktiivisten *x uskoo, että A tai B* tai ehdollisten uskomusten *x uskoo, että jos A, niin B* kanssa. Toinen vaihtoehto näyttäisi olevan yksinkertaisesti väittää, että kompleksiset tilat itse asiassa ovat erillisiä mielentilojaan. Tämä kuitenkin rikkoo Turing-koneen määritelmää, koska kompleksisia representaatioita on mahdollisesti äärettömästi, mutta konetilojen joukko on aina äärellinen. Ennen kaikkea kuitenkin teoria tällöin väittäisi, että mentaalisilla olioilla on äärettömästi toisistaan riippumattomia mentaalisia tiloja, jokainen erillistä halua, uskomusta, pelkoa ja muuta sellaista kohti. Esimerkiksi *x pelkää karhuja* vastaisi omaa konetilaansa, *x pelkää karhuja ja käärmeitä* taas omaansa ja niin edelleen, mikä on absurdia ja teoria fiasko. (*ibid.*, s.175–177)

Edellä olevat argumentit antavat ymmärtää, että vaikka organismit voidaankin kuvata Turing-koneina monella tapaa, on vaikea kuvitella, että psykologiaa voitaisiin kuvata oikein Turing-koneena. Argumenttien taustalla on arkipsykologinen, tai ainakin jonkinlainen representationalistinen, mielenteoria, joten esimerkiksi jonkin sortin eliminativisti ei välttämättä olisi näistä vastaväitteistä kovin huolestunut. Kuitenkin konetilafunktionalismin on tarkoitus olla mielentilojen redusoiva teoria, ja on hankala kuvitella mitä itsenäisiä syitä olisi kannattaa ajatusta, että mieli on Turing-kone. Yllä olevilla argumenteilla on myös hieman yleisempää mielenkiintoa, sillä käsittääkseni ne ovat yleistettävissä mille tahansa seriaalisesti toimivalle virtuaalikoneformalismille, jossa tilojen joukko on äärellinen. Toisin sanoen, jos mieli on komputationaalinen systeemi ja mielentilat propositionaalisia asenteita, niin mielentilat tuskin palautuvat suoraviivaisesti tuon systeemin tiloiksi.

Komputationaalisten systeemien luonnetta tarkasteltaessa valoa asiaan saattaa luoda kuuluisa kolmitasoinen analyysi, joka on peräisin neurotieteilijä David Marrin vuonna 1982 postuumisti julkaistusta teoksesta *Vision* (Marr, 1982, s.19–29). Kolmitasoisen analyysin korkeinta tai abstrakteinta tasoa Marr kutsui *komputationaaliseksi teoriaksi*, joka on kuvaus siitä, mitä systeemi tekee ja miksi. Marr itse, kuten teoksen nimestä voi päätellä, keskittyi näköjärjestelmän toimintaan, jonka komputationaalinen teoria kertoisi millä tavalla näköjärjestelmä esittää maailmaa, eli minkälaisia visuaalisia representaatioita se tuottaa minkäkinlaisesta fotoreseptorien poimimasta informaatiosta. Tarkemmin sanoen tämä kertoo, mitä funktiota näköjärjestelmä itse asiassa suorittaa. Miksi-kysymykseen vastaaminen tyypillisesti edellyttää jonkinlaisen käsityksen siitä, mitä systeemi kokonaisuudessaan pyrkii tekemään, eli mitä varten se on olemassa, miksi se tekee mitä se tekee suhteessa tarkoitukseensa ja mitä ongelmia sillä on päämääränsä saavuttamisessa. Analyysin tällä tasolla systeemiä tarkastellaan oleellisesti mustana laatikkona. Analyysin toinen taso on *algoritminen teoria*, joka nimensä mukaan on kuvaus siitä, miten systeemin suorittama prosessointi täsmälleen ottaen tapahtuu informaation käsittelyn perspektiivistä. Algoritminen kuvaus sisältää selonteon systeemin käyttämistä representaatioista, kalkyylistä sekä proseduraalisista ohjeista. Tämä siis tarkoittaa systeemin virtuaalikonekuvausta. Viimeinen taso taas on *fyysinen implementaatio*, joka kertoo miten algoritmit ja representaatiot ovat fysikaalisesti toteutettu. Tämän kolmijakoisen analyysin tasot jotakuinkin vastaavat tähän mennessä tarkasteltujen mielenteorioiden reduktiivisia ajatuksia.



1° Identiteettiteoria pyrkii samaistamaan psykologiset tilat systeemin fyysisten tilojen kanssa, ja luonnollisesti tällöin mentaaliset prosessit myös ovat fysikaalisia prosesseja keskushermostossa, tai mihin fysikaalisiin ilmiöihin identiteettiteoreetikko sitten katsookaan psykologian palautuvan. 2° Behavioristi taas on kiinnostunut ainoastaan syötteen ja tulosteen, eli ärsykkeen ja reaktion, välisistä suhteista, ja katsoo psykologisten tilojen palautuvan analyysin tälle tasolle. 3° Konetilafunktionalisti puolestaan näkee, että psykologiset tilat ja prosessit palautuvat johonkin kognitiivisen olion etuoikeutettuun algoritmiseen kuvaukseen, ja tässä algoritmi tarkoittaa erityisesti probabilistista Turingin-konetta.

Marrin mukaan komputationaalisen systeemin[40] ymmärtäminen vaatii kaikkien näiden tasojen sisäisen ja välisen toiminnan ymmärtämistä, ja eri tasojen tarkastelu selittää eri asioita systeemin toiminnasta. Mainitut mielenteoriat taas koittavat kukin palauttaa psykologian tiettyyn ilmiöluokkaan. Syy tämänlaiseen suhtautumiseen on, että mielenteoriat koittavat tarjota ontologisen selonteon siitä, mitä mentaaliset tilat itse asiassa ovat, eli

[40]Huomautettakoon, että Marr tarkoittaa *komputationaalisella systeemillä* konkreettista informaatiota käsittelevää laitetta tai organismia. Muualla tässä työssä termiä käytetään enemmänkin viittaamaan siihen, mitä Marr kutsuu algoritmiseksi teoriaksi, eli analyysin tasoon, jota luonnehtii käytetty formaalikieli, kalkyyli ja muu sellainen.

teorioiden tarkoitus on selventää mikä systeemeistä tai organismeista tekee mentaalisia, ja mikä niiden toiminnassa on nimenomaan psykologista.

Mikään ei kuitenkaan pakota muotoilemaan funktionalismia siten, että mielentilat samaistetaan konetilojen ja mentaalinen kausaatio tilasiirtymien kanssa. Ylipäätään mielentilojen funktionaalinen määrittely ei edellytä, että organismeista poimittaisiin joku tietty etuoikeutettu kuvaus, jonka elementtien kanssa psykologiset tilat samaistetaan, vaan mielentilat voidaan määritellä ikään kuin niiden omilla ehdoillaan. Funktionalismin abstraktimpi muotoilu saadaan yksinkertaisesti jo aiemmin todetusta perusideasta, että mentaalinen tila identifioidaan sen kausaalisen roolin perusteella ärsykkeen ja käyttäytymisen välittämisessä. Täsmällisemmin ottaen asia voidaan muotoilla niin sanottujen Ramsey–Lewis-lauseiden avulla seuraavasti.[41] Olkoon T psykologian teoria, joka kertoo mistä mentaaliset tilat $t_1, \ldots, t_n$ aiheutuvat, miten ne suhteutuvat kausaalisesti toisiinsa ja minkälaista käyttäytymistä ne aiheuttavat. Periaatteessa tällainen teoria voidaan esittää n-paikkaisena tiloihin $t_1, \ldots, t_n$ viittaavien termien relaationa $T(t_1, \ldots, t_n)$. Teorian T Ramsey-lause muodostetaan korvaamalla termit $t_1, \ldots, t_n$ eksistenssikvanttorilla sidotuilla muuttujilla $x_1, \ldots, x_n$ relaatiossa T. Toisin sanoen Ramsey-lause on muotoa $\exists x_1, \ldots, \exists x_n : T(x_1, \ldots, x_n)$. Menemättä sen syvemmin siihen, mikä tolkku Ramsey-lauseissa on sinänsä, edellisen avulla voidaan nyt määritellä formaalisti, mitä esimerkiksi kipu teorian T mukaan on: jos x_i on muuttuja, joka korvaa Ramsey-lauseessa teorian T termin "kipu", niin tällöin olio y on kivuissaan, jos ja vain jos $\exists x_1, \ldots, \exists x_n : (T(x_1, \ldots, x_n)$ ja y on tilassa $x_i)$. Edellinen sanoo oleellisesti siis, että "Olio y on kivuissaan" tarkoittaa, että y on tilassa, jolla on tietynlaiset teorian T määrittämät kausaaliset suhteet muihin teorian sisältämiin mielentiloihin $t_1, t_2, \ldots$, joissa y voisi olla. Luonnollisesti y:n täytyy toteuttaa T, jotta teorian Ramsey-Lewis lause voisi päteä sille.

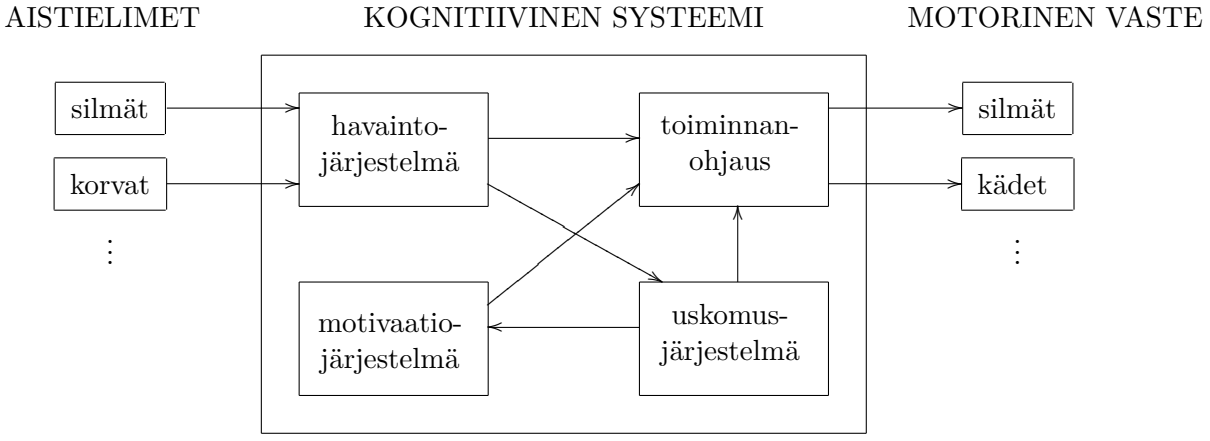
Ajatellaanpa vaikka, että teoria T sanoo kipujen aiheutuvan pistoksista ja aiheuttavan älähtelyä ja ahdistusta. Teoria lisäksi sanoo, että ahdistus puolestaan aiheuttaa irvistelyä. Tällainen kiputeoria on tietenkin esimerkkiä varten ääriyksinkertaistettu. Joka tapauksessa teoriassa on siis kaksi termiä t_1 ="kipu" ja t_2 ="ahdistus". Muut mahdolliset mentaalisiin tiloihin viittaavat termit voidaan jättää tässä huomiotta. Teorian Ramsey-lause siis on: $\exists x_1 \exists x_2 : (x_1$ aiheutuu pistoksista, aiheuttaa älähtelyä ja tilan x_2; tila x_2 aiheuttaa irvistelyä). Nyt siis tämänlaisen määritelmän valossa kipu on predikaatti, joka voidaan määritellä muuttamalla edellinen lause Ramsey–Lewis-lauseeksi: "*y on kivuissaan*, jos ja vain jos $\exists x_1 \exists x_2 : ((x_1$ aiheutuu pistoksista, aiheuttaa älähtelyä ja tilan x_2, x_2 aiheuttaa irvistelyä) ja y on tilassa $x_1)$." Ned Block kutsuu tämänlaista funktionalismia *metafyysiseksi funktionalismiksi* (Block, 1980b, s.172), mutta yleisemmin käytössä lienee termi *kausaaliteoreettinen funktionalismi*, joka paremmin kiteyttää teorian hengen.[42] Idea

[41] Idea on peräisin David Lewisin artikkelista "An Argument for the Identity Theory" (1966), joka hieman hämäävästä nimestään huolimatta ei jäänyt historiaan niinkään identiteettiteorian puolustuksena, vaan kausaaliteoreettisen funktionalismin alkusoittona. Nimitys "Ramsey–Lewis-lause" tulee siitä, että Lewis käytti yleistä niin sanottua Ramseyn menetelmää erityisesti psykologisten predikaattien määrittelemiseen (Lewis, 1972, s.208–212). Tässä esitetty Ramsey–Lewis-menetelmän esitystapa on peräisin Ned Blockin artikkelista "What is Functionalism" (1980b, s.174) eikä mainitusta Lewisin artikkelista, mutta molemmissa idea on oleellisesti sama.

[42] Kausaaliteoreettisesta funktionalismista tarkemmin ks. esim. Block (1980b, s.174–175) tai tuoreempi oppikirjaversio Kim (2006, s.151–172).

tässä on, että yllä olevassa Ramsey-lauseessa ei esiinny psykologisia termejä, mutta psykologinen termi *kipu* sisältyy Ramsey–Lewis-lauseen vasemmanpuoleiseen osaan. Näin ollen Ramsey–Lewis-lause määrittelee mitä kipu on viittaamatta psykologisiin käsitteisiin. Yleisesti ottaen tällä tavalla voidaan siis määritellä mentaaliset käsitteet käyttämättä mentaalisia käsitteitä, ja tässä mielessä kyseessä on mentaalisten tilojen, ja lopulta koko mentalistisella terminologialla muodostetun teorian T, reduktiivinen määritelmä. Äkkiseltään katsottuna tällainen menetelmä pyyhkii mielenteoriasta kaiken sisällön pois, koska Ramsey-lause ilmaisee vain nimettömien muuttujien välisen abstraktin rakenteen. Saattaa herätä kysymys, missä mielessä menetelmä siis redusoi mielentiloja, koska se ei tarjoa mitään ilmeistä redusoivaa teoriaa tai luonnollista luokkaa, jota mentaaliset tilat vastaavat. Kuitenkin idea tässä nimenomaan on, että redusoivan luokan muodostavat kaikki ne oliot, jotka implementoivat teorian T Ramsey-lauseen ilmaiseman kausaalisen rakenteen.

Yleisesti ottaen funktionaalinen analyysi toimii siten, että teoria T määrittää tarkasteltavien olioiden luokan, johon kuuluvat systeemit, joista T on tosi. Funktionaalinen analyysi puolestaan pilkkoo tarkasteltavat systeemit komponenteiksi, ja selittää funktionaalisesti määritellyn komponentin – olipa kyseessä sitten mielentila, elin, instituutio tai mikä hyvänsä – esiintymisen systeemissä osoittamalla, että komponentilla on tietty kausaalinen asema kokonaissysteemissä (Cummins, 1975, s.741). Mitä tämä sitten tarkoittaa nimenomaan psykologisten teorioiden yhteydessä? Jos T on teoria mentaalisista tiloista ja kausaatiosta, joka nyt tarkoittaa samaa asiaa kuin psykologinen teoria käyttäytymisen etiologiasta, niin mentaalisia ovat ne oliot, joista T on tosi. Teorian muodostaminen alkaa esimerkiksi jonkinlaisista esitieteellisistä arkipsykologisista käsityksistä siitä, mitä mielentiloja mentaalisilla oliolla on ja mikä asema näillä tiloilla on mielen toiminnassa. Käytännössä tämä tarkoittaisi jonkinlaista systemaattista analyysiä praktisen syllogismin kaltaisista väittämistä. Arkipsykologian mukaan ihmisillä on heidän käyttäytymistään ohjaavia haluja ja tarpeita sekä uskomuksia muun muassa seikoista, jotka määrittävät toiminnan mahdollisuuksia. Lisäksi ihmiset saavat tietoa ympäristöstään aistijärjestelmien, päättelyn ja sen sellaisen avulla. Periaatteessa tällaisen analyysin voisi laajentaa koskemaan myös vaikkapa kissoja ja koiria tai mitä tahansa muita olioita, joita teorian muodostaja sattuu pitämään mentaalisina. Joka tapauksessa tämän tyyppinen analyysi psykologisista systeemeistä voisi näyttää esimerkiksi seuraavanlaiselta:



Yllä oleva kaavio tavoittanee karkean arkipsykologisen käsityksen mielen rakenteesta, tai ainakin osan siitä. Tarkastellaanpa seuraavaa praktista syllogismia: "Jos x:llä on nälkä, ja hän uskoo, että nurkan takana olevassa ravintolassa tarjoillaan lounasta, niin hän päättää kävellä nurkan taakse, mennä ravintolaan ja tilata annoksen." Tähän luonnollisesti tulee lisätä esimerkiksi x:n uskovan, että kyseinen ravintola tarjoaa ruokaa, jota hän haluaa syödä, sellaiseen hintaan, jonka hän suostuu maksamaan ja niin edelleen. Yksityiskohtaiset halut ja uskomukset eivät nyt ole oleellisia vaan se, miten halut, uskomukset ja sen sellaiset toimivat yhteispelissä.

Kyseinen praktinen syllogismi työssään voitaisiin täsmällisemmin kuvata yllä olevan funktionaalisen mallin mukaan esimerkiksi seuraavanlaisena prosessina: Kun henkilöllä on nälkä, hänen motivaatiolaatikossaan syntyy tavoite poistaa tämä epämiellyttävä tila. Motivaatiosysteemi lähettää pyynnön toiminnanohjaukseen, jotta se arvioisi tilanteen ja tekisi asialle jotain. Toiminnanohjaus toteaa motivaatiojärjestelmän vaatimuksen edellyttävän ravinnon hankkimista. Seuraavaksi toiminnanohjaus konsultoi aistijärjestelmää, jolla ei ole ratkaisua tarjottavanaan, ja lähettää sitten uskomusjärjestelmään pyynnön raportoida mahdolliset ravinnonlähteet lähialueelta. Uskomusjärjestelmä vastaa, että nurkan takana sijaitsee lounasaikaan palveleva ravintola. Toiminnanohjaus lähettää komentoja motoriselle järjestelmälle, jotta se liikuttaisi raajoja siten, että rannekello ja taskussa lojuvat rahat saadaan näköpiiriin. Nämä komennot ovat kognitiivisen järjestelmän tulostetta. Kun komento pääsee motorisen kontrollin piiriin, representaatiot muuttuvat jollain tapaa muun muassa raajojen ja silmien ohjausliikkeiksi. Mikäli ruumis tekee työnsä, aistijärjestelmä muuntaa verkkokalvoilla syntyvät hermoimpulssit representaatioiksi: *kello on 11.35* ja *rahaa on 10 euroa*. Ensimmäinen näistä representaatioista kertoo suoraan toiminnanohjaukselle, että lounasaika on menossa ja suunnitelmaa voidaan jatkaa. Rahamäärä laitetaan muistiin jatkokäyttöä varten. Toiminnanohjaus komentaa nyt motorista järjestelmää siirtämään ruumiin nurkantaa. Ruumis tottelee. Aistijärjestelmä raportoi ravintolan edessä seisovan kyltin perusteella lounaan maksavan 8 euroa ja 90 senttiä. Toiminnanohjaus konsultoi muistia, eli uskomuslaatikkoa, joka raportoi rahaa olevan 10 euroa, vertaa näitä summia, toteaa rahan riittävän ja komentaa ruumiin ravintolaan.

Kun yllä olevaa selontekoa tarkastelee, saattaa herätä kysymys, mitä funktionalismilla oikeastaan on tarjottavanaan. Esimerkissähän on vain rautalangasta väännetty hypoteettinen malli siitä, mitä mielessä tapahtuu tietyssä tilanteessa. Analyysin tarve tulee kuitenkin ilmeiseksi, kun tarkastelemme olioita, joiden mentaalisuudesta saatamme olla hyvinkin eri mieltä. Onko esimerkiksi koirilla uskomuksia, haluja ja muita psykologisia tiloja? Monet varmaankin vastaisivat tähän myöntävästi, mutta miten on varpusten ja kastematojen laita? Jos halut esimerkiksi ajatellaan jotenkin primitiivisiksi ilmiöiksi, joita on helppo ajatella sisältyvän varpusten psyykeeseen, niin onko kyseisillä linnuilla myös esimerkiksi haaveita tulevaisuudesta? Jos ei, niin onko varpusten pitäminen kirjaimellisesti haluavina olioina jotenkin virheellistä, koska haaveilla tuntuu olevan paljonkin tekemistä halujen kanssa? Entä miten on robottien ja tietokoneiden laita? Ovatko mentaaliset tilat biologisia ilmiöitä, tai muuten vain elollisten olioiden yksinoikeus?

Yllä oleva praktisen syllogismin käsittely on vain yksi esimerkki, miten mielen eri komponentit ehkä keskenään vuorovaikuttavat. Mielentilojen funktionaalisen määrittelyn edel-

lytys on, että mielentiloja määrittelevät teoriat kuvaavat mielen rakenteen periaatteessa täydellisesti, jolloin mielen komponenttien funktiot voidaan täsmällisesti määrittää. Yllä olevassa esimerkkimallissa havaintojärjestelmän funktio on tuottaa uskomuksia ja ohjata käyttäytymistä tarjoamalla toiminnanohjaukselle tietoa ympäristöstä. Motivaatio- ja uskomusjärjestelmä molemmat syöttävät informaatiota toiminnanohjaukseen. Ne eroavat karkeasti ottaen siten, että motivaatiojärjestelmä kertoo miten asioiden tulisi olla, siinä missä uskomusjärjestelmän tarkoitus on kertoa miten asiat ovat. Puhtaasti informaatiovirtaa tarkasteltaessa molemmat yksinkertaisesti syöttävät tietoja toiminnanohjaukseen, mutta näiden komponenttien tulosteiden luonteenomaisten kausaalisten vaikutusten takia kutsumme ensimmäisen järjestelmän tulosteita muun muassa haluiksi ja jälkimmäisen uskomuksiksi. Funktionalismin mukaan mielentiloista ei oikeastaan ole enempää sanottavissa niiden välisten kausaalisten vaikutusten lisäksi. Tämä juuri on teorian ydinväite, josta seuraa, että varpusilla ja roboteilla on mielentiloja tasan siinä tapauksessa, että niiden käyttäytymistä ohjaava systeemi koostuu komponenteista, jotka toteuttavat mielentilojen luonteenomaisia psykologisia funktioita. Sana ”funktio” ei siis nimessä ”funktionalismi” ole matemaattinen käsite, vaan viittaa siihen, että mielentilat identifioidaan niiden psykologisten funktioiden perusteella. Joskin komputationalismiin yhdistettäessä psykologisen teorian Ramsey-lause määrittelee kuvauksen ψ, jota voidaan pitää matemaattisessa mielessä funktiona, joka liittää organismin tilan ja ärsykkeen käyttäytymiseen. Tämä seikka voi aiheuttaa sekaannuksia.

Komponentin funktiota ei kuitenkaan voi yksiselitteisesti samaistaa vaikutukseen. Tämän erottelun teki jo funktionaalisen analyysin isä Carl Hempel todetessaan, että sydämen syke saa aikaan verenkierron mutta myös vaimeaa ääntä. Ensimmäinen on sydämen sykkeen funktio, jälkimmäinen taas ei. (Cummins, 1975, s.742) Näin ollen komponentin funktiota ei voida määritellä tarkastelematta kokonaisjärjestelmää, jossa se esiintyy, koska funktio tarkoittaa oleellisesti komponentin tarkoitusta tuossa kokonaissysteemissä. Esimerkiksi näköjärjestelmä tuottaa jälkikuvia muun muassa kirkkaista valonlähteistä, mutta tämä tuskin on minkäänlainen näköjärjestelmän funktio, koska jälkikuvat eivät palvele mitään ilmeistä tarkoitusperää. Näin ollen jälkikuvien synnyttäminen kuuluu kyllä ihmisen näköjärjestelmän kausaaliseen, mutta ei funktionaaliseen kuvaukseen. Tällä huomiolla on suoria seurauksia mielen funktionalistisen teorian muodostamiselle. Esimerkiksi kissojen mieli toimii varmasti monella tapaa erilailla kuin ihmisten. On kuitenkin mahdollista, että kissojen mielet koostuvat komponenteista, joilla on meidän mentaalisten tilojemme kanssa samanlaiset psykologiset funktiot, vaikkakin osittain eri vaikutukset. Mikäli näin ei kuitenkaan ole, funktionalismista seuraa, että kissoilla ei varsinaisesti ole mielentiloja, ja tällöin niiden pitäminen kirjaimellisesti mentaalisina on virheellistä. Tätä ongelmaa voisi kutsua *käsitteelliseksi ongelmaksi*, koska kyse on siitä, mitkä ovat ne kriteerit, joiden perusteella mentaalisia termejä voidaan erityyppisiin systeemeihin soveltaa.

Toinen ongelma esimerkiksi yllä olevassa esityksessä on, että siinä käytetään huoletta intentionaalisia käsitteitä funktionaalisten komponenttien toiminnan selittämisessä. Havaintojärjestelmä tuottaa uskomuksia, toiminnanohjaus tekee suunnitelmia ja niin edelleen. Kuitenkin funktionaalisen analyysin tavoite on määritellä intentionaaliset tai psykologiset termit ei-intentionaalisin käsittein. Muussa tapauksessahan malli ei tarjoa intentionaalisten systeemien reduktiivista teoriaa, vaan intentionaalisuuden suhteen malli

ikään kuin nostaa itsensä ilmaan. Vaadittava ratkaisu tähän on mielen kausaalisen teorian muodostaminen, joka toisaalta voidaan antaa täysin ei-intentionaalisin termein Ramsey-menetelmän avulla, mutta toisaalta tulkita intentionaalisesti käyttämällä Ramsey–Lewis-lauseita. Esimerkiksi nälän psykologisena funktiona voitaneen pitää ravinnon etsimisen aikaansaamista, mutta tämä ei kuitenkaan ole nälän täydellinen funktionaalinen määritelmä tässä tarkoitetussa mielessä, koska ravinnon etsiminen on intentionaalista toimintaa. Tätä toista ongelmaa kutsuttakoon *tieteelliseksi ongelmaksi*, jossa kyse on intentionaalisuuden reduktiivisesta selittämisestä.

Huomautettakoon, että nämä ongelmat ovat funktionalismin sisäisiä teorianmuodostuskysymyksiä, eivät teorian kritiikkiä. Joka tapauksessa ne ovat ongelmia, joihin funktionalismin on tavalla tai toisella tarjottava vastaus. Ramsey–Lewis-menetelmän tarkoitus on vain osoittaa, että tietynlaisen täydellisen psykologisen teorian avulla intentionaaliset mielentilat voidaan määritellä ei-intentionaalisesti, mutta Ramsey–Lewis-määritelmät saadaan vasta tutkimuksen lopputuloksena, eikä menetelmä sinänsä kerro miten teoria mielen rakenteesta pitäisi muodostaa. Toiseksi ei ole selvää, että tällainen lopullinen teoria sisältäisi tiloja, jotka mitenkään ilmeisesti vastaisivat propositionaalisia asenteita. Siirrytään seuraavaksi analysoimaan näitä ongelmia.

3.3 Homunkulaarinen psykofunktionalismi

Funktionaalinen teoria on esiintynyt monessa muodossa, joista seuraavaksi tarkastellaan analyyttistä, psyko-, ja homunkulaarista funktionalismia. Oletetaanpa, että mielen toiminta voidaan kuvata edellä esitetyn vuokaaviomaisen laatikkomallin mukaisesti. *Analyyttisen funktionalismin* vastaus yllä esitettyyn käsitteelliseen ongelmaan on, että laatikot, eli funktionaaliset komponentit, voidaan nimetä esimerkiksi uskomus- tai halujärjestelmiksi yksinkertaisesti sillä perusteella, että ne toimivat kognitiivisessa systeemissä tavoilla, joita pidämme ominaisina uskomuksille ja haluille (Lewis 1966, s.19–23; 1972, s.213–214). Esimerkiksi *haluiksi* kutsuttaisiin sen laatikon sisältöjä, joita organismi pyrkii toteuttamaan, *peloiksi* niitä, joita se pyrkii kaikin keinoin välttämään ja niin edelleen. Kognitiivisen systeemin komponenttien funktionaalinen määrittely edellyttää melko täsmällistä käsitystä propositionaalisten asenteiden kausaalisesta asemasta mielen toiminnassa. Analyyttiset intuitiomme mielentilojen luonteesta pohjautuvat pitkälti arkipsykologiaan, joten analyyttinen funktionalismi edellyttää, että arkipsykologiset käsityksemme mielentilojen luonteesta ovat melko hienostuneita. Lisäksi välttämätön ehto analyysin läpiviemiselle on, että arkipsykologia on tosi teoria, eli yleiset käsityksemme propositionaalisista asenteista ja niiden asemasta mielen toiminnassa ovat oikeita. Mikäli mieli ei toimi arkipsykologisten intuitioiden sanelemalla tavalla, johtaa analyyttinen funktionalismi eliminativismiin, koska kognitiivisessa systeemissä ei ole rakenteita, jotka voitaisiin samaistaa uskomusten, halujen ja muiden sellaisten kanssa. Eliminativistinen johtopäätös saattaa tietenkin olla myös tervetullut, mutta lähtökohtaisesti ei tunnu kovin järkevältä perustaa tieteellistä mielenteoriaa sen varaan, mitä arkiset käsityksemme mielentiloista sattuvat olemaan. Sitä paitsi kenen analyyttisiä intuitioita pitäisi kuunnella? Psykologien, filosofien vai kadunmiesten ja -naisten? Nyt psykokulttuurin aikakautena sivistyneen maallikon

mielenteoriasta monesti löytyy kaikenlaisia psykologisia käsitteitä, kuten *alitajuiset halut sekä uskomukset* ja *defenssimekanismit*. Pitäisikö esimerkiksi näille ilmiöille löytyä paikka funktionalistisesta mielenteoriasta vaiko ei?

Psykofunktionalismin mukaan mielentilojen määrittelyssä tulee käyttä parasta mahdollista empiiristä teoriaa (Block, 1978, s.270–271). Tällöin funktionaalisen analyysin ei tarvitse lähteä kovin hyvistä mentaalisten ilmiöiden luonnetta koskevista käsityksistä, tai edes alustavasti ottaa kantaa siihen, mitä mielentiloja on olemassa. Psykofunktionalisti suhtautuu näihin avoimina empiirisinä kysymyksinä. Funktionalistisen teorian ja empiirisen psykologian suhde on suurin piirtein sellainen, että psykofunktionalismi tarjoaa psykologialle metateorian, jonka mukaan psykologiset teoriat tulee muotoilla organismin sisäisiin funktionaalisiin komponentteihin perustuvan systeemiteoreettisen analyysin avulla, eikä aikaansa kannata tuhlata esimerkiksi käyttäytymistaipumusten tai psykoneuraalisten identiteettien kanssa puuhasteluun. Empiirinen psykologia puolestaan paljastaa mielen tosiasiallisen rakenteen, ja joko vahvistaa tai korjaa mentaalisiin ilmiöihin liittyviä käsityksiämme. Luonnollisesti psykologisen tutkimuksen täytyy pohjautua joillekin mieltä koskeville esitieteellisesille käsityksille, eli jonkinlainen enemmän tai vähemmän analyyttinen puoli kaikessa tutkimuksessa on aina mukana. Mutta kuten empiirisissä tieteissä yleensä, näiden käsitysten oletetaan muuttuvan ja tarkentuvan tutkimuksen edetessä. Näin psykofunktionalistin ei tarvitse hyväksyä kaikenlaisia arkijärkisiä käsityksiä mielen toiminnasta, vaan teoreetikko voi olla valikoivasti eliminativistinen tiettyjen ilmiöiden suhteen. Tutkimus voi myös ottaa kantaa kyseenalaisten ilmiöiden luonteeseen ja olemassaoloon, kuten defenssimekanismeihin, alitajuisiin haluihin ja sen sellaisiin. Lisäksi tämän tyyppinen tutkimus voi tuoda lisäselvennystä kysymyksiin, joihin arkipsykologiset intuitiomme eivät ole tarpeeksi hienojakoisia vastaamaan. Esimerkiksi onko pakokauhussa, pelossa ja huolessa kyse saman psykologisen tilan erityyppisistä esiintymisistä vai kolmesta eri mielentilasta.

Valinta analyyttisen- ja psykofunktionalismin välillä tarkoittaa vain valintaa, painotetaanko mielentilojen määrittelyssä enemmän analyyttisiä intuitioita, eli mentalististen termien merkitysten analyysiä, vaiko empiirisen psykologian tarjoamia mahdollisuuksia. Huomautettakoon, että analyyttisen funktionalismin kantava ajatus ei suinkaan ole, että psykologista tutkimusta voidaan tehdä yksistään nojatuolista käsin vaan, että jos meillä sattuisi olemaan psykologisen teorian Ramsey-lause nimettömine muuttujineen, tarvitsisimme jonkinlaista propositionaalisten asenteiden merkitysanalyysiä, jonka perusteella mentalistisia nimilappuja voidaan muuttujiin liimata. Teknisemmin sanoen merkitysanalyysiä tarvitaan Ramsey–Lewis-lauseiden muodostamiseen, koska empiirinen tutkimus ei sinänsä kerro, mitkä mielentilat ovat esimerkiksi haluja ilman, että meillä on jokin käsitys niiden luonteenomaisista ominaisuuksista.

Näin laveasti määriteltynä kumpikaan funktionalismin muoto ei määrittele erityisen tarkasti mielenteorian sisältöä, eikä juuri muotoakaan. Käytännössä kuitenkin psykofunktionalismi on ollut vahvassa liitossa kognitiivisen psykologian kanssa. Psykofunktionalistit tyypillisesti katsovat, että mentaalisia prosesseja tutkivan psykologian tutkimuskohde rajoittuu kognitiivisen systeemin sisään. Kognitiiviselle systeemille tunnusomaista on, että se operoi mentaalisilla representaatioilla, jotka puolestaan voivat olla esimerkiksi mieli-

kuvia, ajattelun kielen lauseita tai muita vastaavia. Se, mitä esimerkiksi motorisen systeemin ja raajojen, tai muiden ulokkeiden, välillä tapahtuu, ei yleensä katsota kuuluvan psykologian alaan. Vastaavasti aistijärjestelmän katsotaan omalta osaltaan määrittävän psykologian tutkimuskohteen rajoja. Se, miten aistinjärjestelmä muuntaa esimerkiksi verkkokalvolle tulevia ärsykkeitä mentaalisiksi representaatioiksi, ei kuulu psykologian, vaan aistinelinten fysiologian tutkimukseen. Aistielinten kognitiiviselle systeemille tarjoamat tulosteet puolestaan ovat psykologisia ilmiöitä, nimittäin havaintoja. (Fodor, 1983, s.40–43.)

Mikään psykofunktionalismissa ei pakota yllä mainittuihin teoreettisiin sitoumuksiin, mutta jos mielentilojen ajatellaan olevan representaationaalisia ja mentaalisten prosessien representaatioiden käsittelyä, niin teoria kognitiivisesta systeemistä on melko luonnollinen valinta mielenteorian ensisijaiseksi kohteeksi. Kognitiivisen psykologian pioneeri Ulric Neisser kirjoitti teoksensa *Cognitive Psychology* ensimmäiseen lukuun kuvaavasti:

> As used here, the term "cognition" refers to all the processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used. [...] Such terms as *sensation*, *perception*, *imaginary*, *retention*, *recall*, *problem solving*, and *thinking* among many others, refer to hypotetical stages or aspects of cognition. [...] Given such a sweeping definition, it is apparent that cognition is involved with everything a human beign might possibly do; that every psychological phenomenon is a cognitive phenomenon. (Neisser, 1967, s.4)

Siinä missä Neisser katsoi kognitiivisen psykologian koskevan nimenomaan aisti-informaation käsittelyä, käyttöä ja muistiin tallentamista, myöhemmin kognitiivinen psykologia on laajentunut koskemaan myös muun muassa tunteita, motivaatioita, luovuutta ja pääpiirteissään kaikkia mahdollisia yksilötason psykologisia ilmiöitä. Lienee sanomattakin selvää, ettei tämä tarkoita kaiken psykologisen tutkimuksen olevan kognitiivisen systeemin tutkimista, mutta jos mielentilat ja prosessit määritellään kognitiivisen systeemin tiloiksi ja prosesseiksi, tarjoaa kognitiivinen psykologia ontologisen perustan periaatteessa kaikelle psykologialle.

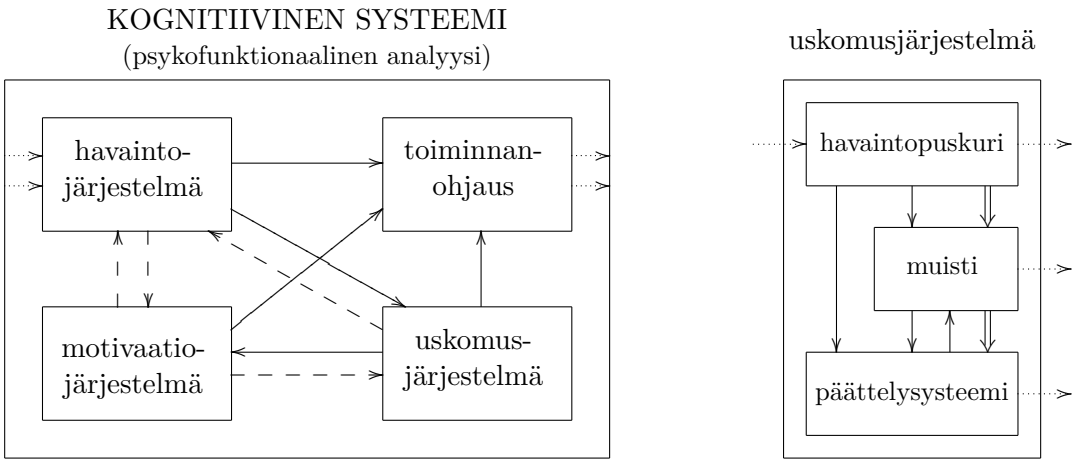
Homunkulaarinen funktionalismi on oikeastaan eräs psykofunktionalismin muoto, jolle ominaista on ajatus, että kognitiivisen systeemin toimintaa tulee analysoida tavallaan kerroksittain. Mielen rakenteesta voidaan tehdä vuokaaviomalli, mutta homunkulaarisessa mallissa laatikoiden sisältä löytyy lisää laatikoita. Toisin sanoen analyysiä jatketaan suorittamalla vastaavanlainen funktionaalinen analyysi ensimmäisessä vaiheessa löydetyille komponenteille. Lisäksi homunkulaariselle analyysille ominaista on, että nämä komponentit katsotaan yleensä suhteellisen autonomisiksi. Komponenttien sisäinen toiminta ei juurikaan riipu siitä, mitä kognitiivisessa systeemissä muualla tapahtuu. Kantava ajatus tässä on, että funktionaalinen analyysi jäljittelee jotakuinkin kognitiivisen psykologian teorianmuodostusprosessia.[43]

[43]Ks. (Fodor, 1968b), (Dennett, 1975) ja (Lycan, 1981, s.26–31).

Esimerkiksi kasvojen tai puheen tunnistusta, työmuistin toimintaa, päätöksentekoa tai vastaavaa tutkiva kognitiivinen psykologi joutuu tyypillisesti eristämään tutkimansa ilmiön muusta kognitiosta, ja ilmiön selitys voi alkaa kuvaamalla miten se syntyy intentionaalisten alasysteemien yhteistoiminnasta. Nimitys *homunkulaarinen* tulee siitä, että analyysin mukaan intentionaaliset systeemit itse koostuvat intentionaalisista systeemeistä (*Homunculus*, lat. *pieni ihminen*). Vastaavanlainen metodologia on perinteisesti ollut kovassa käytössä myös tekoälytutkimuksessa, jossa yleensä pyritään rakentamaan monimutkaisia systeemeitä yksinkertaisemmista, verrattain autonomisista, aliohjelmista. Esimerkiksi tekoälypioneeri Marvin Minskyn *The Society of Mind* (1985) lienee kuuluisimpia homunkulaarisia mielen malleja. Tyypillisesti ohjelmoijat eivät – ainakaan käytännön tasolla – ole pyrkineet mallintamaan koko kognitiivista systeemiä kerrallaan, vaan jotain sen osia, kuten ongelmanratkontaa, hahmontunnistusta ja niin edelleen. Tarve pilkkoa monimutkaiset ohjelmat autonomisiksi aliohjelmiksi oli ilmeistä jo tekoälyn alkuaikoina (Newell, 1962, s.10–11). Kyse ei varsinaisesti ole tekoälyn ominaisluonteesta, vaan käytännön välttämättömyys lähes kaikessa ohjelmoinnissa.

Tällainen analyysi ei heijasta pelkästään kognitiotieteiden metodologiaa vaan myös tuloksia. Tutkimuksen empiirisesti suuntautuneilla aloilla – niin kognitiivisessa psykologiassa (Baars, 1988, s.42) kuin neurotieteissäkin (Kosslyn & Smith, 2000, s.961–963) – on varsin yleisesti hyväksytty käsitys, että monimutkaisemmat kognitiiviset prosessit ovat seurausta yksinkertaisempien ja jokseenkin itsenäisten prosessien yhteistyöstä. Siinä missä psykofunktionalismi siis on vastaus funktionalismin käsitteelliseen ongelmaan, homunkulaarinen funktionalismi on enemmänkin vastaus tieteelliseen ongelmaan, mistä lisää myöhemmin.

Alla vasemmanpuoleinen kaavio on eräiden psykologian tulosten perusteella päivitetty psykofunktionalistinen malli aiemmin esitetystä kognitiivisen systeemin hahmotelmasta. Oikeanpuoleinen kaavio täydentää esimerkkiä tarjoamalla homunkulaarisen funktionaalisen analyysin uskomusjärjestelmän toiminnasta. Yhtenäiset nuolet ovat aiemmin selitettyjä arkipsykologiaa mallintavia käsityksiä mielentilojen välisestä kausaatiosta. Kognitiivisen psykologian termein ilmaistuna nämä nuolet havainnollistavat miten informaatio virtaa systeemissä: havaintojärjestelmä tuottaa uskomuksia, jotka vaikuttavat toiminnan ohjaukseen ja niin edelleen. Katkonuolet kuvaavat alla tarkemmin selostettuja psykologisia ilmiöitä.



New Look -nimellä kulkenut havaintopsykologian suuntaus on eräs edellisessä luvussa mainituista aloista, jotka pohjustivat kognitiivisen psykologian syntyä. Harvardissa 40-luvun lopulla vaikuttaneet tämän koulukunnan edustajat Jerome Bruner ja Leo Postman osoittivat, että havainnot eivät ole yksiselitteisesti ärsykkeen määräämiä, vaan havaintosisällöt määräytyvät myös pitkälti havaitsijan odotusten ja uskomusten perusteella. Yllä olevassa kaaviossa tätä kuvaa nuoli uskomuslaatikosta havaintojärjestelmään. Ehkä hieman yllättäen myös havainnoitsijan arvomaailma sekä sosiaalinen asema vaikuttavat havaintoprosesseihin. Esimerkiksi sanat, jotka kuvaavat havaitsijan arvoja, kyetään tunnistamaan muita nopeammin. (Bechtel et al., 1998, s.20–21.) Näin ollen myös motivaatiojärjestelmä vaikuttaa havaintotapahtumiin. Nuoli motivaatiojärjestelmästä uskomuksiin viittaa 50-luvulla kehitettyyn kognitiivisen dissonanssin teoriaan, jonka mukaan ihmisten ristiriitaiset halut ja uskomukset aiheuttavat tarpeen vähentää tätä ristiriitaa, joko pyrkimällä hallitsemaan haluja tai sitten muutamalla uskomuksia (Cooper, 2007, s.4–9). Klassinen esimerkki tästä on tupakoitsija, joka haluaa säilyttää terveytensä. Hän voi joko koittaa lopettaa tupakoinnin tai sitten yksinkertaisesti ryhtyä uskomaan, ettei se syystä tai toisesta ole hänelle merkittävä terveydellinen riski. Viimeinen nuoli havainnoista motivaatiojärjestelmään taas viittaa havaintojen kykyyn vaikuttaa suoraan haluihin vaikuttamatta uskomuksiin. Vanha viisaus mainonnassa on yhdistää tuotemerkki haluttaviin tai miellyttäviin asioihin, mutta asenteisiin voidaan vaikuttaa myös pelkällä altistamisella. Sosiaalipsykologi Robert Zajonc havaitsi 60-luvulla, että koehenkilöt alkavat suosia hahmoja, joille he ovat altistuneet, olivatpa ne esimerkiksi henkilöitä, symboleita tai melodioita, ilman, että heillä on mitään uskomuksia tai edes muistikuvia niistä (Zajonc, 1986, 1980).

Nämä empiiriset havainnot siis kertovat jotain epäilmeistä siitä, miten uskomukset, havainnot ja motiivit kognitiivisessa systeemissä toimivat. Psykofunktionalismin kannalta tämä tarkoittaa, että nämä tulokset osaltaan määrittelevät uudelleen, mitä mainitut mentaaliset ilmiöt oikeastaan ovat. Teoreettista analyysiä edellyttävä kysymys lisäksi on, mitkä näistä vuorovaikutuksista ovat varsinaisia psykologisia funktioita ja mitkä epätarkoituksenmukaisia psykologisia vaikutuksia. Esimerkiksi tarkoituksenmukaista lienee, että uskomukset vaikuttavat haluihin, mutta tuskin, että halut muokkaavat uskomuksia. Myöskin Zajoncin huomaama altistusvaikutus vaikuttaa äkkiseltään melko epäfunktionaaliselta. Toisaalta tuttu usein korreloinee turvallisen kanssa, joten ilmiö saattaa olla hyvinkin tarkoituksenmukainen.

Kiinnitetään sitten huomiota kaavion oikeanpuoleiseen laatikkoon. Homunkulaarisen analyysin mukaan uskomusjärjestelmä on oma erillinen alajärjestelmänsä, jonka toiminta voidaan – ja pitäisi – myös selittää funktionaalisen analyysin avulla. Uskomukset koostuvat muistikuvista, opituista asioista ja ylipäätään ympäröivää maailmaa koskevista käsityksistä. Kaavio kuvaa, miten uskomusjärjestelmä saattaisi toimia. Systeemin syöte tulee aistijärjestelmältä väliaikaiseen muistiin (kaaviossa ”aistipuskuri”), joka pitää kirjaa ympäristön tapahtumista. Osa tästä syötteestä tallennetaan pitkäkestoiseen muistiin ja osa tarpeen vaatiessa jatkokäsitellään päättelysysteemissä. Kun vaikkapa toiminnanohjaus kysyy muistijärjestelmältä tietoa, niin tämä esimerkkisysteemi toimii seuraavasti: Aluksi tarkastetaan onko tieto saatavissa suoraan aistien avulla ja jos, niin kyselyn vastaus tulostetaan suoraan havaintopuskurista. Mikäli ei, niin havaintopuskuri siirtää kysymyksen

eteenpäin pitkäkestoiselle muistille, joka tulostaa vastauksen mikäli mahdollista.[44] Paksu nuoli oikealla kuvaa tätä kontrollin siirtymistä systeemin komponenteilta toisille. Jos tarvittavaa tietoa ei löydy muististakaan, kysely siirtyy edelleen päättelysysteemille, joka pyrkii tuottamaan vastauksen nojaten aistipuskurin ja pitkäkestoisen muistin sisältämiin tietoihin. Mikäli vastaus löytyy, se tulostetaan toiminnanohjaukselle ja mahdollisesti tallennetaan pitkäkestoiseen muistiin. Mikäli vastausta ei kyetä myöskään päättelemään, uskomusjärjestelmä tulostaa tämän kyselyn vastaukseksi.

Huomion arvoista tässä ei niinkään ole, miten uskomusjärjestelmä oikeasti toimii vaan homunkulaarisen analyysin kerroksittaisuus. Kognitiivisen systeemin kannalta ei ole väliä, mitä uskomusjärjestelmän sisällä tapahtuu, vaan ainoastaan sillä, että kun kyseiselle alasysteemille syötetään kysely *päteekö A?*, niin se yleensä tulostaa kohtuullisen luotettavan vastauksen. Myöskään uskomusjärjestelmän kannalta muiden kognitiivisen systeemin komponenttien sisäisellä toiminnalla ei ole merkitystä. Muut komponentit ovat uskomusjärjestelmästä käsin katsottuna sen toimintaympäristö, jonka kanssa se vuorovaikuttaa syötteiden ja tulosteiden avulla. Luonnollisesti komponenttien funktionaalisen analyysin tulisi olla uskollinen kokonaissysteemin analyysille. Esimerkiksi yllä oleva uskomusjärjestelmän laatikkomalli ei ole täydellinen, koska se jättää huomiotta aiemmin mainitun motivaatiojärjestelmän vaikutuksen uskomuksiin.

Homunkulaarinen analyysi ei johda kehälliseen selitykseen, jossa intentionaalisuutta selitetään intentionaalisuudella. Tämä johtuu siitä, että alasysteemien ajatellaan olevan selitettävää kokonaissysteemiä yksinkertaisempia, siis jotenkin tyhmempiä tai vähemmän intentionaalisia, otuksia. Kun analyysin toisen vaiheen komponentit edelleen analysoidaan funktionaalisiin palasiin, nämä palaset taas ovat kokonaiskomponentteja yksinkertaisempia ja niin edelleen. Kantava oletus on, että analyysissä lopulta saavutetaan taso, jossa komponenttien toiminta on niin yksinkertaista, ettei niiden toiminnan selittäminen enää edellytä intentionaalista selittämistä. (Dennett, 1975, s.80–81) Ensinnäkin näin intentionaalisuus lopulta saa reduktiivisen selityksen, ja toiseksi analyysin pohjimmainen taso osaltaan määrittää kognitiivisen psykologian tutkimuskohteen rajat määrittämällä kognition alusrakenteen. Mikäli analyysiä jatkettaisiin, siirryttäisiin kognitiivisen teorian ulkopuolelle implementaatioteoriaan, ja miten kognitiivisen koneiston pohjataso on implementoitu, on neurofysiologian eikä psykologian ongelma.

Esimerkiksi jos edellä olevan uskomusjärjestelmän analyysiä jatkettaisiin, päättelysysteemin toiminta ilmeisesti vaatisi jonkinlaista selitystä. Selitys voisi olla komputationalistinen kuvaus, jonka mukaan päättelysysteemi operoi lauseiden kaltaisilla mentaalisilla representaatioilla, joihin se soveltaa logiikan päättelysääntöjä heuristisen periaatteiden mukaisesti. Päättelysysteemi olisi siis eräänlainen mentaalisilla representaatioilla operoiva virtuaalikone, joka sisältää päättelysääntöjen muodossa kalkyylin, sekä heuristiikkojen muodossa proseduraaliset ohjeet. Miten nämä komponentit sitten toimivat, eli miten esimerkiksi konjunktion tuontisääntö taas on systeemissä toteutettu? Tähän kysymykseen kognitiotieteilijän ei välttämättä enää tarvitse vastata. Hän voi todeta päättelysääntöjen

[44] Vaikka esimerkki on keksitty, kognitiivisen koneiston tiedonhaku vaikuttaisi oikeasti käyttävän hyväksi ensisijaisesti havaintojärjestelmää. Etu tästä on, että tiedonhaku sekä nopeutuu että käyttää vähemmän työmuistia. (Clark, 2008, s.12–13,120–121.)

soveltamisen olevan psykologisesti primitiivinen operaatio, ja vastaus edellyttäisi selontekoa mentaalisten representaatioiden ja päättelyprosessien neuraalisesta implementaatiosta, mikä ei ole psykologian alaan kuuluva kysymys. Kyse ei ole mielivaltaisesta tieteenalojen työnjaosta, vaan siitä, että kuten aiemmin psykoneuraalisen identiteettiteorian yhteydessä on koitettu osoittaa, neurotieteiden avulla on mahdotonta muotoilla psykologian kannalta mielenkiintoisia ja välttämättömiä yleistyksiä. Toisaalta operaation primitiivisyydelle ei ehkä ole teoreettisia tai empiirisiä perusteita, mutta kysymys voi kuitenkin olla aiheellista ohittaa. Nimittäin joka tapauksessa tiedetään, että perin tyhmät systeemit, kuten diodit ja transistorit, joiden toiminta voidaan selittää ilman intentionaalisia käsitteitä, voivat suorittaa esimerkiksi loogisia operaatioita. Näin ollen vaikka ongelma kuuluisikin psykologeille, kyseessä ei välttämättä ole kovin mielenkiintoinen mysteeri.

Psykohomunkulaarinen malli mielestä muistuttaa edellisessä luvussa mainittua virtuaalikoneiden pinoamista (s.44). Mallin primitiivinen taso, joka paljastaa kognition alkeisoperaatiot, on analoginen tietokoneiden ensimmäisen kertaluvun virtuaalikoneiden kanssa. Primitiivinen taso implementoi erityyppisiä muisti- ja päättelysysteemeitä ja muita vastaavia, jotka edelleen implementoivat esimerkiksi uskomus- ja havaintojärjestelmät, jotka lopulta yhteispelissä tuottavat täysverisen kognitiivisen systeemin. Tietokoneissa ensimmäisen kertaluvun virtuaalikone on konekielen käsittelijä, jonka toiminta voidaan selittää epäintentionaalisesti koneen fysikaalisella rakenteella. Vastaavasti kognitiivisen systeemin perustaso on mentaalisten representaatioiden käsittelijä, jonka toiminta selittyy fysikaalisella implementaatiolla, joka ihmisten tapauksessa on neurobiologinen. Aivojen toimintaa ei tällä hetkellä tunneta perinpohjaisesti, mutta asiaan ei vaikuttaisi liittyvän sen kummempia käsitteellisiä mysteereitä, kuin esimerkiksi tietokoneiden tai vatsan toimintaan. Jos mielen tietokonemetafora otetaan kirjaimellisesti, niin kognitiivisen systeemin perustataso itse asiassa on ensimmäisen kertaluvun virtuaalikone: mentaaliset representaatiot ovat symboleja tai symbolirakenteita, ja perustatason funktionaalisten komponenttien operaatiot proseduraalisia ohjeita. Tämä tarkalleen ottaen on se piste, missä funktionaalinen analyysi ja komputationaalinen mielenteoria hitsautuvat yhteen muodostaen kognitiivisen mielenteorian sen nykyisessä muodossaan.

3.4 Selityksen tasot ja psykologisten tilojen identifiointi

Äkkiä katsottuna psykofunktionalismi näyttäisi lankeavan samaan syntiin, mistä se itse syyttää sekä psykoneuraalista identiteettiteoriaa että konetilafunktionalismia. On luonnollista ajatella, että empiirinen psykologia, johon psykofunktionalismi nojaa mielentilojen määrittelyssä, on ainakin lähtökohtaisesti ihmismielen tutkimusta. Edellä kuvattu vuokaaviomalli on ollut vaikutusvaltainen teorianmuodostusmenetelmä kognitiivisessa psykologiassa. Tämänlaisia malleja on käytetty kuvaamaan ihmismielen toimintaa, mutta periaatteessa ne soveltuvat minkä tahansa olion kognitiivisen systeemin kuvaamiseen. Mikäli mallin tarkoitus ei ole vain luonnehtia erilaisten olioiden psykologiaa, vaan myös määritellä, mitkä mentaalisten tilojen luontaiset ominaisuudet yleisesti ottaen ovat, ongelmia syntyy heti. Nimittäin mikäli analyysi yksikäsitteisesti määrittelee mielentilat jonkin psykologisen teorian perusteella, näitä tiloja ei voi olla olioilla, joista kyseinen teoria on

epätosi. Toisin sanoen psykofunktionalismista näyttäisi seuraavan, että vain psykologisesti ihmistenkaltaiset oliot ovat mentaalisia, olettaen tietysti, että funktionaalinen analyysi koskee ensisijaisesti meidän mielentilojamme. Ongelman ydin on lopulta kaikenlaiseen funktionalismiin sisäänrakennettu holistinen ajatus, jonka mukaan mielentilat tulee määritellä suhteessa niiden muodostamaan kokonaissysteemiin.

Jo aiemmin tavatuilla meritursailla on varmasti hyvin erilainen kokonaispsykologia kuin ihmisillä. Kuitenkin myös ne tarvitsevat ja etsivät ruokaa syödäkseen, pakenevat uhkaavia otuksia ja niin edelleen. Vaikuttaisi siis mielekkäältä ajatella niiden voivan olla muun muassa nälissään ja peloissaan, kuten ihmisetkin. Kuitenkaan pelkkä näennäisesti samankaltainen käyttäytyminen ei oikeuta tekemään tällaista johtopäätöstä. Me ja meritursaat olemme niin erilaisia, että ehkä tarkemmin katsottuna tursailla ei olekaan varsinaisia mielentiloja. Psykofunktionalismin etu on, että se ratkaisee käsitteellisen ongelman olioiden mentaalisuudesta, kun intuitiomme ei luotettavasti osoita suuntaan eikä toiseen. Näin ollen teorian mahdollisesti tarjoama kielteinen kanta meritursaiden mielentiloista pitäisi ehkä laskea sen kunniaksi. Tietenkään psykofunktionalismista ei välttämättä seuraa, ettei tursailla voisi olla jonkinlaisia mielentiloja, mutta teoriasta kyllä seuraa, ettei niillä ole ainakaan inhimillisiä arkipsykologisia tiloja. Tällöin kuitenkaan meillä ei vaikuttaisi olevan käsitteitä, joiden nojalla voisimme ymmärtää tursaita mentaalisina olioina. Toisin sanoen mikäli emme voi soveltaa tyypillisiä intentionaalisia käsitteitä, on epäselvää, mitä olion mentaalisuus edes tarkoittaisi.

Tähtäintä on siis hyvä siirtää vähän lähemmäs ja tarkastella vaikkapa kissoja. Useimmat varmaankin ajattelevat, että kissat voivat olla nälissään, mutta ne eivät voi esimerkiksi pelätä maailmantalouden romahtamista. Ne tuskin edes kykenevät hahmottamaan maailmantalouden kaltaista monimutkaista järjestelmää, saatika sitten sen romahtamisen kauhuja. Mikäli tämä kertoo hyvin oleellisesta erosta ihmisten ja kissojen psykologiassa, niin kuten edellä, tästä seuraa, että joko kissat itse asiassa eivät voikaan olla esimerkiksi nälissään tai sitten psykofunktionalismi on epätosi. Tämä mielestäni alkaa jo olla teorian kannalta melko vakavaa. Toisaalta tästäkin hankalasta tilanteesta saattaa olla ulospääsytie. Jotkut filosofit katsovat olevan psykofunktionalismista riippumattomia syitä olettaa, ettei kissoilla itse asiassa ole varsinaisia intentionaalisia tiloja. Tämä johtuu siitä, että niillä ei ole kieltä. Ajatus siis on, että intentionaalisuus edellyttää kieltä, eikä toisin päin. Tällaisista näkemystä on kannattanut ainakin Donald Davidson (Davidson, 2001, s.126–127) ja ilmeisesti Wittgenstein (Gillett, 1997, s.341). Hieman vastaavaan ajatukseen törmättiin jo johdannossa Sellarsin yhteydessä. Jos kaikki mentaaliset tilat, nälkä mukaan lukien, ovat intentionaalisia, niin tämän näkemyksen mukaan kissat eivät voi olla nälissään. En oikein osaa sanoa onko tämä johtopäätös uskottava, mutta luulisin, että se joka tapauksessa sotii monien funktionalistien intuitioita vastaan. Olipa miten hyvänsä, ilmeisesti kissatkaan eivät siis välttämättä kaada psykofunktionalismia. Jotta tulisi selväksi miten syvälle tämä ongelma menee, lienee syytä turvautua ajatuskokeeseen, jossa verrokin intentionaalisuutta ei voi kyseenalaistaa. On aika kaivaa taas marsilaiset esiin.

Ned Block ja Sidney Shoemaker ovat kumpikin esittäneet argumentin, joka muistuttaa aiemmin tarkastelemaamme identiteettiteoriaa vastaan kehitettyä marsilaistarinaa. Tarinahan meni niin, että joku päivä planeetallemme ilmestyy kopla marsilaisia, jotka päällisin puolin käyttäytyvät kuin me, eli käyvät elokuvissa, kirjoittelevat kirjoja, pohtivat

mentaalisten tilojen identifikaatiokriteereitä ja niin edelleen. Näyttää siis vahvasti siltä, että marsilaiset muun muassa uskovat, haluavat ja iloitsevat siinä missä mekin. Block ja Shoemaker tarkastelivat tilannetta, jossa marsilaiset ja me eroamme psykofunktionaalisen hienorakenteemme suhteen. Jos psykofunktionalistinen analyysi propositionaalisista asenteista edellyttää viittaamista systeemin koko funktionaaliseen rakenteeseen, nämä hienovaraiset erot riittävät johtopäätökseen, että marsilaisilla itse asiassa ei olekaan samoja mielentiloja kuin ihmisillä. (Block 1978, s.310–311; Shoemaker 1981, s.281–282)

Oletetaanpa esimerkin vuoksi, että meidän ja marsilaisten erot löytyvät muistijärjestelmästä, joiden eroavaisuuden paljastaminen vaatii tarkkoja psykologisia koejärjestelyjä. Ihmisen lyhytkestoinen muisti toimii suurinpiirtein siten, että muistiin mahtuu noin seitsemän muistettavaa yksikköä kerrallaan.[45] Kun työmuistista koitetaan palauttaa olioita mieleen, esimerkiksi kysyttäessä esiintyykö joku tietty objekti juuri esitetyssä listassa, lyhytkestoisen muistin sisältö ilmeisesti käydään läpi seriaalisesti ja tyhjentävästi (Sternberg, 1966, 1969). Tämä on hyvin kummallinen tapa käyttää muistia. Jos muisti käydään läpi seriaalisesti, eli yksikkö kerrallaan, muistihaun luulisi loppuvan etsittävän objektin löydyttyä. Ilmeisesti näin ei kuitenkaan tapahdu. Oletetaanpa, että marsilaisten muisti toimii vähän fiksummin, eli muistihaku heidän kohdallaan ei ole tyhjentävä, tai tapahtuu esimerkiksi rinnakkaisena prosessina. Mikäli Block ja Shoemaker ovat oikeassa, psykofunktionalismi sitoutuisi tällöin väittämään, että marsilaisilla ei ole uskomuksia ja haluja eivätkä ne voi esimerkiksi olla kirjaimellisesti janoissaan tai nälissään, mikä kieltämättä olisi absurdia. Tämän kaltaista ongelmaa, jossa analyysi kieltää uskomusten, halujen ja muiden vastaavien mielentilojen olemassaolon olioilta, joilla niitä melko selvästi pitäisi olla, on tapana kutsua *lajisovinismiksi*.

On syytä kiinnittää huomiota siihen, mikä lajisovinismin ongelma oikeastaan on. Jos mielenteoriasta seuraa, ettei tietyillä olioilla ole mielentiloja, vaikka intuitiivisesti katsoisimme niitä niillä olevan, joudumme valitsemaan kahden vaihtoehdon väliltä: teoriaa pidetään kyseisiä intuitioita luotettavampana ja sen seuraukset hyväksytään, tai sitten teoria hylätään sillä perusteella, ettei se ilmiselvästi kykene poimimaan mentaalisten olioiden luokkaa. Jälkimmäinen johtopäätös tarkoittaa, että teoria on kykenemätön selittämään, mikä mentaalisia olioita yhdistää ja mikä niistä tekee mentaalisia. Tällöin teoria ei koske mentaalisuutta eikä mielentiloja, vaan jotain muuta, jos loppujen lopuksi yhtään mitään. Lajisovinismiargumenttia käytettiin hylkäämään sekä psykoneuraalinen identiteettiteoria että konetilafunktionalismi, ja näyttäisi, että sama kohtalo saattaa kohdata psykofunktionalismia.

Austen Clark on esittänyt ilmeisen tien ulos tästä ongelmasta. Se on yksinkertaisesti määritellä psykofunktionaaliset analyysit kattamaan kaikki periaatteessa mahdolliset psykologiset systeemit. Tämä tapahtuu muodostamalla jokaiselle erityyppiselle mentaaliselle oliolle omat Ramsey-lauseensa, jolloin mielentilat määritellään Lewis–Ramsey-menetelmän

[45] Tässä yhteydessä ei tarvitse häiriintyä siitä, mitä ”yksiköllä” tarkoitetaan. Halutessaan lukija voi tutustua esimerkiksi George Millerin artikkeliin ”The Magical Number Seven” (1956), jossa tämä tulos esitettiin alunperin. Millerin havainnolla on käsiteltävän asian kannalta historiallista merkitystä, koska se on yksi merkittävimmistä 1950-luvun tuloksista, jotka käynnistivät niin sanotun kognitiivisen vallankumouksen psykologiassa (Bechtel et al., 1998, s.37–38.).

avulla lajikohtaisesti. (Clark, 1986, s.539–542) Teknisesti ottaen funktionaaliset määritelmät esimerkiksi janolle eivät tällöin olisi muotoa:

> On olemassa funktionaalinen määritelmä M siten, että jokaisella lajilla L tila t on *jano*, jos ja vain jos t toteuttaa määritelmän M lajin L edustajissa,

vaan

> Jokaista lajia kohden L on olemassa funktionaalinen määritelmä M siten, että tila t on *jano*, jos ja vain jos t toteuttaa määritelmän M lajin L edustajissa.

”Laji” tulee tässä yhteydessä ymmärtää ensisijaisesti psykologisena lajityyppinä, eli oliot kuuluvat samaan lajiin, mikäli ne toteuttavat saman psykologisen teorian.

Tällainen menetelmä selvästi vaatisi jonkinlaisen metateorian, joka selventää, mitkä oliot ovat mentaalisia ja miten näiden olioiden kohdalla mikäkin mielentila määritellään. Ongelma tässä on, että psykofunktionalismin nimenomaan pitäisi olla tuo metateoria. Kuten Ned Block on huomauttanut, jos psykofunktionalismi edellyttää jotain erillistä mielentilojen teoriaa, niin meidän kai tulisi valita se mielenteoriaksi psykofunktionalismin sijaan (Block, 1978, s.312). Sitä paitsi fysikalisti tai identiteettiteorian kannattaja voisi käyttää täsmälleen samanlaista menetelmää. Hän voisi ottaa analyysinsä kohteeksi mentaaliset oliot laji kerrallaan ja yksinkertaisesti luetella mitkä fysikaaliset tilat kullakin vastaavat mitäkin mielentilaa.[46] Aiemmin tällaista fysikalismia vastaan esitetty syyte oli, että mikäli teoria ei pysty muodostamaan selontekoa siitä, mikä mentaalisia systeemeitä fysikaalisesti yhdistää, se ei tarjoa reduktionistista analyysiä mielentiloista, vaan ainoastaan periaatteessa äärettömän listan sinänsä tosia psykofyysisiä korrelaatioväitteitä. Nyt jos psykofunktionalisti joutuu suorittamaan analyysinsä vastaavasti laji kerrallaan, ei hän pärjää fysikalistia paremmin tässä suhteessa millään tavalla. Lyhyesti sanottuna siis, jos tilanne on tämä, väitteet ”mielentilat ovat fysikaalisia tiloja” ja ”mielentilat ovat funktionaalisia tiloja” saattavat molemmat olla jossain mielessä tosia, mutta funktionalistiset teoriat eivät määrittele mielentiloja sen enempää kuin fysikalistiset.

On tietenkin mahdollista, että näistä vaikeuksista huolimatta psykofunktionalismi osoittautuu lupauksensa lunastavaksi elinkelposeksi hankkeeksi. Yleisesti ottaen tiede toimii siten, että tutkimus alkaa joidenkin esiteoreettisesti ymmärretytjen ilmiöiden tarkastelusta ja tähtää niitä kuvaavaan enemmän tai vähemmän systemaattisen teorian muodostamiseen. Tämän jälkeen saatetaan huomata, että teoriaa voidaan soveltaa ilmiöihin, joita alunperin ei ajateltu kuuluvan sen alaan. Tässä mielessä tieteenalan tutkimuskohdetta ei tarvitse – eikä aina voi – määritellä tutkimuksen aluksi, vaan se selviää tutkimuksen aikana tai jopa sen jälkeen. Luontainen osa teorianmuodostusta on sen soveltuvuusalan selventämistä, eli sen tutkimista, mitä teoria itse asiassa koskee. (Fodor, 1968a, s.9–12) Lajisovinismin ongelma ei siis välttämättä kieli psykofunktionalismia perinjuurin myrkyttävästä käsitteellisestä ongelmasta, vaan yksikertaisesti normaalista teorianmuodostuksen vaiheesta. Vetoaminen tulevaisuuden tieteen mahdollisuuksiin ratkaista esillä olevat käsitteelliset ongelmat on asiallista argumentointia johonkin rajaan asti. Tässä tapauksessa ongelma koskee kuitenkin psykofunktionalismin ydintä, eli sen periaatteellista mahdollisuutta tarjota pätevä yleinen analyysi mielentiloista, joten se tuntuisi vaativan ainakin jonkinlaisen lupaavan ratkaisuehdotuksen.

[46] Ks. esim. (Kim, 1992, s.19–26).

Austen Clark on ehdottanut tähänkin ongelmaan seuraavaa, myös melko ilmeistä, ratkaisua. Lähdetään siitä, että jokin funktionaalisista komponenteista ja niiden välisistä kytkennöistä muodostuva vuokaaviomalli on oikenlainen kuvaus ihmisten psykologiasta, joku toinen taas kissojen psykologiasta ja niin edellen. Nyt teoreetikko voisi eristää ihmismielen vuokaaviosta vaikkapa janoon liittyvät prosessit poimimalla kaaviosta ne komponentit, joissa tapahtuu systemaattisia muutoksia veden puutteen ja nauttimisen yhteydessä, sekä kytkennät, jotka ilmaisevat yhteyttä homeostaasin sekä veden etsimisen ja nauttimisen välillä. Clark olettaa, että esimerkiksi jano voitaisiin tällä tavalla määritellä viittaamatta ihmisten kokonaispsykologiaan, koska kyseinen ilmiö muodostaa oman alasysteeminsä. Samoin esimerkiksi kissan psykologiaa kuvaavasta vuokaaviosta luultavasti löytyy vastaavanlainen alasysteemi. Mikäli vaikkapa janoa vastaava alasysteemi on eristettävissä, voisi myös olettaa, ettei sen kuvauksessa nouse esiin esimerkiksi pörssikursseja koskevat uskomukset, joten tällaiset seikat eivät ole oleellisia sen kannalta, voiko olio olla janoissaan vai ei. Näin ollen riippumatta kissojen ja ihmisten kokonaispsykologian välisistä eroista, vuokaavioiden tiettyjen samankaltaisten alirakenteiden perusteella on oikeutettua sanoa, että kissat voivat olla janoissaan siinä missä ihmisetkin. (Clark, 1986, s.546–548.)

Vastaava analyysi pätee myös marsilaisten tapauksessa. Blockin ja Shoemakerin marsilaisargumenteissa ei sen tarkemmin määritellä, millä tavalla he ajattelevat muukalaisten eroavan meistä, mutta jos marsilaisten psyykkeestä on löydettävissä samanlaisia alajärjestelmiä kuin ihmismielestä, niin clarkilaisesta psykofunktionalismista ei seuraa kaikkien mielentilojen kieltäminen hypoteettisilta marsilaisilta tai muilta eksoottisilta olioilta. Entä jos marsilaisilta ei löydy esimerkiksi janoa vastaavaa alasysteemiä? Tämä tarkoittaisi, että joko marsilaisten kokonaispsykologia on meihin verrattuna niin erilainen, ettei mitään yhteisiä alasysteemejä ylipäätään ole löydettävissä, tai sitten noilla olioilla veden puute ei aiheuta muun muassa homeostaattista epätasapainoa, veden etsimiseen tähtääviä suunnitelmia tai kohonnutta todennäköisyyttä juoda tilaisuuden sattuessa. Näissä tapauksissa clarkilaisesta analyysistä seuraa, että marsilaiset eivät voi olla janoisia, mutta teoria silti välttänee lajisovinismin, koska ei ole mitenkään selvää, että janoisuuden käsitettä tällöin voisi heihin soveltaa. (*ibid.*, s.551.)

Clarkin analyysi lienee funktionalismin hengen vaan ei täysin pykälän mukainen. Syy, miksi funktionaalinen analyysi alun pitäen ei lähde tällaisesta alasysteemianalyysistä on, että mielentilojen määritelmissä joudutaan viittaamaan muihin mielentiloihin. Esimerkiksi janon analyysi edellyttänee, että janoisilla olioilla on ainakin haluja, mutta mahdollisesti myös vettä koskevia uskomuksia, sen hankkimiseen tähtääviä suunnitelmia ja niin edelleen. Funktionalististen määritelmien holistisuus johtuu siitä, että viittaamalla koko systeemiin kerrallaan, mentalistiset termit pystytään eliminoimaan määritelmistä – luonnollisesti määriteltävä termi pois lukien. Homunkulaarisen funktionalismin yhteydessä todettiin, että mentaalisiin tiloihin viittaaminen voi toisaalta olla myös funktionalismissa perusteltua eikä välttämättä johda ongelmiin. Clarkilainen analyysi kuitenkin herättää kysymyksiä, voiko pelkästään tietyn alasysteemin löytämistä pitää riittävänä kriteerinä, että kokonaisuudella on sitä vastaava mielentila. Jos alasysteemin määritelmissä viitataan esimerkiksi uskomus-, halu-, ynnä muihin järjestelmiin, edellyttäisi alasysteemiä vastaavan mielentilan esiintyminen myös muiden alajärjestelmien olemassaoloa. Tällöin alasysteemin täydellisen määritelmän tulisi kattaa myös noiden muiden systeemien funk-

tionaaliset määritelmät, jotka mahdollisesti edellyttävät muiden systeemien määritelmiä ja tie holismiin on taas avattu. Ongelma palautuu funktionalismin juurille ajatukseen, etteivät mielentilat yksinkertaisesti toteuta mitään psykologista funktiota muuten, kuin osana kognitiivista systeemiä. Toisekseen Clarkin idea kuulostaa kovin järkevältä, koska esimerkkinä hän käyttää yksittäistä mielentilaa, mutta on epäselvää kykeneekö tällainen analyysi kattamaan kokonaisia mielentilojen luokkia, kuten uskomusjärjestelmää ja kaikenlaisia haluja, joiden joukkoon janokin sisältynee. Clarkilainen analyysi on ylipäätään funktionalismia siis vain tietyin varauksin. Tarkat koulukuntakysymykset eivät ole kovin mielenkiintoisia, mutta mielenkiintoista on, kantaako tällainen alasysteemianalyysi kuitenkaan kovin pitkälle ilman holismia, josta se pyrkii eroon. Tarkastellaan seuraavaksi millä tavalla hierarkkinen homunkulaarinen analyysi saattaisi ratkaista esiin nousseet ongelmat tarjoamalla sarjan analyysejä, jotka ehkä hieman paradoksaalisesti ovat holistisia kuitenkaan viittaamatta organismin kokonaispsykologiaan.

Palautetaan aluksi mieliin Marrin kolmetasoinen analyysi informaationkäsittelysysteemeistä. Analyysin kaksi ensimmäistä tasoa ovat hieman hämäävästi nimetty komputationaalinen teoria, joka oleellisesti on kuvaus siitä, mitä systeemi tekee ja miksi, ja algoritminen teoria, joka kertoo mitä representaatioita systeemi käyttää ja miten. Algoritminen teoria on siis oleellisesti systeemin virtuaalikonekuvaus. Kolmas taso on implementaatioteoria, joka kertoo systeemin fysikaalisen toteutuksen. Usein systeemin ymmärtämisellä tarkoitetaan komputationaalisen teorian tuntemista, eli sen tietämistä, minkä tyyppinen systeemi on kyseessä, mitä se tekee ja minkä takia. Kuitenkin lähemmässä tarkastelussa saattaa herätä kysymyksiä, jotka vaativat hieman tarkempaa kuvausta järjestelmän toiminnasta ja toteutuksesta. Ilmeinen tällainen tilanne tulee eteen, kun systeemi toimii virheellisesti. Luonnollisesti virheellistä toimintaa ei voi selittää komputationaalisen teorian tasolla, koska tällöin systeemi tekee jotain mitä sen ei pitäisi, ja komputationaalinen teoria taas kuvaa mitä sen pitäisi tehdä. Virheellisen toiminnan ymmärtäminen edellyttää siis joko algoritmi- tai implementaatiotason selitystä. Syy virheeseen voi olla esimerkiksi ohjelmointivirheessä tai rikkoutuneessa komponentissa. Tietysti systeemi voi olla suunniteltu toimimaan virheellisesti esimerkiksi jotain koulutustarkoitusta varten. Tällöin sanoisimme, että koulutettavalla on järjestelmästä erilainen komputationaalinen teoria kuin sen laatijalla. Komputationaalinen teoria ei siis ole systeemin täydellinen eikä yksiselitteinen kuvaus sikäli, että se ei välttämättä täsmällisesti kerro mitä funktiota systeemi laskee, paitsi ehkä jossain yksikertaisessa ideaalitapauksessa. Komputationaalinen teoria ei siis ole kausaalinen eikä edes matemaattisessa mielessä intensionaalinen, vaan oikeastaan intentionaalinen kuvaus tietojenkäsittelyjärjestelmästä.

Tasojen välinen erottelu on hyvä tehdä silloinkin, kun tarkastellaan normaalisti toimivaa järjestelmää. Ajatellaanpa, että kaksi fyysisesti identtistä konetta laskee samaa funktiota. Nyt jos toinen koneista laskee toista nopeammin, selitys mitä ilmeisimmin on, että koneiden laskennat on toteutettu algoritmisesti eri tavoilla. Vastaavasti jos kaksi konetta suorittaa täsmälleen samaa algoritmia eri tahtiin, syyn täytyy olla fysikaalisen implementaation eroissa. Huomattakoon lisäksi, että tasojen välinen erottelu ei pelkästään auta ymmärtämään jotain tiettyä laitetta, vaan informaatiota käsittelevien systeemien luonnetta ylipäätään. Tästä esimerkkinä mainittakoon melko yleinen käsitys, että tietokoneet eivät tee virheitä vaan yksinkertaisesti sitä, mitä ne on ohjelmoitu tekemään. Tätä huo-

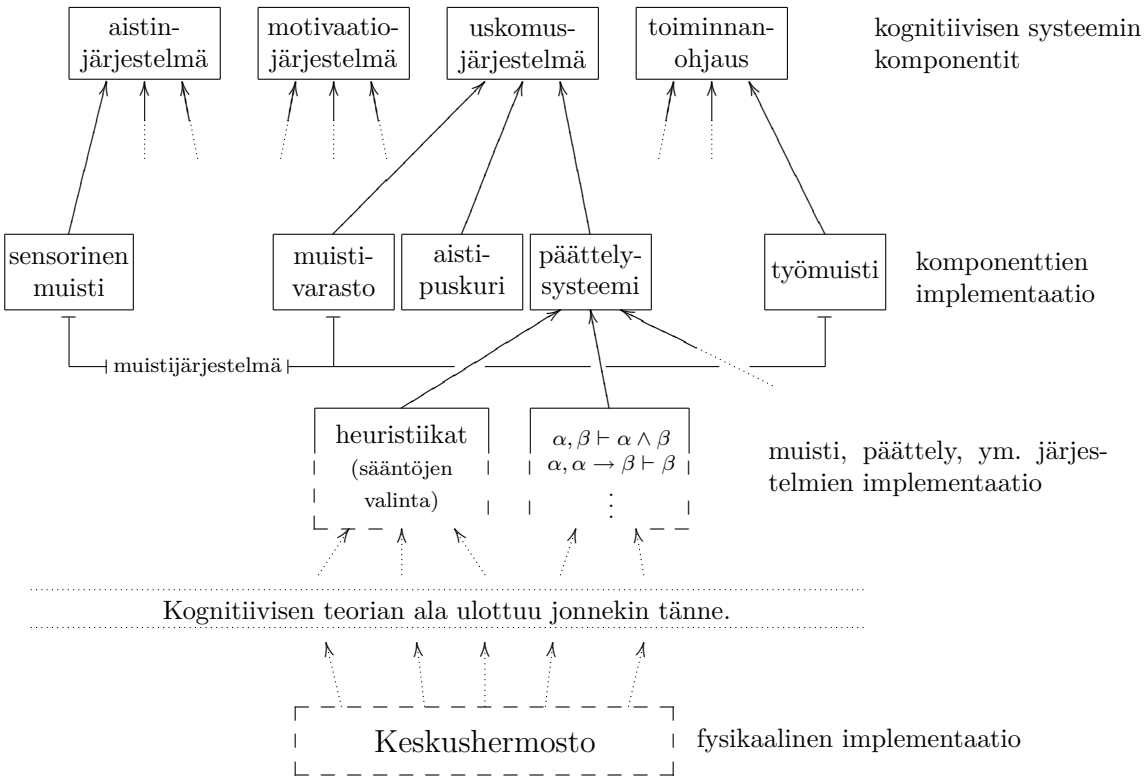
miota käytetään ajoittain tekoälyä vastaan: koska ihmiset tekevät virheitä ja koneet eivät, ihmiset eivät ole koneita. Tätä vasta-argumenttia käsitteli jo Turing aikoinaan (Turing, 1950, s.448–449). Kyseessä on filosofisesti mielenkiintoinen väärinkäsitys, joka seuraa siitä, ettei komputationaalista teoriaa huomata erottaa algoritmisesta. Algoritmitaso on abstrakti kuvaus koneen tosiasiallisista tekemisistä askel askeleelta, ja on totta, että tällä tasolla tarkasteltuna koneet eivät tee virheitä, siis mikäli algoritmit suoritetaan oikein. Tämä kuitenkin johtuu siitä, että virheen käsite ei sovellu toiminnan kausaaliseen, vaan intentionaaliseen kuvaamiseen. Ei esimerkiksi ole mielekästä sanoa, että ”pelaaja virheellisesti potkaisi pallon ohi maalin”, ellei oleteta, että pelaajan olisi pitänyt potkaista pallo maaliin. Tapahtumien puhtaasti kausaalisissa kuvauksissa, irrotettuna tulkinnasta, mitä olisi pitänyt tapahtua, ei virheen käsitteellä ole mitään merkitystä. Komputationaalisen teorian tasolla tarkasteltuna koneet kyllä tekevät virheitä, jos komputationaalinen teoria ajatellaan intentionaaliseksi kuvaukseksi. Koneet kaatuilevat, ohjelmat toimivat miten sattuu, tekevät tyhmiä siirtoja shakissa ja niin edelleen. Sivumennen sanoen samankaltainen erottelu pätee ihmisten toimintaan tarkasteltuna fysikaalisesti ja intentionaalisesti: aivot eivät tee virheitä, ihmiset tekevät.

Tarkastelun yhteydet homunkulaariseen analyysiin lienevät jokseenkin ilmeiset. Yllä olevan selostuksen tarkoitus on teroittaa, että jos jonkin representaatioita käsittelevän systeemin toimintaa ollaan selittämässä, tai sen komponentteja määrittelemässä, on oleellista huomioida mitä analyysin tasoa oikeastaan tarkastellaan. Huomioitakoon, että algoritminen teoria ei tarkoita samaa kuin funktionaalinen analyysi. Algoritminen teoria kuvaa kyllä ohjelman suorituksen kausaalisen rakenteen, mutta se on vuokaaviotyyppistä kuvausta yksityiskohtaisempi pitäen sisällään kuvauksen systeemin käyttämistä representaatioista ja prosesseista. Toisaalta funktionaalinen analyysi ei tuota systeemistä intentionaalista vaan kausaalisen kuvauksen, joten se ei myöskään ole komputationaalinen teoria Marrin tarkoittamassa mielessä. Ohjelman funktionaalinen analyysi osuu johonkin tähän väliin. Mielen algoritminen teoria tarkoittaisi täysveristä komputationaalis–funktionaalista teoriaa, joka sisältää kuvauksen sekä mentaalisista representaatioista että niitä manipuloivista prosesseista. Voidaan sanoa, että funktionaalinen analyysi on algoritmisen teorian abstraktio, jossa ei huomioida täsmällisiä komputationaalisia prosesseja. Tästä syystä sama funktionaalinen rakenne voidaan toteuttaa monilla eri virtuaalikoneilla ja formalismeilla.

Pohjustetaanpa varsinaista asiaa vielä tietokoneanalogialla. Tarkastellaan tilannetta, jossa koneella M suoritetaan tulkin T avulla jotain korkean tason ohjelmointikielellä L muodostettua ohjelmaa O. Esimerkin yksinkertaistamiseksi oletetaan, että konetta käytetään nimenomaan kyseisen ohjelman suorittamiseen, joten systeemin komputaationaalinen teoria palautuu kysymykseksi, mitä ohjelma O tekee. Tässä tapauksessa algoritminen teoria taas ei ole täysin yksiselitteinen. Ohjelman O L-kielinen koodi on täydellinen algoritminen kuvaus O:n toiminnasta, mutta se ei kata koko systeemin algoritmista kuvausta. Ohjelman O suorittaminen ei ole implementoitu suoraan fyysisesti vaan välillisesti tulkin T avulla, joka puolestaan on M:n konekielinen ohjelma. Nyt tulkki voidaan ottaa analyysin kohteeksi, ja sille voidaan muodostaa oma komputationaalinen teoriansa, joka on yksinkertaisesti, että se kääntää L-kielisen ohjelmakoodin koneen M konekielelle käyttöjärjestelmän suoritettavaksi. Tulkin algoritminen teoria taas on sen konekielinen ohjelmakoodi. Edelleen T:n suoritus ei ole implementoitu suoraan fysikaalisesti, vaan suorittamisesta

huolehtii käyttöjärjestelmä, joka on M:n konekielellä toteutettu virtuaalikone, jolle puolestaan on oma komputationaalinen ja algoritminen teoriansa, ja joka lopulta suoritetaan koneen M fyysisesti implementoidulla ensimmäisen kertaluvun virtuaalikoneella.

Edellä oleellista on tasojen M, T ja O muodostama implementaatioiden ketju. Mikäli haluamme tietää, miten systeemi toimii, kysymme luultavasti ohjelman O rakennetta. Se, miten tulkki suorittaa ohjelmaa, ei kuulu asiaan, ellei kysymys sitten nimenomaan koske juuri tätä. Selonteko tulkin ja käyttöjärjestelmän toiminnasta on ohjelman kannalta implementaatioteoriaa, eikä niillä välttämättä ole merkitystä ohjelman toiminnan selittämisen kannalta. Ohjelma O on siis algoritmisesti monitoteutuva, kuten ohjelmat yleensä. Jos taas kysytään nimenomaan tulkin tai käyttöjärjestelmän rakennetta, koneen fyysinen ilmplementaatio on vastaavasti epäoleellista. Lisäksi tulkin ja käyttöjärjestelmän tiloilla ei ole mitään suoranaista tekemistä niiden suorittamien ohjelmien sisäisten tilojen kanssa. Luonnollisesti tulkin tilat ovat vastuussa sen suorittaman ohjelman tilojen implementaatiosta, mutta ydinasia on, että T:n tilat ovat eri asia kuin O:n tilat, jo siitäkin syystä, että T voi implementoida äärettömän monenlaisten ohjelmien ajamisen. Näin ollen O:n täydellinen kuvaus ei sisällä viittauksia tulkin T sekä käyttöjärjestelmän muodostaman virtuaalikoneen tiloihin, ja M, T sekä O ovat pitkälti autonomisia systeemeitä. Koska virtuaalikoneita voidaan periaatteessa pinota mielin määrin, tarinan opetus ei päde vain kolmeen analyysin tasoon, vaan kokonaiseen virtuaalikoneiden hierarkiaan. Edellisessä alaluvussa käsitelty mielenmalli on alla purettu tällaiseksi implementaatioiden ketjuksi.



Sitten takaisin marsilaisiin. Jos marsilaisten psykologia on karkeasti ottaen identtinen meidän psykologiamme kanssa, voisi olettaa, että heidän kognitiivinen koneistonsa olisi purettavissa suurin piirtein samanlaisiksi *korkean tason* systeemeiksi kuin meidän, mistä haluja, pelkoja, uskomuksia ja muita arkipsykologisia tiloja vastaavat komponentit oletettavasti on löydettävissä. Blockin ja erityisesti Shoemakerin argumentti nojaa ajatukseen, että marsilaisten ja meidän väliset erot ovat hienovaraisia ”syvyyspsykologisia” eroja (Shoemaker, 1981, s.280), jotka piilevät jossain pinnan alla, kenties muistijärjestelmän toiminnan yksityiskohdissa. Homunkulaarinen funktionalisti voi kuitenkin todeta, että muistijärjestelmiä koskevat kysymykset eivät ole samalla analyysin tasolla, kuin uskomuksia, haluja ja muita arkipsykologisia tiloja ja prosesseja koskevat. Funktionaalisten määritelmien tulee olla holistisia aiemmin mainituista syistä, mutta tämä holistisuuden vaatimus koskee sitä analyysin tasoa, jonka ilmiöitä ja tiloja määritellään. Ei ole mitään käsitteellistä syytä, miksi määritelmien pitäisi leikata koko systeemin läpi myös ikään kuin vertikaalisessa suunnassa. Toisin sanoen, kun kognitiivista systeemiä tarkastellaan uskomus-, halu-, ynnä muiden järjestelmien kokoelmana, analyysissä ei tarvitse viitata muistijärjestelmiin tai muihin alemman tason systeemeihin. Itse asiassa analyysin ei edes tule viitata tällaisten systeemien yksityiskohtaiseen rakenteeseen, koska ne kuuluvat korkean tason järjestelmän implementaatioteorian alaan. Kun esimerkiksi uskomusjärjestelmästä puhutaan osana arkipsykologisen tason selitystä, se on periaatteessa primitiivinen funktionaalinen komponentti, jonka funktiota voivat toteutta sisäisesti hyvin erilaiset järjestelmät eri systeemeissä.

Jos taas halutaan tietään, miten uskomusjärjestelmä toimii, on aivan mielekästä todeta tämän riippuvan kenen uskomusjärjestelmästä puhutaan. Tällöin kyseistä komponenttia siirrytään tarkastelemaan kokonaissysteeminä, ja kysymykseen vastaaminen edellyttää funktionaalisen analyysin suorittamista tälle järjestelmälle. Vastaavasti, kuin käyttöjärjestelmän tiloilla ei ole mitään suoranaista tekemistä sen suorittamien ohjelmien tilojen kanssa, ei ole syytä olettaa, että uskomus-, ja muiden alasysteemien sisäisillä tiloilla olisi sen enempää tekemistä kognitiivisen koneiston korkean tason tilojen kanssa, ainakaan siinä mielessä, että jälkimmäiset palautuisivat edellisiin. Nyt jos marsilaisten alasysteemit toimivat eri tavalla kuin meidän, on tietenkin totta, että olemme kokonaisuutena katsottuna jossain määrin psykologisesti erilaisia. Tästä ei kuitenkaan seuraa, että olisimme analyysin kaikilla tasoilla erilaisia. Se, mitkä ilmiöt kuuluvat millekin, ja erityisesti samoille, tasoille on kysymys, johon minulla ei ole vastausta, ja johon homunkulaarisen psykofunktionalistin ei lähtökohtaisesti tarvitse vastata. Kyseessä on enemmänkin mielen rakennetta koskeva empiirinen kuin käsitteellinen kysymys.

Huomautettakoon, että eri kysymyksiin vastaaminen voi edellyttää myös systeemin pilkkomista funktionaalisesti eri tavoilla. Mikäli ihmettelemme, millainen on ihmisen muistin rakenne, niin oppikirjavastauksen mukaan se karkeasti ottaen koostuu sensorisesta muistista, pitkäkestoisesta muistista (yllä olevassa kaaviossa ”muistivarasto”) ja lyhytkestoisesta työmuistista. Edellä esitetyssä kaaviossa nämä muistijärjestelmät kuuluvat eri kognitiivisiin komponentteihin, mutta tietyssä mielessä ne muodostavat yhtenäisen muistijärjestelmäksi kutsutun kokonaisuuden. Tässä mielessä funktionaalinen rakenne ei ole välttämättä edes riippumaton siitä, mitä tarkalleen ottaen kysytään.

Sivumennen sanoen tämänlaista argumenttia voidaan käyttää myös vastaamaan tiettyihin eliminativistisiin argumentteihin. Ensimmäisessä luvussa mainittiin Paul Churchlandin kritisoineen arkipsykologiaa siitä, ettei se kykene selventämään tiettyjä psykologisia ilmiöitä, kuten poikkeuksellista luovuutta, tiettyjä mielenhäiriöitä ja yksilöiden välisiä eroja älykkyydessä (s.7). Tästä Churchland päätteli, ettei arkipsykologia voi tarjota kaikenkattavaa psykologista teoriaa, joten se tulisi korvata tarkemmalla ja kattavammalla tieteellisellä teorialla, joka ei perustu uskomus–halu-ontologialle. Voitaneen myöntää, että väitteen premissi pitää paikkansa, mutta edellä esitettyihin tarkasteluihin nojaten johtopäätös ei seuraa. Argumentti on mielekäs ainoastaan, jos ajatellaan, että psykologinen teoria on yksi suuri kokonaisuus, joka pitäisi muotoilla yhdellä kaiken kattavalla käsitteistöllä, ja arkipsykologisen käsitteistön olisi tarkoitus toteuttaa tämä vaatimus. Jos kuitenkin arkipsykologisten käsitteiden alan katsotaan rajoittuvan vain korkeimman tason siivuun kognitiivisessa systeemissä, on tietysti selvää, ettei niillä voida kuvata kaikkia mahdollisia psykologisia ilmiöitä. Samalla kuitenkin pitäisi olla selvää, että tämä ei ole oire arkipsykologian perikadosta, vaan siihen liittyvät käsitteet ja selitykset voivat kuitenkin vastata todellisia mentaalisia tiloja ja prosesseja tietyllä kognitiivisen systeemin analyysin tasolla. Esimerkiksi erot yksilöiden välisessä älykkyydessä saattavat selittyä jollakin päättelysysteemien rakennetta koskevilla teorioilla, joilla ei ole suoranaisesti tekemistä uskomusten ja halujen kanssa.

Painotettakoon lopuksi, että näiden tarkastelujen merkitys ei ole valmistautua muukalaisten kohtaamisen, vaan tutkia niitä teorianmuodostuksen periaatteita, joiden avulla voimme toivottavasti joku päivä ymmärtää omia mieliämme, ja hyvällä onnella mentaalisuutta luonnonilmiönä ylipäätään. Yhteenvetona nämä funktionalismiin liittyvät teorianmuodostuksen periaatteet ovat seuraavanlaiset: 1. *Behaviorismin hylkääminen.* Mentaaliset prosessit ovat organismin sisäisiä prosesseja, joiden tutkiminen on mahdollista ja psykologian kannalta välttämätöntä. Käyttäytymistaipumusten avulla ei ole mahdollista muodostaa moniakaan mielenkiintoisia psykologisia yleistyksiä. 2. *Fysikalismin hylkääminen.* Mentaaliset prosessit ovat fysikaalisesti implementoituja, mutta on epäuskottavaa, että fysiikan avulla pystyttäisiin määrittelemään, mikä mentaalisia olioita yhdistää nimen omaan mentaalisina olioina. Fysikalismin eräs ongelma on, että niin psykologian, kuin monien muidenkin erityistieteiden, yleistyksien kannalta fysiikka on liian yleinen tiede. Monet ilmiöt ovat fysiikan perspektiivistä täysin erilaisia, vaikka ne jonkin muun teorian kannalta saattavat olla täysin identtisiä. Näin ollen useimpien erityistieteiden, ja tässä yhteydessä erityisesti psykologian, lainalaisuuksien muotoileminen fysiikan teorioiden avulla on luultavasti mahdotonta. 3. *Psykoneuraalisten identiteettien hylkääminen.* Mikäli teoria neuraalisista tiloista ja prosesseista ajatellaan kuuluvan osaksi fysiikkaa, edellisestä kohdasta seuraa suoraan, että neurotieteiden käsitteistö on riittämätön määrittelemään mentaalisia tiloja. Syy identiteettiteorian hylkäämiseen ei kuitenkaan ole neurotieteiden liian laaja perspektiivi vaan, että se päinvastoin on liian suppea teoria kuvaamaan kaikkia mahdollisia mentaalisia olioita. 4. *Komputationaalisen mekanismin hylkääminen.* Saattaa kuulostaa kummalliselta, että funktionalistisessa analyysissä tulee hylätä varsinaiset komputationaaliset mekanismit, joiden ajatellaan olevan vastuussa mentaalisesta kausaatiosta. Kuitenkaan funktionalismi ja komputationalismi – vastoin melko yleistä väärinkäsitystä – eivät ole sama teoria, vaikkakin ne ovat hyvin läheisiä. Käsittääkseni konetilafunktiona-

lismia vastaan esitetyt sovinismiargumentit pätevät periaatteessa mihin tahansa formalismiin, joskin Turing-koneiden käyttäminen psykologisten teorioiden redusoimiseen pitää sisällään lisäksi omia erityisiä ongelmiaan. Kuitenkin samanlaiset syöte–tuloste-profiilit voidaan toteuttaa periaatteessa rajattoman monenlaisilla formalismeilla, eikä vaikuta kovin uskottavalta, että mielentilat vastaisivat suoraviivaisesti minkään tietyn virtuaalikoneformalismin tiloja. Se, että kognitiiviset prosessit ovat luonteeltaan komputationaalisia, on kognitivistiseen mielenteoriaan sisältyvä erillinen empiirinen hypoteesi, ei funktionalismiin sisältyvä käsitteellinen väite.

Näiden negatiivisten luonnehdintojen jälkeen lopulta käteen jää teesi, jonka mukaan mentaalisia olioita yhdistää se, että ne toteuttavat jonkin melko abstraktisti määritellyn kausaalisen kuvauksen, missä mentaaliset tilat toimivat välittäjinä kognitiivisen systeemin syötteen ja tulosteen välillä, ja että mentaalisten tilojen identiteetti määräytyy niiden suhteesta tuohon kokonaisjärjestelmään. Psykofunktionaalinen analyysi tarjoaa tältä pohjalta empiiriset kriteerit vertailla, missä mielessä kaksi systeemiä ovat psykologisesti samanlaisia tai erilaisia. On hieman epämääräistä sanoa, että mentaalisia *olioita* yhdistää tietty funktionaalinen kuvaus, ellei mentaalisuuden kriteeriksi sitten valita nimeonmaan tietynlaista funktionaalista organisaatiota, esimerkiksi samanlaista kuin meillä. Tämä olisi kuitenkin hieman omituista, koska tällöin poikkeava mentaalisuus olisi käsitteellisesti mahdotonta. Näin ollen on ehkä parempi pitää psykofunktionaalista samankaltaisuutta mentaalisuuden riittävänä mutta ei välttämättömänä ehtona. Tämä tarkoittaa, että funktionaaliset määritelmät poimivat tiettyjen mentaalisten tilojen luonnolliset luokat, mutta eivät kaikkien mahdollisten mentaalisten olioiden luonnollista luokkaa. Mikäli kuitenkin tämä katsotaan funktionalismin tappioksi, voisimme katsoa mentaalisten olioiden muodostuvan luokasta olioita, joiden funktionaalinen rakenne mahdollistaa sen, että niiden voidaan sanoa muun muassa uskovan ja toivovan. Tämä saattaa olla turhan kapea näkökulma mentaalisuuteen, mutta joka tapauksessa homunkulaarinen analyysi antaa aika paljon liikkumatilaa sille, minkälaisia olioita tähän luokkaan voidaan katsoa kuuluvaksi.

Toisaalta funktionalismin tarjoama opetus on ehkä jotakuinkin sellainen, ettei mentaalisten olioiden luokka itse asiassa ole hyvin määritelty, eikä ole mitään kovin selviä metafyysisiä tai tieteellisiä kriteereitä sille, mitkä oliot ovat mentaalisia ja mitkä eivät. Uskoisin, että kyseinen luokka on rajoiltaan epämääräinen ja lähinnä perheyhtäläinen kokoelma olioita, joita yhdistää löyhät samankaltaisuudet funktionaalisen rakenteen suhteen. Homunkulaarinen psykofunktionalismi soveltuu varsin hyvin tällaisten olioiden luokitteluun juurikin siitä syystä, ettei analyysi anna kovinkaan luontevaa tapaa rajata olioita selkeästi mentaalisiin ja ei-mentaalisiin, mutta silti se tarjoaa käsitteellisiä välineitä vertailla, missä mielessä oliot ovat samanlaisia sisäisen tietojenkäsittelyn ja käyttäytymisen etiologian suhteen.

4 Kognitivistisen teorian olemus ja ongelmat

Kaksi edellistä lukua käsittelivät kognitivistisen mielenteorian pääkomponenttien, komputationalismin ja funktionalismin, kehitystä ja keskeisimpiä ideoita. Näistä ensiksi mainittu on empiirinen hypoteesi, jonka mukaan kognitiivinen systeemi voidaan implementoida jonkinlaisen komputationaalisen systeemin avulla. Teoria ei sinänsä väitä, että mentaalisuus edellyttäisi lainomaisesti tai käsitteellisesti välttämättä jonkinlaista laskentaa. Funktionalismi puolestaan on mielentilojen luonnetta koskeva käsitteellinen tai metafyysinen teoria, jonka mukaan mieli on muun muassa havainnon ja toiminnan suhdetta välittävä kausaalinen järjestelmä, ja mielentilojen identiteetit määräytyvän sen perusteella, mitkä niiden kausaaliset ominaisuudet ovat tuon systeemin toiminnassa. Sinänsä funktionalismi ei ota kantaa kyseisen kausaalirakenteen implementaatioon, joten periaatteessa teoria on komputationalismista riippumaton. Kuten viime luvussa nähtiin, näiden teorioiden sekoittaminen siten, että funktionaaliset määritelmät samaistetaan komputationaalisen mekanismin tiloihin, johtaa helposti ongelmalliseen mielenteoriaan. Kun funktionalismi ja komputationalismi pidetään sopivasti erillään, kokoelma niiden muodostamaan mielenteoriaan liittyviä ongelmia saa melko luonnollisen ratkaisun. Homunkulaarinen funktionalismi poistaa tarpeen muotoilla propositionaaliset asenteet suoraan komputationalistisin termein, ja oikeuttaa mielen reduktiivisessa selittämisessä viittaamisen intentionaalisiin järjestelmiin, jotka ovat selitettävää kokonaissysteemiä yksinkertaisempia. Kun intentionaalisia alasysteemejä puretaan tarpeeksi perusteellisesti yhä yksinkertaisemmiksi palasiksi, lopulta käteen pitäisi jäädä verrattain yksinkertaisia mekaanisia komponentteja, joista mieli lopulta rakentuu. Komputationaalinen mielenteoria täydentää tämän selonteon selittämällä mekaanisten komponenttien olevan luonteeltaan syntaktisilla symbolirakenteilla operoivia prosesseja.

Kun symbolirakenteet yhdistetään ajatusten ja niiden manipulaatiot ajattelun kanssa, tarjoaa komputationaalinen teoria eräänlaisen ratkaisun mieli–ruumis-ongelmaan. Mutta tarkalleen ottaen, miksi olettaa mielen toiminnan olevan nimenomaan eräänlaista laskentaa tai tietojenkäsittelyä? Se, että tietokoneet tarjoavat mielenkiintoisen analogian mielelle, ei tarkoita, että komputationalismi olisi oikea selitys eikä, että analogia tulisi ottaa kirjaimellisesti. Edellisessä luvussa päädyttiin näkemykseen, jonka mukaan abstraktit, jonkinlaisia kausaalisuhteita kuvaavat vuokaaviomallit tarjoavat perustellun tavan kuvata mielen rakennetta ja toimintaa, mutta kuten todettu, mikään tässä metodologiassa ei sinänsä edellytä komputationalismia. Nojatuolista tarkasteltuna komputationalistinen teoria vaikuttaa hyvältä idealta, mutta koska psykologia voi olla kognitiivista olematta komputationaalista, niin millä perusteella kyseessä on oikea idea? Tarkastellaan aluksi hieman täsmällisempiä komputationaalisen teorian muotoiluja, eli sitä, minkälaisen muodon toisen luvun lopussa esitetyt ideat laskennan ja ajattelun samaistamisesta ovat 1900-luvun jälkipuoliskolla oikeastaan saaneet. Tämän jälkeen tarkastellaan kognitivismiin liittyviä ongelmia, ja lopuksi vedetään yhteen, mistä nämä ongelmat nähtävästi kumpuavat, miten ne ehkä olisivat ratkaistavissa ja miltä kognitivismin tila ja tulevaisuus näiden tarkastelujen valossa oikein näyttävät.

4.1 Komputationalismi ja kognitiotiede

Johdannossa (s.10) esiteltiin syitä olettaa, että tarkoituksenmukaisen toiminnan kuvaamiseksi on välttämätöntä käyttää mentalistisia predikaatteja, tarkemmin sanoen propositionaalisia asenteita. Vaikka emme ymmärtäisi käyttäytymistä muuten kuin tällaisten mentalististen käsitteiden avulla, ei tästä tietenkään vielä sinänsä seuraa, että niitä vastaavia prosesseja tai tiloja varsinaisesti löytyisi ihmisen kognitiivisesta systeemistä. Kuitenkin jos arkipsykologia sisältää ontologisesti kelvollisen kuvauksen mielestä, uskomukset ja halut todella majailevat tai tapahtuvat jossain korvien välissä. On syytä kiinnittää huomiota nimenomaisesti mentaalisten representaatioiden ja tilojen väliseen erotteluun sekä representaatioiden ja propositioiden hyvin läheiseen yhteyteen. Yleisesti ottaen proposition olemuksellisena piirteenä pidetään sen totuus- tai toteutumisehtoja, jotka määrittävät sen merkityksen. Lisäksi propositioita voidaan yhdistellä monimutkaisemmiksi kokonaisuuksiksi, esimerkiksi konjunktio- tai ehtolauseiksi. Niiden välillä vallitsee niiden merkityksille perustuvia loogisia ja käsitteellisiä suhteita, jotka muodostavat abstraktin rakenteen, joka puolestaan analyyttisen filosofian perinteisen viisauden mukaan voidaan esittää jonkinlaisena loogisena systeeminä. Jos organismilla on erityisiä mentaalisia suhteita propositioihin ja oletamme, että nuo propositiot kirjaimellisesti ovat jollain tavalla mentaalisten olioiden mielen sisällä, niin ilmeinen – ellei jopa välttämätön – hypoteesi on olettaa organismin ruumillistavan jonkinlaisen rakenteeltaan propositioiden systeemiä vastaavan loogisen systeemin.[47] Materialistisen taustaoletuksen valossa propositioiden järjestelmällä tulee siis lopulta olla organismissa jonkinlainen fysikaalinen, loogisen systeemin kanssa rakenneyhtäläinen implementaatio.

Jos kognitiivinen systeemi on äärellinen mekanismi, jonka toiminta ohjautuu systemaattisesti mentaalisten representaatioiden syntaktisten – eli implementaatiotasolla siis lopulta fysikaalisten – ominaisuuksien perusteella, niin kysymys, miksi pitää mieltä komputationaalisena systeeminä, saa ilmeisen vastauksensa. Mikäli edellä on maalattu oikea kuva mielestä, niin ei ole mitään erillistä empiiristä kysymystä siitä, onko mieli komputationaalinen systeemi, koska tätä laskennalla ainakin näissä yhteyksissä tarkoitetaan. Yhtä mielekästä olisi kysyä, suorittaako taskulaskin laskentoja oikeasti vai jotenkin näennäisesti. Jos taas tämä ei ole oikea kuva mielestä, niin mentaalisuus on täysi mysteeri. Tämä on kuuluisa komputationalismin *the only game in town* -argumentti, jonka mukaan mitään uskottavaa vaihtoehtoa komputationalistiselle teorialle ei ole.[48] On sitten asia erikseen, halutaanko propositioiden sijaitsemiseen mielessä sitoutua näin vahvasti.

Niin sanottu *klassinen komputationaalinen teoria* ottaa edellisen argumentin vakavasti. Lisäksi teoria on sitoutunut kahteen seuraavaan teesiin mentaalisten representaatioiden ja prosessien luonteesta (Fodor & Pylyshyn, 1988, s.12–13). 1. *Mentaalisten representaatioiden kompositionaalinen syntaksi ja semantiikka*: Mentaaliset representaatiot ovat

[47] Ks. (Field, 1978, s.88,114). Johtopäätöksen väistämättömyys riippuu hieman siitä, miten kirjaimellisesti propositioiden, tai oikeastaan niiden representaatioiden, ajatellaan oleilevan organismin sisällä (Dennett, 1982, s.123–126).

[48] Muun muassa Fodor on vedonnut tähän 70-luvulta (Fodor, 1975, s.27) aivan viime aikoihin asti (Fodor, 2008, s.58–61,112–113), joskin argumentti on menettänyt merkittävästi vetovoimaansa sitten 80-luvun, jolloin logiikkapohjaiselle mallintamiselle kehitettiin uskottavia vaihtoehtoisia komputationaalisia malleja, joista lyhyesti lisää myöhemmin.

niin sanotun ajattelun kielen lauseita. Kyseessä on symbolijärjestelmä, jolle ominaista ovat seuraavat piirteet: a) Erottelu atomisten ja molekulaaristen, eli yksinkertaisien ja kompleksisten, representaatioiden välillä, b) molekulaariset ilmaukset voidaan koostaa joidenkin kompositioperiaatteiden mukaan toisista ilmauksista ja c) jokaisen ilmauksen, eli ajattelun kielen lauseen, sisältö määräytyy yksiselitteisesti sen kompositionaalisen rakenteen ja sen sisältämien atomisten rakenneosien merkitysten perusteella. 2. *Prosessien rakenneherkkyys*: Mentaaliset prosessit manipuloivat ajatuksen kielen lauseita, ja nämä prosessit ovat määritelty representaatioiden muodon, eli syntaksin, perusteella. Prosessit eivät pelkästään muuntele symboleita toisiksi, vaan ne manipuoivat myös molekulaaristen representaatioiden rakenteita. Nämä kaksi ominaisuutta ovat pääpiirteissään loogisten kielten ja deduktiosysteemien oleelliset ominaispiirteet mainitussa järjestyksessä.

Klassisen komputationaalisen teorian on tarkoitus selittää kaksi mielen toiminnan ominaislaatuista piirrettä, nimittäin ajattelun systemaattisuuden ja produktiivisuuden.[49] Ajattelun *systemaattisuus* itse asiassa viittaa ajattelun kahteen hieman erilaiseen mutta toisiinsa kytkeytyvään ominaisuuteen. Uskomukset näyttäisivät olevan jossain määrin loogisesti systemaattisia siinä mielessä, että esimerkiksi jos x uskoo, että Jaska jahtaa kissaa, niin x mitä suurimmalla todennäköisyydellä uskoo myös, että Jaksa jahtaa jotakin, eräs kissa on jonkun jahtaama ja ylipäätään joku, tai jokin, jahtaa jotakin. Vastaavasti jos x uskoo väitteet "A" ja "B", niin oletettavasti hän uskoo myös väitteen "A ja B" ja päinvastoin. Tätä voitaisiin kutsua *deduktiiviseksi systemaattisuudeksi*. Tämä huomio on osittain käsitteellinen. Mikäli olion ajattelu ei ole tällä tavalla systemaattista, on hieman epäselvää, onko hänellä esimerkiksi konnektiivin "ja" käsitettä ollenkaan, ja mikäli kissan jahtaamisen tapauksessa kuvattu systemaattisuus pettää, on hyvin vaikea sanoa mitä x oikeastaan uskoo, tai onko hänellä varsinaista käsitteellistä ajattelua itse asiassa alkuunkaan. Mikäli ei, niin on epäselvää voiko uskomisen käsitettä tällöin edes soveltaa. Toista systemaattisuuden muotoa voitaisiin kutsua *kompositionaaliseksi systemaattisuudeksi*. Kyseessä on kyky ymmärtää rakenteellisesti samanlaisia, mutta merkitykseltään erilaisia propositioita. Esimerkiksi jokainen, joka kykenee ajattelemaan, että Jaska jahtaa kissaa, kykenee myös ajattelemaan, että kissa jahtaa Jaksaa. Edelleen, jos kykenee ajattelemaan, että talo on keltainen ja auto punainen, kykenee melko varmasti ajattelemaan myös, että talo on punainen ja auto keltainen.[50] Komputationalismi on hyvin yksinkertainen selitys molemmille systemaattisuuden muodoille.

Produktiivisuus puolestaan tarkoittaa kykyä ajatella periaatteessa ääretöntä määrä erilaisia ajatuksia ja käsitteitä. Komputationaalinen teoria selittää tämän siten, että atomisista representaatioista voidaan rakentaa periaatteessa rajatta uusia ja merkitykseltään erilaisia molekulaarisia representaatioita. Tällaisten kompleksisten representaatioiden merkitys

[49]Ks. (Fodor & Pylyshyn, 1988, s.33–50), (Davies, 1991, 239–250) ja (Marcus, 1998, s.275–277), joista ensiksi mainittu lienee tunnetuimpia asiaa käsitteleviä artikkeleja. Kaikki mainitut lähteet tarkastelevat systemaattisuutta ja produktiivisuutta suhteessa klassiselle teorialle vaihtoehtoisiin komputationaalisiin malleihin.

[50]Miten pitkälle tällainen systemaattisuus oikeastaan ulottuu? Esimerkiksi "kissa nukkuu talossa" on täysin ymmärrettävä, mutta entä "talo nukkuu kissassa"? Jälkimmäinen väite on semanttisesti luonnoton, mutta tavallaan tulkittavissa oleva. Kuka tahansa, jolla on normaali talon, kissan ja nukkumisen käsite, kykenee ymmärtämään, että talot eivät ole asioita, jotka nukkuvat tai mahtuisivat kissojen sisälle. Toisaalta jossain *Liisa ihmemassa* -tyyppisessä tarinassa tällä lauseella voisi olla aivan mielekäs ja käyttökelpoinen sisältö.

palautuu niiden rakenteeseen sekä niiden sisältämien atomisten representaatioiden merkityksiin. Ajatus tässä on oleellisesti sama, kuin kompositionaalisten malliteoreettisten kielten tapauksessa, joissa jokaiselle oikein muodostetulle lauseelle on yksiselitteinen tulkinta, mikäli sen rakenneosien merkitykset on kiinnitetty. Oikeastaan jos mieli todella on eräänlainen deduktiosysteemi, produktiivisuus ja kompositionaalinen systemaattisuus ovat hyvin pitkälti sama asia. Joka tapauksessa nämä ajattelun piirteet selittyvät hyvin elegantisti – ellei jopa ainoastaan – olettamalla kognitiivisen koneiston perustuvan rekursiiviseen, rakenneherkkään symbolirakenteiden käsittelyyn.

Ajattelun produktiivinen ja systemaattinen luonne voi vaikuttaa itsestään selvältä, mutta tarvittaessa perusteluja näille näkemyksille saadaan kielitieteiden ja kielellisen käyttäytymisen puolelta. Tunnetusti jo Descartes katsoi kielen erottavan ihmiset eläimistä ja mekaanisista automaateista (Descartes, 2001, s.154–156). Hänen mukaansa nimittäin kielen käyttö edellyttää järkeä, joka puolestaan joustavana ja universaalisena instrumenttina ei ole mekanisoitavissa. Hieman modernimpi näkökulma kielellisen kyvyn ja ajattelun yhteyteen polveutuu paljolti chomskylaisesta kielitieteestä. Toisessa luvussa (s.24) sivuttiin Chomskyn 50-luvun lopun töitä, jotka viittasivat kielellisen käyttäytymisen – siis niin kielen tuottamisen kuin ymmärtämisen – edellyttävän sisäistä representaatiojärjestelmää sekä sillä operoivia syntaktisia mekanismeja. Tarkalleen ottaen Chomsky ei ollut ensisijaisesti kiinnostunut kielellisen käyttäytymisen taustalla olevista kognitiivisista mekanismeista eikä myöskään ajattelusta sikäli, ettei hänen työnsä varsinaisesti käsitellyt esimerkiksi kielellisten ilmausten merkitystä. Hän tutki kieltä abstraktina lauseiden järjestelmänä ja oli kiinnostunut kielioppisäännöistä, jotka tuottavat kielen sanastosta äärettömän määrän erialisia kieliopillisia ilmauksia sisältävän systeemin. Chomsky suhtautui tähän systeemiin eräänlaisena teoriana, joka määrittää ne kielelliset ilmaukset, jotka kielen käyttäjä periaatteessa tunnistaa kieliopillisiksi. Kyseessä on siis idealisoitu kuvaus kielen käyttäjän kielellisestä kompetenssista (Chomsky 1957, s.49–52; 1965, s.3–9). Toisaalta, jos kielen käyttäjä ymmärtää käyttämänsä sanaston, systeemi samalla kuvaa, mitä ilmauksia kielenkäyttäjä kykenee ymmärtämään ja sitä myötä ajattelemaan. Tämä ei sisällä väitettä, että kaiken ajateltavissa olevan täytyisi olla kielellisesti ilmaistavissa, mutta mikäli kieltä käytetään ajatusten ilmaisemiseen, ja kaikki kieliopilliset lauseet ovat ajateltavissa, täytyy ajattelusysteemin olla vähintään yhtä ilmaisuvoimainen kuin kielen. Kielen produktiivisuudesta kyetään siis päättelemään ajattelun produktiivisuuteen, ja tämän askeleen myös Chomsky on ollut valmis ottamaan (Chomsky, 2006, s.88). On kuitenkin syytä huomata, että chomskylainen teoria tarjoaa idealisoidun kuvan kielellisestä kyvystä, joten argumentti ei takaa ajattelun rajatonta produktiivisuutta.

Ajattelun kieltä innokkaimmin markkinoineet teoreetikot erottavat sen luonnollisesta kielestä. Esimerkiksi Fodor on väittänyt Chomskyn töihin nojaten, että luonnollisen kielen oppiminen edellyttää kielioppia koskevien hypoteesien muodostamista ja testaamista, ja nämä hypoteesit puolestaan täytyy muotoilla jossain mentaalisessa representaatiojärjestelmässä, joka ei tietenkään voi olla tuo oppimisen kohteena oleva kieli (Fodor, 1975, s.56–57). Toisekseen luonnollisen kielen ilmaukset ovat moniselitteisiä. Esimerkiksi lause ”Mies löi poikaa sohvalla” voidaan tulkita kolmella eri tavalla. Ajattelun kielen lauseet taas eivät voi olla moniselitteisiä tai tulkinnanvaraisia, koska ne nimen omaan ovat noita tulkintoja (Pinker, 1997, s.70).

Myös tekoälytutkimuksen varhaisempien vuosien huippuhetket valoivat uskoa ajattelun komputationaaliseen teoriaan. Tekoälyn pioneerit Cliff Shaw, Allen Newell ja Herbert Simon työskentelivät 1950-luvun puolivälissä *Logic Theory Machine*, tai usein vain *Logic Theorist*, nimeä kantavan ohjelman parissa, jonka tarkoituksena oli mallintaa ihmisten ongelmanratkaisua (Newell & Simon, 1956). Ohjelma rajoittui todistamaan Russellin ja Whiteheadin *Principia Mathematican* teoreemoja teoksen loogista systeemiä käyttäen. Vuonna 1956 *Logic Theorist* todisti 38 *Principian* ensimmäisestä viidestäkymmenestä kahdesta teoreemasta. Lisäksi ohjelma jopa johti Eukleideen niin sanotulle aasinsiltateoreemalle yksinkertaisemman todistuksen, kuin mitä Russell ja Whitehead olivat kirjaansa painattaneet. (Boden, 2006, s.323–324) Seuraavaksi Simon ja Newell laativat melko optimistisesti nimetyn ohjelman *General Problem Solver* (GPS). Nimensä mukaisesti sen tarkoitus oli kyetä ratkaisemaan periaatteessa mitä tahansa luonteeltaan formaaleja ongelmia ilman ennakkoon asetettua rajausta tiettyyn ongelmatyyppiin ja formalismiin. Ohjelma kykenikin kohtuullisen monipuoliseen ongelmanratkontaan teoreemojen todistamisesta klassisiin päättelyongelmiin, kuten ”lähetyssaarnaajat ja kannibaalit”.[51]

Olivatpa nämä ohjelmat miten vakuuttavia mielenmalleja tahansa, niiden parissa tehty tutkimus valaisi merkittävästi komputationaalisen mielenteorian ehtoja yleisemmin. Molemmat edellä mainitut ohjelmat käyttivät heuristisia menetelmiä ongelmanratkonnassa. GPS:n toiminta perustui niin sanottuun keino–päämäärä-analyysiin. Menetelmää voi hahmottaa polun luomisena ongelman alkutilanteesta sen ratkaisuun. Ohjelma luo representaation esitetystä ongelmasta sekä hyväksyttävien ratkaisujen joukosta, joka voi olla yksikäsitteinen, kuten tietty teoreema. Ohjelmalla on joukko operaatioita, joita soveltamalla se voi muuttaa alkutilanteen toisenlaiseksi, ja edelleen näin aikaan saadun tilanteen taas uudeksi lopulta lähestyen tavoiteltavaa lopputilaa, eli ratkaisua. Operaatiot vaihtelevat riippuen ratkaistavasta ongelmasta, ja ne voivat käsittää esimerkiksi sallittujen siirtojen joukon shakkipelissä, päättelysäännöt deduktioissa ja niin edelleen. Menetelmä edellyttää jonkinlaista representaatioformaattia, jonka avulla periaatteessa mikä tahansa ongelma, sallitut operaatiot ja niiden soveltamisen aikaansaamat muutokset ovat esitettävissä. Tällaisessa järjestelmässä periaatteessa mikä tahansa ongelma voidaan esittää puuna, jossa oksat muodostuvat mahdollisten operaatioiden ketjuista ja solmukohdat mahdollisista tiloista. Puun päätepisteet taas edustavat ongelman ratkaisuja tai umpikujia, joista ei voida enää edetä. Menetelmän avainongelma on löytää ratkaisuun johtava oksa seurattavaksi. (Ernst & Newell, 1969, s.24–26,248–255)

Kuten ensimmäisessä luvussa mainittiin, mikä tahansa ratkaistavissa oleva formaalisti hyvin määritelty ongelma ratkeaa jollakin algoritmilla. Ikävä kyllä tällä huomiolla on usein enemmänkin teoreettista kuin käytännöllistä merkitystä. Jos algoritmin käyttämä aika suhteessa syötteen pituuteen kasvaa eksponentiaalisesti, ongelmaa voi pitää käytännössä ratkeamattomana, joskin jollain kevyemmällä algoritmilla voidaan mahdollisesti approksimoida ratkaisua (Hopcroft et al., 2001, s.413). Tekoälyn kannalta tämä on hyvin oleellista, koska yllättävän yksinkertaisetkin ongelmat voivat olla laskennallisesti hyvin raskaita. Kauppamatkustajan ongelma on tästä hyvin edustava esimerkki: Joukko kaupunkeja on yhdistetty mielivaltaisen pituisilla tieyhteyksillä, ja ongelmana on löytää lyhin kaikkien kaupunkien kautta kulkeva reitti. Kun kaupunkeja on n-kappaletta, mahdollisten reittien

[51] Ks. esim. (Ernst & Newell, 1969, Luku VI: ”Tasks Given to GPS”, s.125–246).

määrä on $n!$, eli kaupunkien kertoman verran. Koska reittejä kuitenkin on äärellinen määrä, ongelmalla on triviaali ratkaisu: yksinkertaisesti lasketaan kaikkien reittien pituudet ja valitaan niiden joukosta lyhin. Ikävä kyllä, kun n kasvaa suureksi, kaikkien reittien laskeminen vaatii kohtuuttomasti aikaa. Voidaan osoittaa, että tehokkain ongelman ratkaiseva algoritmi vaatii laskenta-aikaa eksponentiaalisesti suhteessa kaupunkien määrään.[52] Tekoälyn yhteydessä tästä aiheutuvaa ongelmaa kutsutaan usein kombinatoriseksi räjähdykseksi, joka tarkoittaa, että keino–päämäärä-menetelmän puut kasvavat hyvin äkkiä liian isoiksi, jotta ratkaisua voitaisiin etsiä yksinkertaisesti käymällä läpi kaikki mahdolliset oksat.

Newellin ja Simonin kehittelemä ratkaisu kombinatoriseen räjähdykseen oli *heuristiset algoritmit*.[53] Ne eivät generoi koko ratkaisupuuta vaan valitsevat joidenkin heuristiikkojen, eli nyrkkisääntöjen, nojalla puusta lupaavimmat polut ja jättävät muut huomiotta. Käytännössä tällaiset ohjelmat usein soveltavat jonkinlaista yritys–erehdys-menetelmää, eli mahdollisia oksia lasketaan vain tiettyyn syvyyteen asti, ja oksat, jotka eivät vähennä alkutilan ja ratkaisun välistä eroa, karsitaan pois. Heuristiset algoritmit pitävät ongelmanratkaisussa vaadittavat resurssit kurissa, mutta ikävä kyllä samalla menetetään takeet, että ohjelma löytää optimaalisen tai yhtään minkäänlaisen ratkaisun. Ohjelmoija joutuu tasapainoilemaan heuristiikkojen laskennallisen vaativuuden ja ongelmanratkaisukyvyn välillä, ja mitä yleisempää ongelmaluokkaa ohjelman on tarkoitus ratkoa, sitä hankalammaksi tämä tyypillisesti käy. Yksittäisissä ongelmissa menetelmä voi olla hyvinkin tehokas, mutta on hankalaa laatia heuristiikkoja, jotka toimisivat tehokkaasti periaatteessa missä tahansa tilanteessa. Yleisyys on ongelmallista myös toisella tavalla. Puuhastelu GPS:n parissa osoitti, että mitä yleisemmän ongelmanratkaisijasta pyrkii tekemään, sitä haastavampaa ohjelman sisäisen representaatiojärjestelmän laatimisesta tulee. Vastoin kuin ihmiset, GPS-tyyppinen ohjelma ei yleisesti ottaen kykene luomaan oikeanlaista representaatiota mielivaltaisesta ongelmasta, eikä päättelemään ratkaisun kannalta oleellisia aukkokohtia ongelman kuvauksessa. Mitä yleisemmän ongelmanratkojan on tarkoitus olla, tyypillisesti sitä täsmällisemmin ongelman ratkaisumenetelmä sille täytyy kuvata. Pahimmillaan ohjelmalle tulee selostaa oleellisesti koko ratkaisu, jotta se kykenisi edes aloittamaan prosessointia.[54] (Ernst & Newell, 1969, s.28–33,269–274)

Heuristiset algoritmit edustavat vähemmän idealisoitua kuvaa kognitiosta kuin esimerkiksi chomskylaiset kielioppiteoriat tai vastaavat kognitiivisten kykyjen teoreettista kompetenssia painottavat mallit. Heuristiset algoritmit toimivat rajoitetun rationaalisesti kuten myös ihmiset, joilla ei yleensä ole aikaa eikä kognitiivista kapasiteettia etsiä ongelmiin

[52] Tarkalleen ottaen kauppamatkustajan ongelma on niin sanottu NP-täydellinen ongelma, ja sikäli kun tiedetään, näiden ongelmien ratkaisu vaatii eksponentiaalisesti aikaa suhteessa syötteen pituuteen. On avoin kysymys, voidaanko NP-täydelliset ongelmat ratkaista polynomiaikaisilla algoritmeilla. (Hopcroft et al., 2001, s.413–421).

[53] Joskin on syytä mainita, että heuristisen ohjelmoinnin idea oli jo Turigilla 40-luvun alussa. Idean myöhemmin itsenäisesti keksineet Newell ja Simon kuitenkin tekivät sen laajemmin tunnetuksi. (Copeland, 2004, s.353–354)

[54] Vrt. ongelmaa esim. Turing-koneisiin. Universaali Turing-kone kykenee suorittamaan minkä tahansa laskennan, mutta tämä edellyttää, että koneelle annetaan varsinaisen syötteen lisäksi myös ratkaisualgoritmi. Toisaalta tiettyä funktiota laskeva kone ei tarvitse muuta kuin varsinaisen syötteen, koska algoritmi on koodattu siirtymäfunktioon. Toisaalta tällainen kone ei puolestaan kykene laskemaan mitään muuta funktiota, joten koneen autonomisuus on tietyllä tavalla käänteisessä suhteessa sen yleisyyteen.

parasta mahdollista ratkaisua. Itse asiassa muodollisesti parhaan ratkaisun etsiminen voi olla käytännössä irrationaalista: kauppamatkustaja ehtii vanhentua kuoliaaksi laskiessaan lyhintä reittiä sadan kaupungin läpi, mutta ei välttämättä menetä mitään merkittävää valitessaan melko mielivaltaisesti jokseenkin järjellisen vaihtoehdon. Heuristinen ohjelmointi on oivallus, joka nousee koneiden, ei ohjelmien, rajallisuudesta ja tuo komputationaalisen teorian askeleen lähemmäksi psykologista todellisuutta. Uskoa siihen, että heuristinen keino–päämäärä-analyysi todella on oikeansuuntainen malli ihmisen ongelmanratkonnasta, lisäsi Newellin ja Simonin koejärjestelyt, joissa he pyysivät opiskelijoita todistamaan lauselogiikan teoreemoja ja samalla kertomaan ääneen mitä he toimituksen aikana oikeastaan tekevät. Kognitiivisen psykologian metodina tällainen protokolla-analyysiksi kutsuttu menetelmä ei ehkä ole täysin aukoton, mutta mielenkiintoista kyllä, opiskelijat ja GPS vaikuttivat etsivän teoreemojen todistuksia suurinpiirtein samoilla tavoilla (Newell & Simon, 1961, s.289–293).

Muun muassa näistä tuloksista innostuneina Newell ja Simon päätyivät esittämään kuuluisan *fysikaalinen symbolisysteemi* -hypoteesinsa: fysikaalisella symbolisysteemillä on riittävät ja välttämättömät välineet yleisluontoiseen älykkääseen toimintaan. Tässä ”fysikaalinen symbolisysteemi” oleellisesti tarkoittaa tietokonetta ja ”yleisluontoinen älykäs toiminta” inhimillistä kykyä toimia jotakuinkin päämäärärationaalisesti periaatteessa missä tahansa tilanteessa (Newell & Simon, 1976, s.116). Oleellista tässä on tietokone*metaforan* korvaaminen väitteellä, että symbolirakenteiden manipuloiminen on älykkään toiminnan välttämätön edellytys. Lisäksi heillä oli tarjottavanaan hypoteesi mentaalisten prosessien täsmällisestä luonteesta, nimittäin heuristiset hakualgoritmit. Tähän on syytä kiinnittää huomiota, koska kyseessä on melko vahva väite, jonka mukaan kaikki mentaalinen toiminta on ongelmanratkaisua, tai ainakin laadullisesti sen kaltaista (*ibid.*, s.125–126). Sitten konetilafunktionalismin harvempi teoreetikko on ollut näin suorasanainen ja kirjaimellinen komputationalismin kannattaja. Hypoteesi ei suoraan samaista funktionalismia ja komputationalismia, mutta kiinnittää ne yhteen hyvin tiiviisti: funktionaaliset kausaalirakenteet ovat nomologisesti välttämättä komputationaalisia prosesseja.[55]

Edellä käsiteltyjen seikkojen lisäksi komputationalistinen teoria pyrkii myös selittämään, miten mentaalinen kausaatio voi perustua mielentilojen sisällöille. Muistutuksena ensimmäisestä luvusta, keskeinen oivallus tässä on, että kognitiivisen systeemin ajatellaan olevan fysikaalisesti implementoitu formaali systeemi, jonka toiminta voidaan laatia noudattamaan implementoidun kalkyylin semanttisia periaatteita. Tämä järjestely mahdollistaa merkityksiä noudattavan mekanismin olemassaolon. Olen toistaiseksi vältellyt kysymystä mentaalisten representaatioiden semantiikasta, eli siitä, että jos ajatukset ovat jonkinlaista formaalia koodia, niin miten tämä koodi oikein saa merkityksensä? Kovan linjan formalisti voi koittaa sivuuttaa tämä ongelman sillä perusteella, että jos formaalikielille on olemassa jokin yleinen semanttinen teoria, kuten esimerkiksi tarskilainen malliteoria, niin kysymys on sitä myöten ratkaistu, koska ajattelun kieli on vain formaalikielten erityistapaus. Jos taas mitään tyydyttävää yleistä teoriaa ei ole, niin samasta syystä tämä

[55]Tarkemmin hypoteesia käsitellään Newellin huomattavasti laajemmassa artikkelissa ”Physical Symbol Systems”, jossa esitetään, että mahdollisuus liittä periaatteessa mikä tahansa vaste (reaktio) mihin tahansa syötteeseen (havainto) on älykkään käyttäytymisen edellytys. Tässä yhteydessä Newell myös väittää, ettei ole käsitteellinen vaan empiirinen tosiasia, että tällainen universaalisuus edellyttää symbolisysteemiä (Newell, 1980, s.147,154–155).

ei varsinaisesti ole kognitiotieteilijöille kuuluva ongelma. Heidän tehtävänsä on selittää, miten mieli toimii tavalla, joka noudattaa rationaalisuutta ja mentaalisten representaatioiden oletettuja sisältöjä. Komputationalismi selittää nämä seikat, mutta teorian mukaan varsinaisesta kausaalista työtä eivät tee merkitykset – ainakaan jos merkityksellä tarkoitetaan representaation viittaussuhdetta maailmaan – vaan päänsisäiset syntaktiset komputationaaliset prosessit, joiden kannalta symbolien merkitykset ovat samantekeviä. Näin ollen semantiikalla ei ole mitään suoranaista tekemistä mentaalisen kausaation eikä käyttäytymisen etiologian kanssa, joten kognitivistin ei tarvitse olla merkityksistä kovin kiinnostunut.[56] John Haugelandin sanoin formalistin motto kuuluu, että jos pidetään huoli syntaksista, niin semantiikka pitää huolen itsestään.

Tällainen suhtautuminen ei kuitenkaan ole kovin tyydyttävää. Ensinnäkin esimerkiksi tarskilainen malliteoria määrittelee vain loogisten vakioiden tulkinnat, eikä ota kantaa kielen atomisten ilmausten merkityksiin, eli teoria sallii mielivaltaisen tulkinnan formaalikielten ei-loogista sisältöä kantaville symboleille. Käytännössä tämä tarkoittaa, että esimerkiksi kissoihin ja koiriin viittaavat symbolit voidaan periaatteessa aivan vapaasti tulkita viittaavan vaikka sammakoihin ja alkulukuihin. Malliteoria sinänsä ei siis määrää termien referenttejä. Mitä taas tulee kalkyylin noudattamiin semanttisiin periaatteisiin, niin se vain tarkoitta, että deduktiosysteemin päättelysäännöt noudattavat loogisten vakioiden tulkintaa. Ei-loogisten symbolien semanttisiin ominaisuuksiin tämä ei liity mitenkään. Kuitenkin juuri ne ovat mielenkiintoisia, koska niiden yleensä ajatellaan vastaavan varsinaisia käsitteitä. Näin ollen formaalikieliin liittyvät matemaattiset teoriat eivät erityisemmin auta selventämään mielenteorian kannalta mielenkiintoista merkityksen ongelmaa.

Toisaalta väite, ettei kognitivistin tarvitse antaa selontekoa mielensisällöistä, on hieman uskottavampi mutta ongelmallinen sekin. Ensinnäkin on epäselvää, miten kognitiivinen teoria voisi selittää merkityksille perustuvan mentaalisen kausaation ilman minkäänlaista teoriaa mielensisällöistä. Miten reduktiivinen selitys olisi mahdollinen ilman minkäänlaista käsitystä siitä, mitä ollaan redusoimassa? Mikäli mielenteoria ei tarjoa selontekoa mentaalisten representaatioiden propositionaalisesta sisällöstä, on vaikea nähdä miten sillä voisi olla mitään tekemistä propositionaalisten asenteiden kanssa, jolloin herää kysymys miksi ylipäätään kannattaa kognitivismia? Tällöin kyseinen teoria vaikuttaisi ylimääräiseltä kyhäelmältä jossain aivotutkimuksen ja intentionaalisen psykologian välissä. John Searle lienee tunnetuin kriitikko, joka on koittanut osoittaa, ettei komputationaalinen teoria periaatteessakaan kykene selittämään mielensisältöjä. Hänen mukaansa komputationaaliset systeemit ovat läpikotaisin syntaktisia, eikä pelkkä syntaksi riitä antamaan symbolirakenteille merkityksiä.[57] Hänen mukaansa mentaalinen kausaatio kuitenkin perustuu merkityksille, joten komputationalismi ei voi olla oikea mielenteoria. (Searle 1980,1984,

[56]Ks. (Stich, 1983, s.170–183) ja myös (Putnam, 1975b), missä merkitysten esitetään olevan kognitiivisista prosesseista erillisiä ja (Fodor, 1981, s.233–241,253), jossa tästä päätellään, ettei mentaalisten representaatioiden semanttisilla ominaisuuksilla näin ollen ole tekemistä psykologisten yleistysten kanssa.

[57]Ks. (Searle, 1984, s.39), missä hän ilman kummempia perusteluja sanoo tämän olevan käsitteellinen tosiasia. Ilmeisesti Searle tarkoittaa, että suoraan määritelmällisesti symboli on merkki, jonka referentti on mielivaltainen. Juuri tästä syystä tarskilainen semantiikka ei voi määritellä atomisten symbolien referenttejä, koska mikään merkissä c tai merkkijonossa ”kissa” ei pakota näitä symboleja viittaamaan

s.38–41.) Searlen argumentit ovat herättäneet melko laajaa keskustelua, johon en tässä aio kajota. Joka tapauksessa kantavaa hänen kritiikissään on, ettei mielensisältöjä voida kognitivismissa sivuuttaa eikä olettaa, vaan ne tulisi jotenkin todella selittää. En myöskään ryhdy sen syvällisemmin käsittelemään kysymystä mentaalisesta representaatiosta, koska mielensisältöjen teoriat ovat jokseenkin ongelmallisia ja niiden tarkempi käsittely menee äkkiä hyvin monimutkaiseksi. Tehdään kuitenkin pikainen katsaus pariin perusteoriaan.

Eräs melko tyypillinen tapa hoitaa ongelma pois päiväjärjestyksestä on oleellisesti malliteoreettinen. Ajatus on, että mentaaliset representaatiot jakaantuvat atomisiin ja molekulaarisiin, vastaavasti kuin esimerkiksi logiikan lauseet, ja samoin kuin logiikassa, molekulaaristen representaatioiden semantiikka palautuu niiden rakennetekijöiden sisältöihin. Esimerkiksi ilmauksen ”siivekäs hevonen” merkitys voidaan palauttaa termien ”siivekäs” ja ”hevonen” merkityksiin. Erityistä semanttista teoriaa tässä vaatii atomisten representaatioiden tulkintojen kiinnittäminen, joka vastaa jotakuinkin samaa, kuin malliteoreettisten kielten tulkintafunktion määrittely. Erona tässä on vain, että mentaaliset representaatiot viittaavat maailman – eikä jonkin mielivaltaisen mallin – olioihin, olioiden ominaisuuksiin ja niiden välisiin relaatioihin. Molekulaaristen representaatioiden palauttaminen rakenneosiinsa taas perustuu kognitiivisen systeemin suorittamaan symbolirakenteiden syntaktiseen analyysiin, joten jos atomisten symbolien merkitys saadaan kiinnitettyä, loppu on puhdasveristä kognitivismia.

Ehkä suosituin tapa tällaisten teorioiden puitteessa on perustaa atomisten representaatioiden semantiikka jonkinlaiselle kausaalisuhteelle. Karkeasti ottaen tämä tarkoittaa, että esimerkiksi representaatio α viittaa kojootteihin, jos ja vain jos α:n esiintymät kognitiivisessa systeemissä riippuvat lainomaisesti tiettyjen kojooteille ominaisten ominaisuuksien esiintymistä maailmassa. Yleisesti ottaen siis representaatio α tarkoittaa y:tä, jos y:n esiintymät lainomaisesti aiheuttavat α:n esiintymiä. Tämän suuntaista lähestymistapaa ovat kehitelleet esimerkiksi Fodor (1990) ja Fred Dretske (1981).

Pääasialliset ongelmat tällaisissa teorioissa liittyvät abstrakteihin käsitteisiin ja niin sanottuun disjunktio-ongelmaan. Jälkimmäinen tarkoittaa, että representaation esiintymä voi lainomaisesti aiheutua muustakin kuin sen oletetusta referentistä. Esimerkiksi joku voisi systemaattisesti erehtyä luulemaan pimeässä näkemiään kettuja koiriksi. Tällöin representaation α esiintyminen on lainomaisesti riippuvainen koirista, mutta myös pimeällä esiintyvistä ketuista. Lisäksi α:n esiintymä voitaisiin ehkä systemaattisesti aiheuttaa myös vaikkapa stimuloimalla aivoja mikroelektrodeilla tai joillain psykedeelisillä huumeilla. Tällöin siis teorian mukaan representaatio α tarkoittaa koiraa tai kettua pimeällä tai tietynlaista aivojen stimulointia ja niin edeleen. Koska teoria perustaa representaatioiden viittaussuhteen lainomaiselle suhteelle, joka käytännössä tarkoittaa ainakin kontrafaktuaalien noudattamista, niin ei ole väliä miten representaation esiintymät ovat kullakin tosiasiassa syntyneet. Jos esiintymät lainomaisesti voitaisiin aiheuttaa tällaisilla erikoisilla tavoilla, nämä erikoiset kausaaliset lähteet osaltaan sisältyisivät esimerkiksi koiran käsitteeseen, mikä on absurdia. Tästä syystä kausaalisuhde sinänsä tuskin riittää representaatioiden sisällön määrittämiseen.[58] Abstraktien käsitteiden tapauksessa taas kausaalisuhde

kissoihin tai yhtään mihinkään. Vastaavasti esimerkiksi se, että luonnollisen kielen sanojen viittauskohteet eivät ole täysin mielivaltaisia, seuraa sosiaalisista käytännöistä, ei näistä sanoista itsestään.

[58] Ks. (Fodor, 1990, 38–42,90–100) ongelman mahdollisesta ratkaisusta.

ei voi olla välttämätön. Kukaan tuskin on koskaan ollut kausaalisessa suhteessa esimerkiksi demokratiaan tai alkulukuihin. Tietysti olemme kausaalisessa vuorovaikutuksessa esimerkiksi demokratiaan kuuluvien instituutioiden kanssa, esimerkiksi äänestyskopissa, mutta nämä tapahtumat eivät varmasti lainomaisesti aiheuta representaation *demokratia* esiintymiä mielissämme. Sitä paitsi tämä on eri asia, kuin olla kausaalisessa kosketuksessa demokratiaan sinänsä. Samoin alkulukujen käsitteemme on ehkä peräisin kirjoista, koulusta tai mistä nyt kullakin, mutta mielemme eivät varmaankaan ole kausaalisesta yhteydestä itse lukuihin. Kukaan ei halua teoriaa, jonka mukaan ajatuksilla alkuluvuista ei itse asiassa ole mitään tekemistä lukujen ja jaollisuuden vaan kirjojen ja peruskoulun kanssa. Näin ollen kausaalisuhde representaation ja sen referentin välillä ei voi olla välttämätön ainakaan kaiken mentaalisen sisällön kannalta.

Toinen suosittu mielensisältöjen teoriakehitelmä kumpuaa logiikan todistusteoriasta ja funktionalismista. Tällaiset teoriat kulkevat nimillä *proseduraalinen* tai *funktionaalinen semantiikka*.[59] Näissä teorioissa representaatioiden sisältö määräytyy sen perusteella, miten niitä käytetään kognitiivisessa systeemissä. Käytännössä tämä tarkoittaa, että sisältö riippuu representaation deduktiivisista suhteista muihin representaatioihin. Loogisten käsitteiden kanssa tämä tekniikka toimii vallan mainiosti. Esimerkiksi päättelysäännöt, joiden mukaan kaavoista α ja β voidaan päätellä $\alpha \wedge \beta$ ja kaavasta $\alpha \wedge \beta$ sekä α että β, määrittelevät melko yksiselitteisesti, että symboli $\wedge$ vastaa konnektiivia ”ja”. Ajatellaanpa, että kognitiivinen systeemi sisältää seuraavat päättelysäännöt: jos x on hevonen ja x on siivekäs, niin x on α; ja jos x on α, niin x on hevonen ja x on siivekäs. Tällöin ilmeisesti predikaatti α vastaa Pegasuksen käsitettä. Joidenkin päättelysääntöjen täytyy olla sisäsyntyisiä, vastaavasti kuin loogisissa systeemeissä joidenkin päättelysääntöjen tulee olla perustavanlaatuisia, jotta mitään päättelyjä ylipäätään voidaan tehdä. Tällaiset säännöt, tai oikeastaan niitä vastaavat mekanismit, muodostavat kognition perusarkkitehtuurin. Oletettavasti nämä mekanismit vastaavat lähinnä loogisia päättelysääntöjä, koska esimerkiksi Pegasuksen käsite tuskin on sisäsyntyinen ja sama lienee totta lähestulkoon kaikista ei-loogisista käsitteistä. Systeemiin voidaan lisätä päättelysäätöjä teorioiden muodossa. Jos teoria sisältää implikaatiolauseen $\alpha \rightarrow \beta$, niin logiikan deduktiolauseen perusteella tämä vastaa täsmälleen päättelysääntöä $\alpha \vdash \beta$. Näin ollen proseduraalisessa semantiikassa käsitteet ovat olemukseltaan eräänlaisia pienoisteorioita.

Puhtaassa muodossaan proseduraalinen semantiikka on täysin syntaktinen teoria, jonka mukaan kaikkien käsitteiden sisällöt määräytyvät niiden välisten deduktiivisten suhteiden perusteella. Käsitteet siis muodostavat eräänlaisen päättelysuhteiden verkoston, ja representaation sisällön määrittää sen paikka tuossa verkossa. Kyseessä on siis lähes samanlainen teoria mielensisällöistä kuin funktionalismi on mielentiloista. Vastaavasti teoria perii funktionalismin yhteydessä esiin nousevat holismiin liittyvät ongelmat. Mikäli kahdella oliolla käsitteiden välinen päättelyverkosto poikkeaa toisistaan, niillä ei teorian kirjaimellisen tulkinnan mukaan voi olla mitään samoja käsitteitä. Lisäksi kaikenlaiset eri-

[59] Engl. *procedural*, *conceptual role* tai *functional role semantics*. Tarkempia lähteitä ei ole mainittu kappaleissa, mutta hyviä perusesityksiä teoriasta tarjoavat esimerkiksi (Harman, 1987) ja (Rapaport, 1995). Jälkimmäinen sisältää melko pitkälle viedyn teoriakehitelmän syntaksiin pohjautuvasta semantiikasta ja ottaa kantaa aiemmin sivuttuihin Searlen argumentteihin. Huomautettakoon, että proseduraalisen semantiikan ydinajatukset muistuttavat kovasti myös Wittgensteinin myöhäistä kielikäsitystä.

koislaatuiset yksilölliset uskomukset kuuluvat osaksi käsitteiden verkkoa. Esimerkiksi jos joku uskoo, että lehmät ovat vaarallisia, hänellä ilmeisesti on sääntö ”jos x on lehmä, niin x on vaarallinen”, ja tällöin vaarallisuus kuuluu hänen lehmän käsitteeseensä. Näin ollen proseduraalisessa semantiikassa on hyvin vaikea erotella käsitteellistä sisältöä mielivaltaisista uskomuksista. Yhdessä holismin kanssa tällä on ikävä taipumus myrkyttää koko käsitejärjestelmä, koska periaatteessa mikä tahansa uskomus vaikuttaa koko käsitteiden verkostoon.[60]

Hieman vakavampi ongelma liittyy käsiteverkoston ja sen tulkinnan mielivaltaisuuteen. Ajatellaanpa, että kognitiivinen systeemi sisältää esimerkiksi sivulla 31 esitellyn teorian aurinkokunnasta. Nyt ongelmaksi muodostuu, että teorian mallit voisivat hyvin olla isomorfisia vaikkapa jonkin lennokkikerhon osallistujia kuvaavien mallien kanssa. Teorian termit voidaan aivan hyvin tulkita esimerkiksi, että a =”lennokkikerhon vetäjä”, $P(x)$ =”x on kerholainen”, $L(x, y)$ =”x on pitempi kuin y” ja niin edelleen. Jos eri teorioiden mallit ovat isomorfisia, ei ole mitään keinoa erottaa kumpi teoria koskee kumpaakin ilmiötä. Lisäksi jos sama teoria soveltuu kuvaamaan kahta eri ilmiöluokkaa, proseduraalinen semantiikka ei pysty erottelemaan kumpaan ilmiöluokkaan teorian termit viittaavat. Toisaalta tässä kohtaa holismista on hyötyä. Nimittäin aurinkokuntia ja lennokkikerhoja koskevat pienoisteoriat sijoittuvat jollain tavalla organismin koko käsiteverkossa, ja esimerkiksi planeetan ja lennokkikerholaisen käsitteillä lienee erilaiset deduktiiviset suhteet useimpiin käsiteverkoston osiin, joten ne eivät ole keskenään vaihdettavissa. Tällöin oleellinen kysymys kuuluu, voidaanko koko käsitejärjestelmä tulkita mielivaltaisilla tavoilla. Kovin monimutkaisessa järjestelmässä tämä ei liene mitenkään itsestäänselvää. Suhteellisen suosittu ratkaisu näihin vaivoihin on yhdistää kausaaliteorioita proseduraaliseen semantiikkaan. Tällaisen niin sanotun kaksoisfaktoriteorian mukaan joidenkin representaatioiden referentit ankkuroidaan kausaaliteoreettisen mallin mukaisesti ja varsinainen käsitteellinen sisältö taas määritellään proseduraalisen semantiikan tapaan käsitteiden välisten deduktiivisten suhteiden perusteella. Koska tietyt käsitteet ovat kiinnitetty, käsitteiden verkkoa ei voi tulkita aivan miten tahansa (Block, 1986, s.108–109).

Tämä esitys riittänee antamaan jonkinlaisen käsityksen, miten kognitivismin puitteissa mielentilojen sisällön ongelmaa on koitettu ratkoa, tai vähintään uskottelemaan, että ongelma on ainakin joissain piireissä otettu vakavasti. Lukija voi perehtyä keskusteluun annettuja viitteitä seuraten ja vetää johtopäätöksensä näiden teorioiden uskottavuudesta. Jatkon kannalta on kuitenkin tarpeen huomata, miten mentaalisen representaation ongelma on kognitivismin puitteissa pyritty ratkaisemaan oleellisesti loogisin välinein. Voidaanko mentaalisia prosesseja pitää kirjaimellisesti laskentana, riippuu osittain siitä, kyetäänkö mielensisällöt selittämään komputationalismin puitteissa vai ei.

Vielä vuonna 1968 Jerry Fodorin vaikutusvaltaisessa teoksessa *Psychological Explanation* komputationaalista psykologiaa käsittelevä luku oli nimeltään ”The Logic of Simulation”. Jotkut taas, kuten Newell ja Simon, ovat kannattaneet komputationaalisen teesin kirjaimellista tulkintaa jo 50-luvulta lähtien, joskin hekin olivat lausunnoissaan hieman varovaisempia ennen vuonna 1975 pitämäänsä *Turing Award* luentoa (Newell & Simon, 1976), jossa he esittivät symbolisysteemihypoteesinsa. Esimerkiksi merkkiteoksessaan *Hu-*

[60]Ks. (Block, 1995), jossa tätä ongelmaa pyritään myös ratkaisemaan.

man Problem Solving vuodelta 1972 he esittivät melko nöyrästi, että ihmisten toimintaa voidaan kuvata symbolirakenteiden käsittelynä ainakin silloin, kun he ratkovat ongelmia (s.9,788). Kuitenkin 70–80-lukujen taitteessa tietokoneanalogia ei näyttäytynyt enää pelkästään metodologiana psykologian tekemiselle tai heuristiikkana mielen ja ruumiin suhteen ymmärtämiselle. Muun muassa ajattelun kielen teorian ja symbolisysteemihypoteesin myötä kognitivismin filosofinen ydin kiteytyi: mielet ovat kirjaimellisesti symboleja käsitteleviä virtuaalikoneita. 80-luvun aikana tämä itsemäärittely sai lisäpontta uudenlaisten komputationaalisten mallien noustessa kognitiotieteen valtavirtaan. Uudet neuroverkkomallit enteilivät jonkinlaista kognitiotieteiden vallankumousta, ja korvasivat kognitivismin avainsanat *symbolirakenteet*, *deduktiosysteemit* ja *heuristiset hakualgoritmit* uusilla: *hajautetut representaatiot*, *lineaarialgebra* ja *hahmontunnistus*. Uusi kognitiotiede oli kuitenkin ytimeltään vanhaa komputationalismia. Näin ollen niin vanhan koulukunnan kuin uuden aallon edustajatkin ryhtyivät määrittelemään, mikä heidän kannattamassan teoriassa on oleellista sekä miten vanha ja uusi komputationalismi tarkalleen ottaen eroavat toisistaan.[61] Filosofisesti kaikki tämä on kovin hedelmällistä, koska mitä kirjaimellisemmin tietokoneanalogia otetaan ja mitä selvemmin kognitivistisen teorian ydin määritellään, sitä helpommaksi tulee sen heikkojen kohtien havaitseminen ja kritiikin kohdistaminen. Perehdytään seuraavaksi tähän puoleen.

4.2 Kaksi askelta eteenpäin, yksi taaksepäin

Edellisessä luvussa oltiin kovin huolissaan mustekalojen ja marsilaisten mielentiloista, mutta miten lienee omamme laita? Nykyisistä kognitiivisen psykologian oppikirjoista löytyy monipuolisesti tietoa havaintojärjestelmien toiminnasta, hahmontunnistuksesta, ongelmanratkaisusta, päätöksenteosta ja sen sellaisista. Poissaolollaan loistaa propositionaalisten asenteiden teoria, mielen yhtenäisteoria sekä varsinainen ajattelun teoria siinä mielessä, mitä klassisen komputationalismin kannattajat tällä tarkoittavat. Yleisesti ottaen mitä korkeammista kognitiivisista prosesseista ja mitä keskeisimmistä representaatioista on kyse, sitä spekulatiivisemmaksi kirjoittelu käy. Edellisen luvun sivulla 76 esiteltiin uskomusten ja halujen vuorovaikutuksia koskevia tutkimuksia, mutta ne eivät varsinaisesti kuulu kognitiivisen- vaan sosiaalipsykologian alaan, missä uskomukset ja halut oletetaan eikä niinkään selitetä. Mielenfilosofian kannalta psykologiaa mielenkiintoisempi tutkimusohjelma kuitenkin on tekoälytutkimus. Tekoälyä voi luonnollisesti tarkastella puhtaasti teknologiana, mutta sivutetaan tämä näkökulma ja kiinnitetään huomiota tekoälysysteemeihin mielen toiminnan malleina tai ainakin osittaisina mielenteorioina. Tekoälyn filosofinen mielenkiinto perustuu pitkälti siihen, että mieltä mallintava ohjelmoija joutuu tekemään hyvin täsmällisiä hypoteeseja mielen arkkitehtuurista sekä tosiasiassa muodostamaan ne tietorakenteet ja heuristiikat, joita systeemi käyttää. Tekoälyohjelman laatimista ja ajamista voidaan pitää ohjelman mallintamaa psykologista ilmiötä koskevan teorian muodostuksena ja testaamisena. Erityisesti varhainen tekoälytutkimus oli paljolti suuntautunut korkeisiin kognitiivisiin prosesseihin, terveen järjen mallintamiseen, kielen-

[61] Esim. (Fodor & Pylyshyn, 1988) on eräs tunnetuimmista neuroverkkomallien kritiikeistä. Jälkikäteen katsottuna tämän artikkelin filosofisesti kantavin sisältö kuitenkin liittyy klassisen komputationalismin ydinteesien mahdollisimman täsmälliseen määrittelyyn. Aiheesta tarkemmin ks. myös (Boden, 1991).

käyttöön ja muuhun yleisluontoiseen älykkyyteen liittyvään toimitaan, mikä on filosofisen analyysin kannalta erityisen mielenkiintoista.

Lisäksi kiinnostavaa on varhaisten tekoälytutkijoiden optimismi ja hankkeen mahdollisuuksia koskevien väitteiden provokatiivisuus. Jo Turing ennusti, että Turing-testissä jokseenkin hyvin pärjäävä kone on mahdollista rakentaa 1900-luvun loppuun mennessä. Hän ei kuitenkaan ollut yltiöpäisen optimistinen, vaan arveli, että noin 70 % koneen kuulustelijoista ei pysty viidessä minuutissa erottamaan keskusteleeko hän ihmisen vai koneen kanssa (Turing, 1950, s.442,455). Vuonna 1957 Simon nosti panoksia väittämällä, että jo tuolloin oli olemassa ajattelevia koneita, ja kymmenen vuoden sisään tietokone on voittanut shakin maailmanmestaruuden, keksinyt ja todistanut uuden merkittävän matemaattisen tuloksen sekä säveltänyt esteettisesti merkittävän teoksen (Simon & Newell, 1958, s.7–8). Vaikka mihinkään näistä koneet eivät kyenneet määräajassa, vielä vuonna 1965 Simon esitti, että 80-luvun puoliväliin mennessä tietokoneet pystyvät mihin tahansa työhön mihin ihminenkin (Boden, 2006, s.716). Lisäksi vuonna 1970 *Life*-lehden haastattelussa merkittävä tekoälytutkija Marvin Minsky ennusti koneen kykenevän muutaman vuoden sisällä kaikkeen mihin ihminenkin ja hyvin pian jopa ylittävän inhimillisen älykkyyden (Darrach, 1970, s.58D).

Yllä olevista lupauksista yksi on tavallaan toteutunut, joskin kolme vuosikymmentä myöhässä. Tämä tapahtui 11. toukokuuta vuonna 1997, kun IBM:n Deep Blue -kone voitti vallitsevan shakin maailmanmestarin Garri Kasparovin kuuden ottelun pelissä (Boden, 2006, s.15). Kriteerit esimerkiksi matemaattisen tuloksen merkittävyydelle, sävellyksen esteettisyydelle ja yleisluontoiselle älykkyydelle ovat jokseenkin epäselvät, mutta selvää on, ettei tekoäly ole lunastanut lähimainkaan lupauksiaan yleisluontoisen älykkyyden tai ihmisjärkeen verrattavan käyttäytymisen saralla. Puhtaasti teknologisena projektina tekoäly lienee laskettavissa menestystarinoihin, mutta mielenteorian kannalta oleellisen tekoälytutkimuksen paikoillaan polkeminen on kylvänyt tappiomielialaa ja epäuskoa kognitivismiin sekä erityisesti klassiseen komputationalistiseen teoriaan.[62] Voi tietenkin olla, että kognitivistinen taustateoria sinänsä on oikeansuuntainen, mutta tavoitteeseen pääsy on vain osoittautunut oletettua työläämmäksi. Toisaalta Newellin ja Simonin ensimmäisistä ”ajattelevista koneista” on kulunut jo yli puoli vuosisataa, jona aikana koneet ja ohjelmointitekniikat ovat kehittyneet valtavasti, kuten on myös tietämys ihmisen kognitiivisesta systeemistä. Tästä huolimatta tekoälyn laitimmainen tavoite on jäänyt kauas saavuttamattomiin, eikä kognitivistisen teorian puitteissa tehty tutkimus ole sanottavasti valottanut yleisluontoisen inhimillisen älykkyyden eikä kaikkein korkeimpien kognitiivisten prosessien syvintä olemusta. Mikäli mielenteorian pätevyyttä tulee lopulta arvioida näyttöjen eikä lupaavien ideoiden perusteella, kognitivismin tila ja tulevaisuus ei vaikuta aivan parhaalta mahdolliselta. Tekoälytutkimuksen paljastamat kognitivismin keskeisimmät ongelmat voidaan nähdäkseni lohkoa kolmeen ryhmään: tunteiden asema ajattelussa ja rationaalisuudessa, taidokas käyttäytyminen ja proseduraalinen tieto sekä relevanssin ongelma.

Äkkiseltään katsottuna representationaalinen mielenteoria, ja sen päälle rakentuva kognitivismi, ei vaikuta kovin luontevalta teorialta kattamaan tuntemuksia, tunnelmia ja mielia-

[62] Ks. (Boden, 2006, s.1105–1109) pikaista tilannekatsausta varten.

loja. On tietysti mahdollista, että esimerkiksi peloissaan oleminen edellyttää jonkinlaisia mentaalisia representaatioita esimerkiksi pelon kohteesta, joten kognitivisti voinee ottaa työhypoteesikseen, että pelkäämisen vaikutus käyttäytymiseen välittyy normaaliin tapaan pelkoon liittyvien representaatioiden komputaatioina. Pelkoon liittyvät ulkoiset reaktiot, kuten pakeneminen tai välttely, voivat kuulua komputationaalisen teorian piiriin, mikäli niitä pidetään kognitiivisen järjestelmän tulosteena. Pelkoon tietysti voi liittyä myös kaikenlaisia fysiologisia ilmiöitä, kuten pulssin nousua ja käsien tärinää, jotka eivät liity kognitiivisen systeemin toimintaan eivätkä tällöin varsinaisesti kuulu psykologian alaan. Näin kognitivisti voisi ehkä kyetä eristämään tunteista nimenomaisesti psykologiset ulottuvuudet ja onnistua osoittamaan, että ne voidaan uskottavasti kuvata kognitiivisen teorian puitteissa. Toisaalta taas esimerkiksi päiviä tai viikkoja kestävä yleinen huolestuneisuus, tai aikaskaalan toisessa päässä akuutti paniikki, voisivat olla tiloja, joihin mahdollisesti liittyvät mentaaliset representaatiot ovat melko epäoleellisia mutta fysiologiset ilmiöt taas hyvinkin oleellisia. Nämä siis ovat esimerkkejä mentaalisista episodeista, jotka melko selvästi eivät ole kognitiivisia prosesseja, eli symbolirakenteiden käsittelyä. On vaikea sanoa ovatko huoli ja paniikki varsinaisia mentaalisia prosesseja, mutta ainakin ne selvästi vaikuttavat mielen toimintaan sekä käyttäytymiseen. Kognitivismin kielelle käännettynä monet tunnetilat siis vaikuttavat syötteiden käsittelyyn ja lähes kaikkiin tulosteisiin. Mikäli näin on, kognitivismi on puutteellinen mielenteoria, koska kaikki mentaaliset ilmiöt eivät ole kognitiivisia prosesseja, joten käyttäytymistä koskevia lainomaisia yleistyksiä ei voida johtaa yksin kognitiivisen systeemin toiminnasta.[63]

Toisaalta pakoteitä tästä ongelmasta löytyy parikin. Ensinnäkin mielialat eivät välttämättä ole kognitiivisia prosesseja sinänsä, mutta ehkä ne liittyvät siihen, miten kognitiivisia prosesseja suoritetaan, esimerkiksi miten toiminnanohjaus priorisoi haluja ja päämääriä. Välittömän vaaran pelko ohittanee pitkän tähtäimen tavoitteiden toteuttamiseen, joten mahdollisesti tunteita ja mielialoja voitaisiin mallintaa tällä tavoin. Ikävä kyllä tällaista teoriaa ei juurikaan ole tarjolla. Itse asiassa varhaisessa tekoälytutkimuksessa erinäköisiä emotionaalisia prosesseja pyrittiin mallintamaan, mutta ohjelmat olivat hyvin yksinkertaistavia, psykologisesti epäuskottavia ja katosivat varhain valtavirrasta (Boden, 2006, s.368–394). Tehtävä osoittautui liian vaikeaksi ja se pääpirteissään unohdettiin. Asiaan on onneksi ajoittain palattu ja esimerkiksi Rosalind Picard (1997) on kirjoittanut kattavahkon teoksen tunteista kognitiivisina ilmiönä.

Toinen vaihtoehto on yksinkertaisesti kieltää tunteiden olevan ongelma kognitivismille. Kognitivisti voi katsoa teoriansa koskevan ajattelua ja järjenkäyttöä, joilla tunnetusti ei ole tunteiden kanssa juuri muuta tekemistä, kuin että viimeksi mainitut ajoittain sumentavat edelliset. Harmi kylläkin, ettei tämä pidä paikkaansa. Ihmiset, joiden tunnemaailma on elimellisesti vammautunut esimerkiksi aivovaurion seurauksena, saattavat säilyttää perinteisesti kognitiivisiksi mielletyt kykynsä lähes ennallaan ja järkeillä jokseenkin normaalisti, mutta heillä on taipumus toimia hyvin irrationaalisesti. Tyypillisiä ongelmia ilmenee riskien hallinnassa, päätöksenteon vaikeudessa, toiminnan mielekkyyden ja seurausten arvioinnissa, virheistä oppimisessa ja niin edelleen (Damasio, 2006, s.35–36,52–53). Tunteilla vaikuttaisi olevan toiminnanohjauksessa heuristiikkojen kaltainen rooli ohjata käyttäytymistä eliminoimalla huonoon tai toiminnan päämäärän kannalta epäoleelliseen lopputu-

[63] Tämänsuuntaisen kritiikin esittäjistä tunnetuimpia lienevät John Haugeland (1978, s.270-272; 1985, s.230–238) ja Antonio Damasio (2006).

lokseen johtavat toimintavaihtoehdot. Tunteet ilmeisesti siis ikäänkuin kerovat kognitiolle mikä on hyvä ja mikä huono idea.

Kognitivismin suhteen tässä hyvin ongelmallista on, että tunteet ja tuntemukset ovat kognition lisäksi hyvin tiiviissä suhteessa ruumiiseen. Fysiologiset reaktiot eivät vain korreloi tunteiden kanssa, vaan hyvin pitkälle konstituoivat ne (Picard 1997, s.30–35, Damasio 2006, s.165–189). Näin ollen mikäli kognitiivinen systeemi ei kirjaimellisesti ulotu vatsan pohjaan saakka, tunteet eivät palaudu kognitiivisiin prosesseihin. Tällöin taas kognitivismi ei edusta vain rajallista kuvausta mentaalisista ilmiöistä, vaan myös puutteellista teoriaa mielekkään käyttäytymisen etiologiasta. Toisin sanoen kognitiivisen teorian puitteissa ei voida muodostaa päteviä yleistyksiä käyttäytymisestä edes niillä alueilla, joissa kognitivismin pitäisi olla vahvimmillaan, eli päätöksen teossa, toiminnan ohjauksessa ja niin edelleen. Huomautettakoon, että tämä ei ole vain komputationalismin ongelma vaan asettaa myös funktionalismin perin epäilyttävään valoon. Jos kognitiivisen systeemin kausaalinen rakenne on riippuvainen tietynlaisesta fysiologiasta, on epäselvää voivatko ruumiillisesti hyvin erilaiset oliot, ääritapauksessa esimerkiksi ihminen ja tietokone, toteuttaa samoja psykologisia predikaatteja. Onko esimerkiksi mahdollista, että haluja voi olla ilman tunteita? Homunkulaarisen funktionalismin mukaisesti korkea kognitiivinen toiminta ei edellyttäne kuitenkaan kaikilta samanlaista kehoa, kunhan sen funktionaalinen rooli kognitiivisessa toiminnassa korvataan jollain tavalla. Näin ollen tekoäly ei välttämättä edellytä tekokehoa, mutta tutkimuksen ongelmat saattavat ainakin osittain kummuta tunteiden, ja sitä myötä mahdollisesti myös kehollisuuden, merkityksen huomiotta jättämisestä inhimillisessä mielessä normaalin kognition edellytyksenä.

Toinen kognitivismin ongelmakohta liittyy tieto- tai tekijän taitoon. Ajatellaanpa vaikka pianon soittamista. Ensikertalaisella lienee syytä olla joku käsitys, mitä hän on tekemässä ja mitä hänen pitäisi tehdä, jotta toimen aloittaminen ja siinä kehittyminen olisi mahdollista, mutta taitava suoritus ei perustu niinkään propositionaaliseen tietoon vaan pitkälti taidokkaaseen motoriseen kykyyn. Tällaisissa taidoissa kyse tuskin on heuristiikoista, eikä soittamista varmaankaan voi edes kuvata kovin mielekkäästi symbolien manipulointina. Motoriset kyvyt eivät kuitenkaan ole kognitivistille mikään ongelma, koska teoria ei väitä, että inhimillinen toiminta olisi kaikkinensa kognitiivista. Esimerkiksi syöminen ei selvästikään ole symbolien käsittelyä. Autolla ajaminen tuntuisi olevan lähempänä kognitiivista toimintaa kuin soittaminen, koska kuskin täytyy tietää jotain auton ja sen ohjauslaitteiden toiminnasta, tuntea liikennesäännöt ja ymmärtää milloin hän rikkoo niitä, tunnistaa tien tapahtumia ja niin edelleen. Kaikki tämä kuitenkin katoaa harjoituksen myötä, ja jäljelle jää automaattinen mutta sujuva motorinen suoritus. Itse asiassa lähes kaikessa motorisessa toiminnassa suorituskyky ilmeisesti on jotakuinkin käänteisessä suhteessa vaadittavaan kognitiiviseen toimintaan, ja huomion kiinnittäminen toimintaan ei niinkään paranna vaan enemmänkin häiritsee suoritusta (Wulf, 2007, s.3–6).

Kaikki taidot eivät tietenkään ole samanlaisia motorisilta ja kognitiivisilta vaatimuksiltaan. Ääripäitä edustavat ehkä pianon soitto ja shakin pelaaminen, joissa ensimmäinen on pitkälti motorinen siinä missä jälkimmäinen perinteisessä mielessä lähes yksinomaan kognitiivinen kyky. Taitojen luokittelu motorisiin ja kognitiivisiin ei kuitenkaan ole aivan suoraviivaista, koska molemmat komponentit ovat läsnä monissa toimissa kenkien sito-

misesta autolla ajamiseen, joten on vaikeaa vetää rajaa milloin jokin taito on oleellisesti kognitiivinen ja milloin ei. Toiseksi tällainen kahtiajako on puutteellinen. Mukaan tulisi ottaa ainakin hahmontunnistuskyvyt. Esimerkiksi taidokas kokkaus edellyttää ruoan kypsyyden tunnistamista värin, rakenteen, tuoksun tai jonkin muun piirteen perusteella. Myös hahmontunnistuskyvyt ovat kehittyvää sorttia, ja aloittelija joutuu hahmontunnistuksen sijaan tyytymään ohjeiden ja nyrkkisääntöjen soveltamiseen.

Dreyfusin veljekset ovat analysoineet kognition ja taitojen suhdetta, ja esittäneet, että käänteinen suhde, joka havaitaan sujuvan motorisen suorituksen ja vaadittavan kognitiivisen toiminnan välillä, on yleistettävissä myös muihinkin kuin motorisiin taitoihin. Heidän mukaansa tie aloittelijasta ekspertiksi kulkee kyvystä riippumatta suunnilleen samalla tavalla. Esimerkiksi shakissa aloittelijan tulee aluksi tietää, mitä mikäkin nappula tekee sekä mitkä pääpiirteissään ovat hyviä ja mitkä huonoja siirtoja. Hän on kyvytön hahmottamaan laudan, ja vielä vähemmän pelin, kokonaistilannetta ja näkemään mikä kokonaisuudessa on oleellista ja mikä ei. Hänen suoritustaan on parasta arvioida tarkastelemalla, miten hyvin hän kykenee noudattamaan erinäisiä nyrkkisääntöjä, ei esimerkiksi miten joustavasti hän soveltaa ja rikkoo niitä eritystilanteissa, joita hän tämän mallin mukaan ei vielä kykene edes hahmottamaan. Suoritus paranee, kun sääntöjä ei enää sovelleta kontekstittomasti, vaan aloittelija on saavuttanut vaiheen, jossa hän pystyy hahmottamaan erilaisia kokonaistilanteita ja soveltamaan sääntöjä mielekkäästi tilanteen edellyttämällä tavalla. Lopulta tilanteita opitaan tarkastelemaan eri perspektiiveistä, tietoinen sääntöjen soveltaminen vähenee ja lopulta valinnat tehdään enemmänkin intuitiivisesti kuin harkiten. Samalla tunne valintojen merkityksestä, vastuusta ja sitoutumisesta kasvaa, huonojen tai epäoleellisten valintojen huomiointi katoaa, kuten myös toimijan tietoisuus omien valintojensa perusteista. Dreyfusit eivät viittaa termillä ”intuitio” mihinkään mystiseen vaan normaaliin mentaaliseen kykyyn kaiken arkisen toimintamme taustalla, joka ei edellytä tietoista ajattelua ja jota kognitiivinen teoria heidän mukaansa ei ole onnistunut valaisemaan. Kohtuullisen hyvin tämän asiantuntemuksen mahdollistaman intuition idean voi tiivistää toteamukseen, että kun asiat etenevät normaalisti, asiantuntijat eivät ratko ongelmia vaan tekevät sitä, mikä yleensä toimii. (Dreyfus & Dreyfus, 1986, s.21–35)

Heuristiset ongelmanratkontasysteemit ovat suunniteltu toimimaan juuri siten, kuin aloittelija yllä kuvatun mallin mukaan toimii, eli ongelmat pyritään jakamaan erillisiin manipuloitaviin komponentteihin, joita käsitellään hyvin määriteltyjen nyrkkisääntöjen avulla. Tätä ogelmanratkonta, ja sitä myötä oleellisesti kaikki kognitiivinen toiminta, Newellin ja Simonin mukaan on. Dreyfusin veljeksen väittävät, että tekoälysysteemit – esimerkiksi shakkiohjelmat – parhaimmillaan vaikuttavat toimivan aloittelijaa paremmin, mutta tämä johtuu yksinkertaisesti siitä, että koneiden laskentakapasiteetti mahdollistaa heuristisen algoritmien nopean ja tehokkaan käytön. Kuitenkin tällaiset ohjelmat ovat laadullisesti vangittu aloittelijan asteelle. (*ibid.*, s.63–65) Yleisluontoisesti älykkään käyttäytymisen etiologian kannalta yllä olevat huomiot ovat merkittäviä, olettaen, että Dreyfusit ovat suurin piirtein oikeassa. On ehkä ylipäätään luonnotonta mieltää normaalia inhimillistä toimintaa ongelmien ja ratkaisujen sarjana, mutta jos näille sanoille annetaan tietynlaiset tekniset merkitykset, voitaneen erilaisia arkisia tilanteita työhaastattelusta ystävän kanssa kahvilla käymiseen mieltää eräänlaisina ongelmatilanteina, ainakin hieman mielikuvitusta käyttäen. Joka tapauksessa vaikka tämä käsitteellistys olisi mielekäs ja edellä

GPS -ohjelman yhteydessä huomioidut yleisyyteen liittyvät tekniset ongelmat saataisiin ratkaistua, Dreyfusien analyysin mukaan heuristiset symbolienkäsittelijät eivät siltikään kykene inhimillisessä katsannossa pätevään, yleisesti älykkääseen ja mielekkääseen käyttäytymiseen. Vaikkakin oikein ohjelmoidut koneet voivat laskentavoimansa avulla toimia pätevän oloisesti tiettyyn rajaan asti, ne Dreyfusien mukaan kuitenkin toimivat kankeammin ja laadullisesti eri tavalla kuin ihmiset.

Nämäkään huomiot eivät mitenkään ilmeisesti kognitivimia kaada. Taidokkaaseen toimintaan liittyvät motoriset tai muuten epäkognitiiviset komponentit eivät varsinaisesti kuulu teoria tutkimuskohteeseen, ja mitä Dreyfusien analyysiin tulee, on tietenkin mahdollista, että intuition kehittyminen ja nyrkkisääntöjen tietoisen noudattamisen katoaminen tarkoittaa vain, että ihminen siirtyy käyttämään tekokkaampia heuristisia menetelmiä alitajuisesti. Kuitenkin nämä tarkastelut saattavat avata portit hyvinkin radikaalille antikognitivismille. On melko ilmeistä, että ihmiset joskus käsittelevät symboleita ja symbolirakenteita. Erityisesti väite, että meitä voi pitää symbolien käsittelijöinä ainakin joskus kun ratkomme ongelmia, lienee oikean suuntainen. Kysymys kuitenkin kuuluu, miten oleellinen osa mielekkäästä, taidokkaasta ja älykkäästä käyttäytymistä todella perustuu symbolirakenteiden käsittelylle tai edellyttää propositionaalisia representaatioita ylipäätään. Esimerkiksi robotiikkaa kognitiotieteen ytimeen työntänyt Rodney Brooks on yksinkertaisten autonomisesti toimivien robottien parissa tekemänsä työn perusteella vetänyt sellaiset johtopäätökset, että yksinkertaista älykkyyttä vaativassa toiminnassa mallien ja representaatioiden käytöstä on vain haittaa, ja on parempi antaa maailman suoraan toimia omana mallinaan. Tällä hän tarkoittaa, että tehokas toiminnanohjaus perustuu havainnon ja toiminnan muodostamaan kehään eikä niinkään kontekstittomien propositionaalisten representaatioiden luomiseen ja käsittelyyn. Edelleen hän on esittänyt, että representaatio on ylipäätään väärä tapa käsitteellistää merkittävää osaa älykkäiden systeemien toiminnasta (Brooks, 1991, s.80–81). Huomautettakoon myös, että sensorimotoriset taidot, hahmontunnistus ja motorinen kontrolli vaikuttavat vaativat merkittävästi suurempia laskentaresursseja, kuin järkeily (Moraveck, 1988, s.14–16). Lisäksi ainakin tietoinen ja tahdonalainen symbolienkäsittely vaikuttaa olevan ihmisille kohtuullisen hankalaa ja tehotonta. Suurin osa tehokkaasta kognitiosta on alitajuista ja automaattista, ja tämä alue on huonosti tunnettua ja hankalasti mallinnettavissa (Minsky, 1985, s.29). Näin ollen on hyvinkin mahdollista, että merkittävä osa mielekkäästä käyttäytymisestä perustuu jollekin aivan muulle kuin symbolirakenteiden käsittelylle.

Eräs selitys pätevän toiminnan kehittymiselle saattaisi hyvin olla, että alussa toimintaa harjoitellaan kankeiden, mutta pääpiirteissään toimivien, nyrkkisääntöjen avulla, kunnes hahmontunnistusjärjestelmät oppivat erottamaan tilanteista oleelliset piirteet ja liittämään niihin hyväksi havaitun päätöksen tai toiminnan. Tällöin mentaaliset representaatiot ovat kyllä työssään, mutta ne mahdollisesti ovat tilanne-, tarve- ja perspektiiviriippuvaisia, mikä melko pitkälle on oleellisuuden määritelmä, eivätkä luonteeltaan kontekstittomia propositionaalisia esityksiä. Tämä olisi eräs selitysehdotus intuitiolle. Mikäli tämänkaltainen käsitys mielen toiminnasta on oikeansuuntainen, niin heuristiikat ja kontekstittomat propositionaaliset representaatiot kyllä edistävät yleisluontoista älykästä käyttäytymistä saattamalla alkuun uudentyyppisten kykyjen kehityksen, mutta sujuvan ja mielekkään käyttäytymisen kehitys on siirtymistä poispäin yleisyydestä ja kohti eri-

koistunutta havainto–toiminta-sykliä. Pätevällä toimijalla on tarvittaessa kuitenkin mahdollisuus ottaa ongelmiin abstrakti, käsitteellinen ja harkitseva näkökulma, esimerkiksi kun intuitiivinen kyky osoittautuu eriytymisensä takia puutteelliseksi. Laadullisesti tämä siirto ei kuitenkaan ole asiantuntemuksen tuoma kyky ajatella asioita abstraktisti vaan tilapäistä ja tarkoituksenmukaista taantumusta aloittelijan asteelle. (Dreyfus & Dreyfus, 1986, s.36–40) Tässä mielessä asioiden katsominen aloittelijan silmin saattaa olla yllättävän osuva ja melko kirjaimellisesti pätevä metafora luovasta ongelmanratkaisusta.

Lisäksi päättelyn ja ongelmanratkaisun psykologia asettaa käsityksen mielestä ensisijaisesti deduktiosysteeminä melko arveluttavaan valoon. 1960-luvun loppupuolella Peter Wason raportoi mielenkiintoisia tuloksia kokeista, joissa koehenkilöitä pyydettiin ratkomaan jokseenkin abstrakteja mutta yksinkertaisia ongelmia. Erään tunnetun kokeen vakiomuotoilussa koehenkilöille kerrottaan, että jokaisessa kortissa, jossa on kirjain A, toisella puolella on numero 3, ja henkilöille näytettään kortit A,D,3 ja 7. Kun sitten kysytään, mitkä kortit ovat välttämättä käännettävä väitteen totuuden testaamiseksi, alle 10 % ymmärtää, että kortti 7, mutta ei 3, on esitetyn väitteen kannalta ratkaiseva. Suoritus kuitenkin paranee merkittävästi, jos loogisesti täysin sama ongelma esitetään vähemmän abstraktissa muodossa. Esimerkiksi jos väitetään, että eräissä nuorten kotibileissä kaikki juopuneet ovat täysi-ikäisiä, niin lähes kaikki ymmärtävät, että juopuneiden ikä ja alaikäisen juopumustila on syytä selvittää, mutta täysi-ikäiset ja selväpäiset juhlijat ovat väitteen kannalta epäoleellisia. (Evans et al. 1993, s.99–101; 2003, s.456)

Tämä epäsuhta on melko kummallinen, mikäli päättely perustuu ongelman formaalin representaation rakentamiseen mielessä ja sen pohjalta suoritettuun deduktioon. Toisaalta jos mielen normaali toimintatapa on kiinnittää huomio oleelliseen, ei korttikokeen paljastama ajattelun epäloogisuus vaikuta kovin erikoiselta. Mahdollisesti henkilöt pitävät kortteja A ja 3 oleellisina – ensimmäistä oikein ja jälkimmäistä virheellisesti – koska nämä kortit mainittiin tehtävän kuvauksessa eikä heillä ole edeltävää kokemusta tällaisten tehtävien ratkonnasta, joka ohjaisi huomiota oikeasti oleelliseen. Kotibile-esimerkissä taas ihmisillä on apunaan sosiaaliset normit, eli ettei alaikäisten pitäisi olla humalassa mutta täysi-ikäisten osalta tämä on heidän oma asiansa. Näin ollen kaikki tämän kulttuurisen normin jakavat yksilöt näkevät yleensä välittömästi, mikä väitteen kannalta on oleellista ja mikä ei kiinnittämättä huomiota ongelman yksinkertaiseen loogiseen rakenteeseen. (Evans et al., 1993, s.112–114.) Logiikkaa harrastaneiden luulisi puolestaan pärjäävän korttitehtävässä paremmin, koska heidän pitäisi suoraan tunnistaa se yksinkertaisena implikaatiolauseen testaamistehtävänä, mutta ikävä kyllä edes logiikan opiskelu ei välttämättä paranna suoriutumista merkittävästi (Evans et al., 1993, s.107–109). Mikäli ihmisten looginen ajattelu on näin heikkoa, ettei edes logiikkaa opiskelleet kykene soveltamaan *modus tollens* päättelysääntöä systemaattisesti oikein, nakertaa tämä uskottavuutta erääseen komputationalismin kulmakiveen, eli ajattelun deduktiiviseen systemaattisuuteen.

Kuten robotikko Brooks, jotkut kognition filosofit ovat päätyneet hylkäämään representationalismin kokonaan ja ryhtyneet etsimään mielen metaforia tietokoneiden sijasta muun muassa dynaamisten systeemien teoriasta, säätöjärjestelmistä ja muista vastaavista systeemeistä.[64] Tällainen radikaali antikognitivismi on kuitenkin helppo viedä turhan pit-

[64]Ks. esim. (van Gelder, 1995) ja (Bechtel, 2001), jossa käsitellään representaation käsitettä dynaamisissa ja neuraalisissa systeemeissä.

källe. Esimerkiksi kontrafaktuaalinen päättely varmaankin edellyttää jonkinlaisia representaatioita, ja kuten edellä on todettu, ilmeisesti ihmiset ainakin ajoittain käsittelevät symbolirakenteita esimerkiksi päättelyjä tehdessään ja luonnollista kieltä käyttäessään. Ei ole mitään pakottavaa syytä pitäytyä ajattelun kielen teoriassa tai hylätä sitä kokonaan, koska on aivan mahdollista, että kognitio hyödyntää samanaikaisesti useanlaisia representaatioita.

Viime aikoina onkin löytynyt jonkinlaista empiiristä näyttöä, joka viittaisi siihen, että järkeily pohjautuu kahdenlaisille mekanismeille, joista toinen koostuu erikoistuneista, automaattisista hahmontunnistusjärjestelmistä ja toinen puolestaan muistuttaa klassisen komputationaalisen teorian mukaista yleistä deduktiomekanismia (Evans, 2003). Eräs tämän näkemyksen mukainen hypoteesi kognition arkkitehtuurista perustuu virtuaalikonehierarkialle, missä assosiatiiviset rinnakkaisprosessointiin perustuvat hahmontunnistusjärjestelmät muodostavat kognition perusrakenteen ja toisaalta myös implementoivat seriaalista symbolienkäsittelyä suorittavan virtuaalikoneen. Tällaisten hahmontunnistusjärjestelmien yleensä oletetaan olevan jonkinlaisia neuroverkkoja.

Menemättä tässä sen kummemmin neuroverkkojen formalismiin, kyseessä on eräs komputationaalisten systeemien luokka, jotka eivät käsittele symbolirakenteita, vaan muuntelevat numeerisia vektoreita matriisilaskennan avulla. Tyypillisesti neuroverkkojen tulkitaan luokittelevan syötteitä eri joukkoihin, missä syötevektori on tunnistettavan hahmon vektorikomponenteiksi koodattu representaatio ja tulostevektorin komponentit puolestaan kertovat mihin luokkiin verkko syötteen sijoittaa. Siinä missä klassiset komputationaaliset systeemit manipuloivat muistissa olevia symbolirakenteita, neuroverkoissa syötettä ja sen käsittelyä on yksinkertaisempi ajatella signaalina, joka virtaa hajautettuna verkon läpi matkalla muuntuen.[65] Neuroverkot ovat yleisesti ottaen varsin tehokkaita hahmontunnistusjärjestelmiä, mutta koska ne eivät käytä kompositionaalisesti rakennettuja symbolisia representaatioita, eivät ne kovin luontevasti sovellu symbolirakenteiden rekursiiviseen käsittelyyn (Bechtel & Abrahamsen, 2002, s.116). Neuroverkot kuitenkin kuuluvat Turing-täydellisten systeemien luokkaan (Siegelmann & Sontag, 1991), joten tämä rajoitus on luonteeltaan käytännöllinen, ei periaatteellinen. On siis ainakin mahdollista, että kognition ensimmäisen kertaluvun virtuaalikone koostuu neuroverkoista, jotka puolestaan implementoivat jonkinlaisen symbolienkäsittelysysteemin.

Paul Smolensky lienee tunnetuimpia tällaisen kognitionteorian kehittelijöitä. Hänen mukaansa kognitiivisen systeemin neuroverkot eivät varsinaisesti implementoi symbolisysteemeitä, vaan simuloivat tai approksimoivat sellaisia, jolloin suorituskyky voi poiketa huomattavastikin simuloidusta virtuaalikoneesta.[66] Muun muassa muistirajoitusten takia teknisesti ottaen kaikki fysikaaliset symbolisysteemit approksimoivat varsinaisia deduk-

[65]Neuroverkkojen perusmekaniikka on varsin yksinkertaista, mutta formalismin kognitiotieteellisten ulottuvuuksien esittely menee äkkiä niin monimutkaiseksi, ettei siihen tässä yhteydessä ole mahdollisuuksia. (McClelland et al., 1986a) on klassinen esitys neuroverkkomallinnuksen perusteista, joskin (Bechtel & Abrahamsen, 2002) on ehkä hieman ajankohtaisempi perusteos. Syy nostaa neuroverkot esiin tässä kohtaa on vain tarve huomauttaa, että kappaleessa käsiteltävät klassiselle komputationalismille vaihtoehtoiset mallit eivät ole hypoteettisia vaan aktiivisen tutkimuksen ja kehittelyn kohteita.

[66]Ks. (Smolensky, 1988) ja esim. (Clark, 1989, 127–141,150–152), (Rowlands, 1994). Ks. myös (Bechtel & Abrahamsen, 2002, s.103–109), jossa esitetään, että looginen päättely ja sen oppiminen voisi perustua suoraan hahmontunnistukselle ilman tarvetta edes simuloida symbolista virtuaalikonetta.

tiosysteemeitä. Kuitenkin tietokoneohjelmat todella suorittavat simuloimansa formalismin kalkyylia, jolloin on mielekkäämpää puhua implementaatiosta. Jos taas systeemin käyttämät representaatiot ja komputationaaliset operaatiot sekä suurpiirteinen käyttäytyminen poikkeaa merkittävästi simuloidusta systeemistä, ei viritystä varsinaisesti voi pitää implementaationa. Smolenskyn hypoteesin mukaan tällainen siis on kognitiivisen systeemin perusarkkitehtuurin ja sen suorittaman symbolisen prosessoinnin suhde. Teoria pyrkii selittämään, miten monipuolinen ja mielekäs käyttäytyminen on mahdollista, vaikka looginen tai muuten formaali ajattelumme on melko heikkoa, ja toisaalta mihin ilmeinen kykymme kuitenkin jonkinlaiseen propositionaaliseen järjenkäyttöön perustuu.

Toisaalta taas ihmiset pystyvät kohtuullisen monimutkaiseen ja abstraktiin symboliseen ajatteluun. Muun muassa matemaatikot tienaavat leipänsä tällaisella toiminnalla. Muodostaako tämä jonkinlainen antiteesin edellisille tarkasteluille? Symbolirakenteiden manipuloinnin ei kuitenkaan tarvitse tapahtua päässä vaan esimerkiksi paperilla. Yleisesti ottaen deduktioita ja laskutoimituksia on huomattavasti helpompaa tehdä kynän ja paperin kanssa kuin ilman. Tehokkaan symbolisen ajattelun ei tarvitse olla mentaalisten vaan konkreettisten, esimerkiksi kirjoitettujen, symbolirakenteiden manipulointia. Mahdollisesti tämä kyky voidaan osittain sisäistää, jolloin ulkoisia apuvälineitä ei enää tarvita kaikkein yksinkertaisimpiin toimituksiin.[67] Allekkain laskeminen lienee tästä havainnollisimpia esimerkkejä, mutta mahdollisesti näin käy myös luonnollisen kielen kanssa.

Voi olla, että kielellinen kyky syntyy aluksi ulkoisena toimintana, eli kommunikaationa, olipa kyseessä sitten puhuminen, viittominen tai mikä hyvänsä toiminta, ja vasta myöhemmin kielen käyttö opitaan sisäistämään. Tällöin kieli ei olisi alkujaan ajatusten ilmaisua, vaan kommunikaation väline, jota kehittyvä lapsi oppii käyttämään kognitiivisena työkaluna. Tällainen käsitys on yhteensopiva esimerkiksi Lev Vygotskin kielen ja ajattelun kehitystä koskevien teorioiden kanssa (Vygotski, 1982, s.91–94). Kirjoitetun kielen ja muiden ulkoisten representaatiojärjestelmien hallinta laajentaa kognitiivisia kykyjä merkittävästi, koska esimerkiksi työmuistin rajoitukset voidaan näin ohittaa. Jos ajattelun käsitettä ei rajoiteta koskemaan ainoastaan mielen sisäisiä tapahtumia, vaan käsitteen alaan katsotaan kuuluvan ylipäätään representaatioiden käsittelyprosessit joissa kognitiivinen systeemi on tavalla tai toisella osallisena, niin esimerkiksi kirjoittamista itse asiassa voidaan kirjaimellisesti pitää eräänlaista ajattelua.[68] Jos tämänlainen käsitys kognitiosta ja symbolisesta ajattelusta on oikeansuuntainen, mihin itse ainakin varovaisesti uskon, on kognitivistinen mielenteoria hyvin ongelmallinen. Tällöin nimittäin ihmisen kognitiivisiin kykyihin kyllä sisältyy symbolirakenteiden manipuloiminen, mutta tähän liittyvät prosessit eivät pääpiirteissään tapahdu mielen sisällä, vaikkakin tällainen käyttäytyminen oletettavasti edellyttää jonkinlaista kognitiivista systeemiä. Mahdollisesti kognitivismin syvällisin virhe onkin symbolisen ajattelun ja kognitiivisten prosessien samaistaminen ja koko mielenteorian perustaminen tälle erheelle.

Kolmas ja viimeinen ongelma, johon tartun, koskee ymmärtämistä ja tervettä järkeä. Näistä kolmesta kognitivismin ongelmasta tämä on kaikkein epämääräisin ja ehkä syvälli-

[67] Ks. esim. (Clark, 1989, s.127–139) ja (Rumelhart et al., 1986, s.38–52).

[68] Andy Clark on kirjoittanut viimeaikoina kokonaisen kirjan, jonka kantava teema on, että merkittävä osa niin sanotusta korkeasta kognitiosta ei itse asiassa tapahdu pään sisässä, vaan ulkoisia representaatioita manipuloimalla. Ks. (Clark, 2008)

sin. Jossain määrin tämä muistuttaa taidokkaan käyttäytymisen ongelmaa mahdollisesti ollen sen yleistys. Tietyssä mielessä klassisten tekoälyohjelmien voi sanoa ilmentävän jonkinlaista ymmärtämistä, jos tällä tarkoitetaan yksinkertaisesti asiayhteydestä riippuvaa mielekästä käyttäytymistä. Havainnolliesia esimerkkejä tästä edustavat tekoälyn kultavuosien tunnetuimmat ohjelmat: Terry Winogradin 60-luvun lopussa laatima SHRDLU[69] (Winograd, 1971) ja Robert Schankin sekä Robert Abelsonin 70-luvun loppupuolen *Script Applier Mechanism* (SAM) (Schank & Abelson, 1977).

SHRDLU toimii simuloidussa "laatikkomaailmassa", joka koostuu pienestä tasosta, jonka päällä on erivärisiä ja muotoisia pyramideja, laatikoita ja kuusitahokkaita. Ohjelma pystyy siirtelemään ja pinoamaan näitä kappaleita. Sen käyttöliittymä on toteutettu englannin kielellä, joten käyttäjä pystyy muun muassa antamaan systeemille käskyjä, tietoja ja määritelmiä sekä esittämään sille kysymyksiä luonnollisella kielellä. Ohjelman kanssa käyty kommunikaatio rajoittuu laatikkomaailmaa koskeviin asioihin, mutta se kykenee muun muassa ratkomaan kielen käyttöön liittyviä epämääräisyyksiä, vastailemaan yllättävän mielekkäästi monelaisiin kysymyksiin ja tulkitsemaan melko joustavasti sille annettuja käskyjä (Winograd, 1971, s.35–60).

SAM taas perustui uudenlaisille tietorakenteille, jotka muistuttavat hieman mallien ja teorioiden välimuotoa. Oleellisesti nämä rakenteet ovat stereotyyppisiä kuvauksia erilaisista arkisista tilanteista, kuten syntymäpäiväjuhlista, ravintolassa käymisestä ja muista sellaisista. Ohjelman tarkoituksena on kyetä tulkitsemaan lyhyitä tarinoita tai tilanteen kuvauksia, ja vastailemaan niitä koskeviin kysymyksiin. SAM:in tietorakenteet sisältävät vakio-oletuksia siitä, mistä tapahtumista tällaiset stereotyyppiset tilanteet koostuvat sekä miten eri tapahtumat, toimijat ja muut seikat eri asiayhteyksissä liittyvät ja vaikuttavat toisiinsa (Schank & Abelson, 1977, s.11–17). Esimerkiksi ohjelman representaatio ravintolassa käymisestä koostuu sisääntulosta, paikan etsimisestä, tilaamisesta, joka sisältää muun muassa tarjoilijan kanssa asioimista, syömisestä, maksamisesta ja poistumisesta sekä yksityiskohtaisempaa tietoa siitä, mitä kaikkea näihin tapahtumiin puolestaan liittyy (*ibid.*, s.42–46). Ohjelmalle kerrotut tarinat voivat muuttaa tietorakenteiden vakio-oletuksia tapahtumien kulusta, ja sen pitäisi kyetä päättelemään, mitä tällöin tapahtuu, ilman, että kaikkea kerrotaan sille suoraan. Ohjelma kykenikin kohtuulliseen suoritukseen. SAM:ille esitettiin muun muassa tarina "John meni ravintolaan. Hän tilasi hampurilaisen. Tarjoilija kertoi, ettei heillä ollut sellaisia. John pyysi nakkisämpylän. Kun nakkisämpylä tuli, se oli palanut. Hän poistui ravintolasta." Kun ohjelmalta sitten kysyttiin, söikö John nakkisämpylänsä, SAM vastasi kielteisesti, koska se tiesi palaneen ruoan olevan syömäkelvotonta, ja päätteli, ettei syöminen tässä tilanteessa edellä ravintolasta poistumista, vastoin kuin normaalisti. (*ibid.*, s.190–204). Tällaisten ohjelmien laatiminen ei ole aivan yksinkertaista, koska jo pelkästään ilmausten "hän" ja "se" referentin selvittäminen vaatii jonkinlaista tulkintaa. Lisäksi ohjelman täytyy sisältää jokseenkin paljon kaikenlaista yksityiskohtaista tietoa sekä pystyä päättelemään, mikä on kysytyn asian kannalta oleellista ja mitä kaikenlaisia seurauksia odottamattomilla tapahtumilla voi olla.

Ymmärryksen ongelma tässä yhteydessä siis tarkoittaa oleellisen poimimista, mielekästä toimintaa epämääräisen ja epätäydellisen tiedon varassa, sekä terveen järjen hallintaa.

[69] Ohjelman nimi ei ole akronyymi, vaan eräänlaista huumoria. Ks. (Boden, 2006, s.685).

Ihmisillä on käytössään suunnattomasti kaikenlaista tietoa, jota sovelletaan tarpeen mukaan automaattisesti tai oletusarvoisesti tai sitten vaihtoehtoisesti jätetään tarpeettomana soveltamatta. Esimerkiksi kaikki olettavat, että luentosaleissa on muun muassa lattia, seinät ja katto. Tästä syystä kukaan normaalijärjellä varustettu ihminen ei mainitse näitä asioita kysyttäessä, mitä luentosalissa tänään on, vaan tämän kysymyksen tulkitaan automaattisesti koskevan jotain muuta. Oleellisen poiminen tarkoittaa siis myös epäoleellisen huomiottajättämistä. Kuitenkin esimerkiksi Turing-testissä pärjäävän ohjelman pitäisi sisältää tällaista epäoleellista tietoa suunnattomasti, jotta se osaisi vastata oikein epätyypillisiin kysymyksiin, esimerkiksi syökö ravintolan asiakas suullaan vai korvallaan. Kun asiat menevät pieleen tai etenevät muuten epänormaalilla tavalla, tällainen itsestään selvä epäoleellinen tieto voikin osoittautua toiminnan kannalta ratkaisevaksi. Toiseksi tiedon oleellisuus riippuu hyvin paljon tilanteesta. Esimerkiksi jos luentosali on remontissa, niin se, että salissa juuri tänään on lattia ja katto, saattaa olla hyvinkin oleellista tietoa. Huomautettakoon, että proseduraalisen semantiikan mukaan normaalit käsitteet edellyttävät kykyä päätellä tällaista sirpaleistakin tietoa. Koneella tuskin voi katsoa olevan esimerkiksi asiallista syömisen käsitettä, ellei se tiedä millä ruumiinosalla tämä toimitus tapahtuu. Joka tapauksessa terve järki siis edellyttää sunnattomasti kaikenlaista triviaaliakin tietoa, mutta myös kykyä sen asianmukaiseen käyttöön, mikä puolestaan edellyttää ymmärrystä siitä, mikä milloinkin on relevanttia. Ymmärrys ja terve järki siis muodostavat hyvin samankaltaisen ongelman kuin edellä taidokkaan toiminnan yhteydessä mainittu oleellisen poimiminen epäoleellisesta. Kyse on kuitenkin hieman yleisemmästä ilmiöstä. Aiemmin relevanssin ongelma esiintyi ongelmanratkonnan yhteydessä, mutta terve järki edellyttää myös sen ymmärtämistä, mikä käsillä oleva tilanne tai ongelma oikeastaan on, ja kykyä irtaantua tietystä tulkinnasta tilanteen muuttuessa.

Hubert Dreyfus on puuttunut tähänkin ongelmaan väittäessään, että heuristiset ohjelmat ovat kyvyttömiä erottamaan oleellista epäoleellisesta, ja tässä mielessä ne eivät varsinaisesti ymmärrä mitä ne ovat tekemässä. Hänen mukaansa heuristiset algoritmit toimivat laadullisesti eri tavalla kuin ihmismielet, ja tästä syystä klassiseen komputationalismiin perustuvat tekoälysysteemit eivät voi saavuttaa inhimillistä suorituskykyä useissakaan tehtävissä (Dreyfus, 1965, s.83–85). Hän kritisoi muun muassa Newellin ja Simonin väitettä, että GPS:n toiminta vastaa ihmisten ongelmanratkontaa. Dreyfus huomautti, että kun ohjelman ja ihmisten raportteja vertaa hieman tarkemmin toisiinsa, huomataan, että ajoittain ne itse asiassa poikkeavat oleellisesti. GPS suorittaa melko sokeita yritys–erehdys-hakuja ratkaisupuussa, mutta Newellin ja Simonin koehenkilöt ajoittain selvästi karsivat epäoleellisia ratkaisuyrityksiä. Erityisesti Dreyfus kritisoi heidän väitettään, että myös koehenkilöt tosiasiassa tekivät päättelyssään näitä turhia askelia, mutta eivät vain raportoineet tai edes tiedostaneet niitä. Hän huomauttaa, että varsinaisen empiirisen aineiston perusteella koehenkilöt käyttivät ajoittain laadullisesti hyvin erilaista tietojenkäsittelystrategiaa kuin GPS (*ibid.*, s.24–30). Jos heuristiset algoritmit todella ovat ihmisiin verrattuna tällä tavoin rajallisia, herää kysymys, miten ne kykenisivät avoimessa tilanteessa arvioimaan mikä on oleellista ja mikä ei, jos ne eivät kykene tähän edes kun päämäärät ja käytössä olevat keinot ovat täsmällisesti määritelty.

SHDRLU ja SAM puolestaan ovat paraatiesimerkkejä vanhan koulukunnan tekoälytutkimuksesta, jotka saattavat äkkiseltään vaikuttaa Dreyfusin analyysin antiteeseiltä. Mutta

erityisesti SAM:in näennäisesti mielekäs toiminta perustuu siihen, että sille esitettyjen tarinoiden ja kysymysten konteksti on ennalta määrätty annetussa tietorakenteessa, jota se käyttää tarinoiden tulkitsemiseen. Erityisesti tietorakenteissa on määritelty kaikki oleelliset seikat sekä millä tavalla asiat voivat poiketa normaalista järjestyksestä. Dreyfus onkin huomauttanut, että SAM on ohjelmoitu etäisesti imitoimaan inhimillistä ymmärrystä, mutta tarkemmin katsottuna ohjelma ei sisäistä esimerkiksi ravintolareissuihin liittyviä piirteitä alkuunkaan normaalilla tavalla. Ohjelma ei osaa vastata esimerkiksi kysyttäessä, käveleekö tarjoilija etu- vai takaperin tai syökö asiakas suullaan vai korvallaan. Se ei myöskään itse osaa erotella oleellisesta epäoleellisesta, vaan ohjelmoija määrittelee, mikä on relevanttia sisällyttää tietorakenteisiin, ja ohjelmoijan epäoleellisena pitämistä asioista SAM ei siis tiedä yhtään mitään (Dreyfus, 1992, s.40–43). SHRDLU puolestaan ei ole edes suunniteltu sisäistämään minkäänlaista tervettä järkeä tai ymmärrystä, vaan kyseessä on oikeastaan demonstraatio luonnolliseen kieleen perustuvasta käyttöliittymästä. Ohjelman näennäisesti mielekäs käyttäytyminen on seurausta juurikin hyvin suunnitellusta käyttöliittymästä sekä siitä, että ohjelman palikkamaailma-toimintaympäristö on niin yksinkertainen, että ei pelkästään oleelliset asiat, vaan ylipäätään tuon maailman kaikki tosiasiat on hyvin helppo luetella ja tarvittaessa tarkistaa (Haugeland, 1985, s.188–195). Kontekstin rajaaminenkaan ei toisaalta vaikuta tuovan suunnatonta helpotusta terveen järjen ohjelmointiin. Esimerkiksi Turingin testi itse asiassa on järjestetty rajoitettuna versiona vuosittain vuodesta 1991 lähtien. Keskustelun aihe päätetään etukäteen, mikä rajaa kontekstin johonkin yksinkertaiseen aihepiiriin, kuten jalkapallo-otteluihin, josta kuulustelijoilla ei ole lupaa poiketa. Myöskään kysymyksiä, jotka on selvästi tarkoitettu testaamaan, onko keskustelukumppani kone vaiko ihminen, ei sallita. Tästä huolimatta ohjelmat eivät pärjää testissä lähimainkaan ihmisen veroisesti (Boden, 2006, s.1353–1354).

Ymmärryksen tai maalaisjärjen ongelma on luonnollisesti ollut keskeinen teoreettinen kysymys tekoälyssä jo tutkimuksen alkuajoista lähtien. Tähän liittyvät luonteeltaan filosofiset ongelmat on huomioinut myös ehkä kaikkein paatunein logiikkatekoälyn kannattaja ja merkittävä tietokoneteknologian pioneeri John McCarthy. Hänen tavoitteensa on jo 50-luvulta lähtien ollut koittaa ohjelmoida koneisiin tervettä järkeä predikaattilogiikan kaltaista kieltä ja deduktiosysteemiä käyttäen (McCarthy, 1969). Varsin äkkiä hän huomasi, että pelkkä tietomassa deduktiosääntöjen lisäksi ei saa ohjelmaa toimimaan mielekkäästi, koska melkein minkä tahansa toiminnan kannalta käytännössä lähes ääretön määrä tietoa voi osoittautua oleelliseksi. Logiikan kannalta tämä ei varsinaisesti ole ongelma, mutta se tekee hyvin vaikeaksi ohjelmoida koneeseen suurta määrä tietoa, ilman että sen päättelysysteemi rampautuu tarpeettoman tiedon tulvasta. Yhdessä Patrick Hayesin kanssa he nimesivät tämän *kehysongelmaksi*. Tämä seikka ei kuitenkaan McCarthya ja Hayesia lannistanut, vaan he panivat uskonsa uusiin loogisiin välineisiin, kuten modaaliseen ja epämonotoniseen logiikkaan (McCarthy & Hayes 1969, s.43–63; McCarthy 1986, s.198,216–217). Myöhemmin Hayes koitti replikoida terveen järjen lähtemällä perustoista. Hän pyrki formalisoimaan kansanfysiikan, eli ihmisten arkisen toiminnan taustalla olevan käsityksen kappaleiden käyttäytymisestä, voimista ja muista fysikaalisista ilmiöistä. Ajatus oli päästää tekoälysysteemit ulos laatikkomaailmoista järkeilemään inhimilliseen tapaan oikean maailman tapahtumista (Hayes, 1979). Toiset taas, kuten aiemmin tavatut

Schank ja Abelson, keskittyivät tietorakenteiden luomiseen. Tässä vahvasti mukana oli myös Marvin Minsky kehysteorioineen (Minsky, 1974).

Kehykset ovat hierarkkisia tietorakenteita, jotka määrittelevät erilaisia stereotyyppisiä tilanteita ja olioita, hieman SAM:in tietorakenteiden tapaan, ja niitä voi pitää jonkinlaisina käsitteiden tai psykologisten skeemojen malleina. Tietorakenteen ylin taso sisältää muuttumattomia määreitä, jotka määrittelevät kehyksen sovellusalan välttämättömät piirteet. Esimerkiksi kaupankäynnissä on välttämätöntä, että jokin hyödyke siirtyy A:lta B:lle ja vastavuoroisesti jonkinlainen korvaus B:ltä A:lle, joten tällainen tieto tulisi olla kaupankäyntikehyksessä korkealla tasolla. Hierarkian alimmalla tasolla taas on vakioarvoisia muuttujia, joiden arvot voivat vaihdella tilanteen mukaan. Esimerkiksi vakioinen korvaus kaupankäyntikehyksessä voisi olla raha, tarkemmin sanoen eurot esimerkiksi suomalaisessa kehyksessä, mutta myyjä voi tyytyä muunkinlaiseen korvaukseen ilman, että tapahtuma lakkaa olemasta kaupankäyntiä. Kehysongelman kannalta keskeistä tässä on, että muuttujat eivät voi saada aivan mitä tahansa arvoja. Jos tilanteen edellyttämät arvot eivät sovi kehykseen, jota ohjelma käyttää tilanteen tulkintaan, ohjaa tämä systeemin valitsemaan jonkun toisen paremmin soveltuvan kehyksen. Kehykset voivat myös viitata toisiinsa ja näin muodostaa vielä korkeamman asteen tietorakenteita. (Minsky, 1974, s.95–97) Kehysteoria muistuttaa kovasti siis jonkinlaista olio-ohjelmoinnin esiastetta. Kehysratkaisu poikkeaa puhdasverisestä logiikka- ja heuristiikkaperustaisesta tekoälystä siten, että ydinkysymys ei ole tiedon käsittely vaan organisointi. Ratkaiseeko kehysteoria kehysongelman, riippuu osittain siitä, kykeneekö systeemi valitsemaan kehyksiä mielekkäästi eri tilanteiden edellyttämällä tavalla. Eräs heti silmiinpistävä ongelma tässä on, että kehyksen on tarkoitus nimenomaan määrittää, mikä tilanteessa on oleellista ja mikä ei, mutta toisaalta kehyksen valinta pitkälti edellyttää tilanteen tulkintaa, eli oleellisen erottamista epäoleellisesta. Näin ollen kehällisyyden vaikutelmalta ei voi välttyä.

Edellä käsitellyt ongelmat eivät sinänsä osoita klassista tekoälyä mahdottomaksi, vaan niihin voi suhtautua myös haasteina. Projekti vielä elää, ja ellei voi hyvin, niin ainakin kituuttaa eteenpäin. Esimerkiksi vielä vuonna 2005 McCarthy julisti matemaattisen logiikan tarjoavan parhaan perustan inhimilliseen suorituskykyyn tähtäävälle tekoälylle (McCarthy, 2005). Kunnianhimoisin elossa oleva kehyspohjainen tekoälyprojekti taas lienee Douglas Lenatin luotsaama Cyc. Perustuen eri arvioihin, Lenat ja Edward Feigenbaum esittivät vuonna 1987, että vajaat miljoona kehystä riittää kuvaamaan aikuisen ihmisen tietomäärän, mikä puolestaan riittää antamaan koneelle valmiudet inhimilliseen järjenkäyttöön (Lenat & Feigenbaum, 1987, s.1178,1180). Heidän arvionsa mukaan ohjelman toteuttaminen vaatisi noin 50 miljoonaa dollaria ja vuosikymmenen. Rahoitus on ollut arvion mukaista, mutta tulokset eivät, ja veikkaukset tarvittavien dollarien ja kehysten määrästä ovat aikain saatossa muuttaneet kertaluokkaansa (Boden, 2006, s.1008–1110). Tekoälyn varhaisten vuosien optimismia näyttää siis vielä jostain löytyvän, joskin vuosikymmenten aikana on jotain opittukin. Siirrytään seuraavaksi vetämään yhteen mitä kognitivismin ongelmat paljastavat teorian taustaoletuksista ja miltä funktionalismin ja komputationalismin tulevaisuudennäkymät nyt vaikuttavat.

4.3 Kognitivismin tila ja tulevaisuus

Kognitiivinen mielenteoria voidaan tiivistää muutamaan ydinkohtaan: 1°*a*) Teorian ensisijainen tutkimuskysymys on tarkoituksenmukaisen käyttäytymisen etiologian selittäminen, joka *b*) edellyttää teoriaa, jonka ontologia perustuu propositionaalisille asenteille, ja jonka yleistykset oletettavasti muistuttavat hyvin pitkälle arkipsykologisia vastaavia. 2°*a*) Uskottavin teoria mielentilojen luonteesta on funktionalistinen systeemiteoreettinen malli, ja *b*) paras teoria mentaalisesta kausaatiosta on komputationalistinen. Tarkemmin sanoen mielentilat ovat propositionaalisten representaatioiden ja organismin välisiä komputationaalisia relaatioita, joten 3°*a*) mentaaliset representaatiot ovat propositioiden kaltaisia ajattelun kielen lauseita, ja *b*) paras malli organismin mentaalisista prosesseista perustuu loogisten deduktioiden kaltaisille ajattelun kielen heuristisille komputaatioille.

Edellisessä alaluvussa nähtiin syitä epäillä, että tässä kuviossa on jotain vialla. Vaikuttaa siltä, että kognition lisäksi käyttäytymistä ohjaavat muutkin tekijät, kuten tunteet ja toistaiseksi huonosti ymmärretyt alitajuiset intuitiiviset prosessit. Lisäksi kognitiivisen systeemin rooli on mahdollisesti liioiteltu ja väärinymmärretty myös päättelyjä ja muita symbolisia representaatioita vaativissa toimissa. Komputationaalinen mekanismi saattaa olla ainoa tapa selittää naturalistisesti, miten rationaalinen toiminta on mahdollista, mutta tämä oivallus ei ole kovin paljoa auttanut ymmärtämään, miten valtavasta tietomäärästä oleellinen pystytään seulomaan epäoleellisesta tehokkaasti tilanteen vaatimalla tavalla. Tämä on osoittautunut ongelmalliseksi jopa rajatuissa konteksteissa, mutta yleisluontoisesti älykäs käyttäytyminen edellyttää tätä vieläpä elävässä todellisuudessa, jossa tilanteet ovat luonteeltaan avoimia ja aika monesti kortilla. Ongelman ydin ei nähdäkseni piile komputationalismissa sinänsä, vaan enemmänkin kognitivismin – tai tarkemmin sanoen tekoälyn ja kognitivistisen filosofian – sitoutumisesta propositionaaliseen representaatioformaattiin ja siitä kumpuavaan deduktiomalliin. Abstrakti matemaattinen logiikka ei ylipäätään vaikuta kovin luontevalta mallilta kuvaamaan sitä, miten organismien olettaisi representoivan maailmaa ja käyttävän tietoaan. Kuitenkin kun kiinnittää huomiota laskennan mallien ja komputationalistisen mielenteorian syntyyn, on helppo nähdä, miten matemaattisen logiikan ideat on salakuljetettu kognitivistisen teorian ytimeen.

Laskennan mallit, erityisesti niiden arkkityyppi Turing-kone, syntyivät teoreettisen matematiikan, eivät suinkaan psykologian kysymyksistä. Putnamin huomiot, että fysikaalisesti implementoiduilla komputationaalisilla systeemeillä on tiettyjä mielelle ominaisia piirteitä, ovat kuitenkin mielenkiintoisia ja filosofisesti arvokkaita. Lisäksi mielen suhde ruumiiseen vaikuttaa ainakin osin samankaltaiselta kuin ohjelman suhde sitä suorittavaan koneeseen. Näin ollen on varsin luonteva hypoteesi, että mieli on eräänlainen komputationaalinen systeemi, tai vähintään, että tällaisten systeemien filosofinen analyysi voisi luoda valoa mentaalisuuteen liittyviin käsitteellisiin ongelmiin. Tämä oivallus ei kuitenkaan kerro vielä juuri mitään mentaalisista representaatioista tai mielen rakenteesta. Logiikka puolestaan tarjoaa ilmaisuvoimaisen formaalikielen propositioiden ja teorioiden representoimiselle sekä kalkyylin päättelyille. Kun looginen formalismi implementoidaan fysikaalisesti, saadaan systeemi, jossa propositionaalisten representaatioiden esiintymien väliset kausaaliset suhteet mallintavat propositioiden välisiä deduktiivisia suhteita. Kun

tämä yhdistetään representationaaliseen mielenteoriaan, jossa mielentilojen katsotaan olevan propositionaalisia asenteita, vaikuttaa logiikka tarjoavan ainakin oikeansuuntaisen formalismin mielenteorialle. Toisaalta, samoin kuin laskennan mallit, myöskään moderni matemaattinen logiikka ei syntynyt psykologisista, vaan matemaattisista kysymyksistä, tarkemmin sanoen tarpeesta luoda varma perusta matemaattiselle päättelylle. Vaikka logiikka on perinteisesti edustanut myös yleistä järkeilyn teoriaa, ei sitä voi pitää deskriptiivisenä teoriana, joka kertoo mitä järjenkäytön periaatteita ihmiset todella noudattavat, vaan enneminkin kyseessä on normatiivinen teoria, jonka tehtävänä on määrittää, mitä periaatteita rationaalisen ajattelijan ei ole lupa rikkoa.

Edellisessä alaluvussa nähtiin, miten ihmisten looginen päättelykyky on varsin heikkoa. Tämä antaa syyn olettaa, ettei logiikka tarjoa kovin hyvää mallia ajattelusta. Logiikkaan perustuvassa mallintamisessa myös propositionaalinen representaatioformaatti on psykologisesti varsin epäilyttävä. Propositioiden ajatellaan olevan jonkinlaisia sisällönkantajia, missä sisällön katsotaan palautuvan niihin ehtoihin, joiden vallitessa propositio on tosi. Propositio on siis eräänlainen kontekstiton ja subjektiton asiaintilojen kuvaus, sikäli, että propositionaalisen sisällön pitäisi olla periaatteessa sama kaikille ja kaikissa mahdollisissa tilanteissa. Proposition merkitys on siis perspektiivistä riippumaton. Loogisessa viitekehyksessä käsitteen käsite on vastaavalla tavalla subjektiton. Logiikasta ponnistavassa kielen- ja mielenfilosofiassa käsitteiden katsotaan tyypillisesti olevan määriteltävissä antamalle ne riittävät ja välttämättömät ehdot, joiden pätiessä olio tai ilmiö lankeaa sen alaan. Tällaisessa viitekehyksessä käsitteillä ei ole juuri mitään tekemistä sen kanssa, miten oliot maailmaa hahmottavat, vaan käsitteet ovat hyvin epäpsykologisia otuksia,[70] joiden ala paljastuu loogisen merkitysanalyysin, ei psykologisen tutkimuksen avulla. Loogisessa analyysissä ei sinänsä ole mitään vikaa, mutta jos mentaalisia ilmiöitä mallinnetaan logiikkaa hyväksikäyttäen, kulkeutuu tällainen logiikan kannalta hyvin oleellinen, mutta psykologian kannalta irrelevantti käsitteen käsite ikäänkuin vaivihkaa mielenteorian ytimeen. Jos mielen ensisijainen tehtävä on ohjata organismia monimuotoisessa ympäristössä, olisi perin kummallista, jos mentaaliset representaatiot pohjimmiltaan olisivat riippumattomia organismille ominaisista tarpeista sekä sen kyvyistä aistia ja manipuloida ympäristöään. Lisäksi tuntuisi melko epätarkoituksenmukaiselta perustaa käsitteellistäminen ja toiminnanohjaus ensisijaisesti abstrakteihin, kontekstista ja perspektiivistä riippumattomiin, representaatioihin.

Kognitiivisen psykologian valtavirtaa edustavien näkemysten mukaan käsitteet perustuvat jonkinlaisiin prototyyppeihin tai esimerkkeihin. Tällaisissa teorioissa käsitteen ala on epämääräinen ja siihen kuuluminen määräytyy jonkinlaisen samankaltaisuuden perusteella, missä mittapuuna on käsitteen prototyyppinen tai tyypillinen edustaja (Eysenck & Keane, 2000, s.288–293). Esimerkiksi varpuset ja varikset edustavat ainakin useimmille tyypillisempiä *lintu*-käsitteen instansseja, kuin strutsit ja pingviinit. Prototyyppiteorioiden mukaan ihmisten ensisijaisia käsitteellisiä kategorioita muodostavat keskikokoiset objektit, joita yhdistää jotkin selkeät luokittelua helpottavat havaittavat piirteet, tai joita voidaan käyttää tai manipuloida samoilla tavoilla. Hyvä perusesimerkki tälläisesta käsitteestä on vaikkapa *tuoli*, jonka alaan lankeavat sunnilleen samanmuotoiset objektit, joille

[70] Ks. esim. (Eysenck & Keane, 2000, s.285–288).

voi kätevästi istua, joita voi yleensä helposti siirrellä ja joita yleensä löytyy talojen sisältä. Peruskäsitteistä voidaan sitten johtaa täsmällisempiä toissijaisia käsitteitä liittämällä niihin joitain ominaisia lisämääreitä, esimerkiksi nojatuoli, kiikkutuoli ja niin edelleen. Prototyypeistä voidaan johtaa myös abstraktimpia käsitteellisiä luokkia, kuten esimerkiksi *huonekalu*, jotka yhdistävät useita prototyyppisiä käsitteitä jonkin samankaltaisuusperiaatteen mukaan. (Rosch, 1978, s.30–34) On toisaalta selvää, että prototyyppiteoriat eivät kuvaa täydellisesti ihmisten käsitteellistyskykyä. Esimerkiksi valaat kuuluvat nisäkkäiden, eivätkä kalojen kategoriaan, vaikka ne lähinnä muistuttavat jälkimmäisiä. Lisäksi monien abstraktien käsitteiden, esimerkiksi käsitteen *alkuluku*, kannalta nimenomaisesti määritelmät ovat oleellisia, eivät prototyypit. Mahdollisesti kategorisointi perustuu kahteen periaatteeseen, määritelmiin ja prototyyppeihin, jotka painottuvat eri tavalla eri abstraktiotason käsitteissä.[71]

Prototyyppiteorian huomattavan kehittäjän Eleanor Roschin teoriassa mielenkiitoista on, että käsitteet eivät ensisijaisesti perustu abstrakteihin määritelmiin, vaan kategorisoinnin perusta on organismin sensorisissa ja motorisissa kyvyissä. 1980-luvulta lähtien kognition mallinnuksessa on esiintynyt kasvavaa kiinnostusta käsitteiden kehitystä kohtaan, ja esimerkiksi Jean Piagetin vanhat kehityspsykologiset ideat, joilla on tiettyjä yhtäläisyyksiä prototyyppiteorian kanssa, ovat nousseet marginaalista valtavirtaan (Boden, 2006, s.253–255). Piagetilaisen teorian mukaan kehittyvä vauva hahmottaa maailmaa aluksi sensorimotoristen skeemojen avulla, joiden varaan abstraktimpi ajattelu myöhemmin kehittyy. Skeemat ovat eräänlaisia käsitteiden esiasteita, jotka eivät varsinaisesti poimi olioita ja ilmiöitä vaan havainnossa systemaattisesti esiintyviä rakenteellisia säännönmukaisuuksia. Tällaisia esikäsitteitä voisivat esimerkiksi olla muodot sekä eräänlaiset relationaaliset määreet, kuten *sisä-* ja *ulkopuolella*, *alla* ja *päällä*, tietynlaiset rakenteet, kuten *polku*, *tuki* ja niin edelleen (Mandler, 2004, s.78–85). Skeemat eivät välttämättä ole näin primitiivisiä. Esimerkiksi yksi varhaisimmista vauvan oppimista kategorisoinneista on toimijoiden ja passiivisten objektien erottelu, joka suunnilleen vastaa käsitteellistä jakoa eläimien ja muiden olioiden välillä. Tämä erottelu saattaa vaikuttaa abstraktilta, mutta kehittyvän lapsen tarpeiden ja mahdollisuuksien perspektiivistä se lienee melko primitiivinen, mikäli kyse on pohjimmiltaan manipuloitavien olioiden erottamisesta niistä, jotka itse manipuloivat ympäristöä (*ibid.*, s.93–99). Luonnollisesti lapsen kannalta varsin oleellisia ovat ainakin sellaiset oliot, jotka voivat tehdä jotain sen tarpeiden hyväksi.

Kehityspsykologinen tutkimus on kognition mallintamisen kannalta oleellista, koska se tarjoaa mahdollisuuden tarkastella, mille perustalle mentaalisten representaatioiden järjestelmä rakentuu. Jos käsitteet todella pohjautuvat organismin ruumilisisita kyvyistä ja tarpeista riippuviin skeemoihin, niin primitiiviset representaatiot eivät ole luonteeltaan abstrakteja ja propositionaalisia, vaan organismin biologiseen olemukseen ja perspektiiviin sidottuja. Mutta miten kyky abstraktiin ajatteluun, esimerkiksi suunnitteluun, ennustamiseen ja kontrafaktuaaliseen päättelyyn, voi nousta tältä pohjalta? Kuten edellisen alaluvun alussa todettiin, tämä alue psykologiasta on melko tuntematonta. Tehdään kuitenkin pieni vilkaisu pariin spekulatiiviseen, mutta filosofisesti mielenkiintoiseen teoriaan.

[71] Ks. (Armstrong et al., 1983, s.234–238), jossa osoitetaan, että prototyyppikategorisoinnille tunnusomaisia ilmiöitä esiintyy hieman yllättäen jopa tarkkarajaisten matemaattisten käsitteiden kategorisoinnissa.

Mikäli perspektiivisidonnaiset representaatiot syntyvät yhdessä maailman havainnoimisen ja manipuloimisen kanssa, lienee mahdollista, että abstraktit representaatiot kehittyvät samaan tapaan opittaessa enenevässä määrin havainnosta riippumattomia taitoja, jotka syntyvät yleistyksenä useista perspektiivisidonnaisista kyvyistä. Mahdollisesti tämä prosessi kulminoituu enemmän tai vähemmän objektiivisen näkökulman kehittymiseen, eli kykyyn tehdä arvostelmia, päätelmiä ja käsitteellisiä erotteluja tietystä perspektiivistä riippumatta. Tälläista teoriaa on kehitellyt muun muassa Adrian Cussins (1990). Kognitiotieteilijä Ronald Chrisley on onnistunut demonstroimaan tämän ilmiön laatimalla neuroverkon, jonka avulla simuloitu robotti pystyi opettelemaan paikasta toiseen navigoimista. Aluksi systeemin toiminta perustui maamerkkien havainnointiin, mutta oppimisen edetessä riippuvuus havainnoinnista väheni, ja lopulta robotti kykeni suunnittelemaan reittinsä ennakolta. Tämä perustui siihen, että systeemi oppi ennustamaan, mitä se havaitsisi mistäkin paikasta käsin, jolloin se pystyi soveltamaan perspektiivisidonnaista suunnistuskykyään varsinaisesti menemättä paikan päälle. (Chrisley, 1990; Chrisley & Holland, 1995)

Mark Johnson ja George Lakoff ovat kehitelleet mielenkiintoista teoriaa, jonka mukaan keskeinen abstraktion mekanismi on metafora. Ajatus tässä on, että abstraktit käsitteet, kuten *ymmärtää* ja *tietää*, ovat metaforisia johdannaisia alkujaan kehollisesti tai sensorimotorisesti muodostuneista käsitteistä, kuten *tarttua* ja *nähdä* (Lakoff & Johnson, 1999, 45–59). *Metafora* ei tässä yhteydessä tarkoita epäkirjaimelliseksi tarkoitettua kielellistä ilmausta, vaan tapaa hahmottaa uusia – pääasiassa abstrakteja – ilmiöitä jonkin jo edeltä ymmärretyn ilmiön sisäisen logiikan perusteella. Esimerkiksi paikasta toiseen siirtymiseen liittyviä käsitettä voidaan käyttää ymmärtämään vaikkapa ihmissuhteita. Johnson ja Lakoff esittävät, että esimerkiksi ilmaukset ”suhde on umpikujassa”, ”. . . polkee paikoillaan”, ”. . . etenee vauhdilla” tai ”. . . ei johda mihinkään” eivät ole runollisia kielikuvia, vaan heijastavat tapaamme ymmärtää ihmissuhteisiin liittyviä ilmiöitä soveltamalla matkaamiseen liittyvää käsitteistöä. Tällainen kokonaisen ilmiöluokan ymmärtäminen toisen avulla on eräänlainen korkeamman asteen metafora, tässä yhteydessä *rakkaus on yhteinen matka*. Tätä samaa rakennetta voi luonnollisesti soveltaa myös vaikkapa työuran käsitteellistämiseen. Korkean asteen metaforat muistuttavat hieman teorioita, sikäli että molemmat ovat kognitiivisia välineitä, jotka mahdollistavat abstraktin ymmärtämisen ja päättelyn. (Lakoff & Johnson, 1999, s.61–73)

Ikävä kyllä näihin teorioihin syventyminen ei liene tässä yhteydessä mahdollista, muttei se toisaalta olisi asian kannalta myöskään kovin oleellista. Yllä olevien esimerkkien tarkoitus on lähinnä tuoda esiin vaihtoehtoja abstraktiin päättelyyn liityvälle propositionaalisten mielensisältöjen teorialle, tai ainakin huomauttaa, että tällaisia teorioita on tarjolla. Näin ollen klassinen komputationalismi ei suinkaan ole ainoa malli, mitä markkinoilta löytyy. Lisäksi näissä teorioissa mielenkiintoista on, miten ne tarjoavat selonteon mentaalisten representaatioiden sisällöistä juurruttamalla peruskäsitteet organismin toiminnallisiin ja kehollisiin valmiuksiin.

Toisaalta jotkut vanhan koulukunnan komputationalistit ovat painottaneet, että heidän teoriansa koskee nimenomaisesti korkeita kognitiivisia prosesseja, eli lähinnä abstraktia käsitteellistä ajattelua. Mikäli tällainen kyky noudattaa jotain omia periaatteitaan, ei sen

kehittymisen tai implementaatioteorian tutkiminen välttämättä ole tarpeen. Muun muassa ehkä hieman yllättäen Fodor on painottanut eri teoksissaan, ettei hänen kannattamansa komputationalismi yritäkään olla kattava teoria mentaalisesta toiminnasta (Fodor 1983, s.126–129; 2000, s.1). Toisekseen tekoälytutkija voi katsoa työnsä koskevan älyllisen toiminnan yleisiä edellytyksiä, eikä inhimillisen ajattelun erityispiirteitä. Heuristiset algoritmit voisivat edustaa tällaisen tutkimuksen tuloksia. Koska algoritmien komputationaalinen vaativuus on sama kaikille äärellisille mekanismeille, luultavasti mikä tahansa olio joutuu käyttämään heuristiikkoja ratkoessaan eksponentiaalista aikaa vaativaa ongelmaa. Kuitenkin edellisessä alaluvussa esitettiin syitä epäillä, ettei inhimillisesti toimivaa kognitiota voi mallintaa ymmärtämättä sen erityispiirteitä. Näin ollen ajattelun mallintaminen tunteista, sensorimotorisista kyvyistä, hahmontunnistuksesta ja intuitioista irrallisena ilmiönä voi olla mahdotonta. Mikäli järjenkäytön yleisten reunaehtojen tutkiminen tarkoittaa ruumiittomien, tunteettomien ja perspektiivittömien tietojenkäsittelyprosessien tarkastelemista, vaikuttaa inhimillinen äly jäävän tutkimuksen ulkopuolelle. Klassinen komputationalismi on keskittynyt ongelmien rakenteiden ja ratkonnan analyysiin, jonka oletetaan heijastavan mentaalisten representaatioiden ja prosessien luonnetta. Kognitivismin tavoitteiden kannalta olisi kuitenkin syytä keskittyä mielen ja maailman vuorovaikutukseen. Vastaukset mielen rakenteen ja toiminnan kysymyksiin löytynevät keskittymällä siihen, mitä tapahtuu mielen ja maailman välissä, ei pelkästään ensiksi mainitun sisällä.

On äkkiseltään vaikea sanoa, ratkaiseeko sensorimotorisiin vuorovaikutustaitoihin perustuva mielentutkimus kolmea aiemmin käsiteltyä kognitivismin ongelmaa, mutta ainakin tällainen tutkimus huomioi olioiden ruumiillisuuden ja kiinnittää huomiota myös muihin psykologisiin kykyihin, kuin järkeilyyn. Näin ollen lähestymistapa saattaa luoda valoa tunteiden, taidokkaan toiminnan ja intuition kysymyksiin. Epäselvää kuitenkin on, miten tällaiset painotukset millään tavalla koskettaisivat kehysongelmaa. Toisaalta terveen järjen taustalla olevat mekanismit tunnetaan tällä hetkellä niin huonosti, että on äkkiseltään hyvin hankala sanoa, minkälainen tutkimus niitä voisi juurikaan koskettaa. Kuitenkin mikäli abstraktin ajattelun perusta lepää sensorimotoristen valmiuksien päällä, saattaa tämänkaltainen tutkimus valaista uudella tavalla mentaalisten representaatioiden sekä tiedon organisoinnin ja käytön luonnetta. Ehkä avaimet tekoälyn ongelmiin löytyvätkin robotiikasta eikä logiikasta. Robotiikka- ja keinoelämäpohjainen tekoälytutkimus on varsin nuorta, joten sen tuloksia ja mahdollisuuksia on tässä vaiheessa hankala arvioida. Lisäksi jos ruumiillisten kykyjen ja tarpeiden huomioinnin ottaa vakavasti, saattaa täysverisen inhimillisen kognition mallintaminen edellyttää nimenomaisesti inhimillisillä kyvyillä ja tarpeilla varustetun systeemin toteuttamista, mikä lienee tällä haava käytännössä mahdotonta.

Yllä esitetyt kriittiset huomiot ovat pääpiirteiltään antikomputationalistisia, tarkemmin sanoen klassiseen komputationaaliseen teoriaan suuntautuvia. Kuitenkin myös funktionalismin keskeiset teesit vaikuttavat hieman arveluttavilta näiden tarkastelujen valossa. Ensinnäkin jos mentaaliset representaatiot eivät ole yleisesti ottaen propositionaalisia, eivät myöskään mielentilat pääpiirteissään ole propositionaalisia asenteita. Funktionalismin kantava idea kuitenkin on tarjota reduktiivinen teoria juuri tällaisille mielentiloille. Toisekseen argumentit funktionalismin puolesta perustuvat pitkälti siihen, että muunlaiset teoriat eivät kykene selittämään, miten erilaiset oliot voivat jakaa samanlaisia usko-

muksia, haluja, pelkoja ja vastaavia. Jos mielentilat eivät kuitenkaan ole tällaisia, herää kysymys, onko mitään erityistä syytä kannattaa funktionalistista mielentilojen teoriaa. Toisaalta taas ehkä jotkin mielentilat ovat propositionaalisina asenteina, ja homunkulaarinen funktionalisti voi katsoa teoriansa koskevan juuri tällaisia mentaalisia ilmiöitä ja niiden vuorovaikutuksia. Kuitenkin jos psykologiset lainalaisuudet edellyttävät viittaamista niin sanotusti alempiin tai suorastaan epäkognitiivisiin psykologisiin ilmiöihin, kuten tunteisiin ja sensorimotorisiin kykyihin, eivät korkeamman kognition lainalaisuuden ole muotoiltavissa täysin itsenäisesti. Mikäli näin on, ei ole mitenkään selvää, että mieltä voi todella analysoida kerroksittain homunkulaarisen funktionalismin mukaisesti. Tällöin funktionalismin keskeiset ongelmat palaavat, ja vieläpä hankalammassa muodossa, koska mahdollisesti myös ruumiillisesti, eikä vain kognitiivisesti, erillaiset oliot edellyttävät erilaisia funktionaalisia kuvauksia analyysin korkeimmalle tasolle saakka. Tällöin vaikuttaa perin toivottomalta, että funktionalismin puitteissa kyettäisiin määrittämään mentaalisia olioita yhdistävää kuvausta.

Ehkä vahvin syy olettaa mielentilojen olevan propositionaalisia asenteita on se, että ilmeisesti tarvitsemme niitä vastaavia intentionaalisia käsitteitä ymmärtämään tarkoituksenmukaista käyttäytymistä ja poimimaan toiminnasta nimenomaisesti psykologiset piirteet. Mutta vaikka näin olisikin, ei tästä vielä seuraa, että mielentilat ja psykologiset prosessit välttämättä vastaisivat mitenkään suoraviivaisesti arkipsykologian ontologiaa ja yleistyksiä. Muun muassa Daniel Dennett on tunnettu homunkulaarisen funktionalismin kehittäjä, mutta ehkä vielä tunnetumpi puoli hänen ajattelussaan on hänen mielenkiintoinen teoriansa arkipsykologisesta selittämisestä. Dennett kutsuu *intentionaalisiksi systeemeiksi* olioita, joiden toimintaa voidaan selittää ja ymmärtää soveltamalla arkipsykologiaa (Dennett, 1971, s.87). Hän erottaa persoonallisen intentionaalisen psykologian alipersoonallisesta kognitiivisesta psykologiasta, joista ensimmäinen kuvaa systeemin toimintaa uskomus–halu-käsitteillä ja jälkimmäinen taas laskennan ja tietojenkäsittelyn termein. Dennettin mukaan nämä selityksen tasot ovat toisistaan riippumattomia. Syy tähän on, että järjellisen oloisesti toimiviin systeemeihin voidaan soveltaa arkipsykologista selittämistä riippumatta siitä, katsotaanko niillä varsinaisesti olevan haluja, uskomuksia tai ylipäätään yhtään mitään mielentiloja.

Esimerkiksi shakkiohjelma voi pyrkiä hallitsemaan laudan keskialuetta ja saattamaan kuningattaren mahdollisimman aikaisin peliin sisältämättä mitään varsinaisia sääntöjä näihin päämääriin tähtääville siirroille. Tällainen toiminta ei siis edellytä minkäänlaista uskomus- tai halujärjestelmää, joka sisältäisi propositiot ”hallitse keskialuetta” tai ”siirrä kuningatar peliin ajoissa”, vaan mainittu säännönmukainen käyttäytyminen voi yksinkertaisesti kummuta ohjelman heuristisista sääntöjen monimutkaisesta vuorovaikutuksesta. Tästä huolimatta ohjelman toimintaa voidaan ennustaa ja ymmärtää vetoamalla sen oletettuihin uskomuksiin ja haluihin. Erityisesti jos ihminen pelaa konetta vastaan, voidaan molempien toimintaa selittää samalla tavalla ilman, että niiden tietojenkäsittely- ja päätöksentekomekanismeilla olisi juurikaan mitään yhteistä. Jos kone pelaa tarpeeksi rationaalisesti, voidaan sanoa, että se esimerkiksi *pyrkii* syömään vastustajan kuningattaren ja tilaisuuden tultua tekee niin, jos se *uskoo*, että tästä siirrosta koituu sille enemmän hyötyä kuin haittaa. Tällainen suhtautumistapa saattaa olla jopa välttämätöntä. Shakkikone voi olla niin monimutkainen, että sen ohjelmakoodista on äärimmäisen vaikeaa

päätellä, mitä se missäkin tilanteessa tekee, jolloin paras tapa ennustaa sen käyttäytymistä on vedota sen uskomuksiin ja pyrkimyksiin, jotka oletettavasti heijastavat sitä, mikä shakissa yleisesti ottaen on mielekästä. Toiminnan ennustamisen kannalta ohjelman sisäisellä rakenteella ei ole merkitystä, jos sen käyttäytyminen on tarpeeksi rationaalista ja systemaattista, jotta sen pitäminen intentionaalisena systeeminä tuottaa pääasiassa oikeita ennustuksia. (Dennett 1971, s.87–93; 1987, s.107; 1981, s.58–65) Dennettin käsitys arkipsykologian soveltamisesta on siis vahvasti instrumentalistinen.

Näin ollen arkipsykologia ei erottele mentaalisia olioita ei-mentaalisista, koska selitysmallin soveltamisen kannalta on samantekevää, ajatellaanko tarkasteltavalla systeemillä oikeasti mielentiloja vaiko ei. Tästä seuraa käsiteltävän asian kannalta keskeinen johtopäätös: Jos ohjelman tapauksessa arkipsykologisen selityksen toimivuudesta ei voida päätellä sen sisäisten tilojen olevan propositionaalisia asenteita, aivan vastaavasti tällaista päätelmää ei voida tehdä myöskään ihmisten tapauksessa. Näin ollen intentionaalinen selittäminen ei varsinaisesti koske systeemin sisäisiä ilmiöitä, vaan propositionaalisia asenteita käytetään poimimaan tiettyjä säännönmukaisia käyttäytymispiirteitä. Tämä on oikeastaan aika ilmeistä, jos katsotaan hieman tarkemmin esimerkiksi praktista syllogismia. Ajatellaanpa, että Riitalla on nälkä hän uskoo, ettei hänen jääkaapissaan ole ruokaa, mutta naapurikorttelin ravintola tarjoilee lounasta käypään hintaan. Tämä olisi aivan pätevä selitys sille, miksi Riitta päätyi asioimaan mainitussa ravintolassa. Selitys kuitenkin ohittaa täysin ne mielenkiintoiset mentaaliset prosessit, joiden mallintaminen on osoittautunut tekoälytutkimuksessa hyvin vaikeaksi. Miksi Riitta ei lähde varastelemaan ruokaa kaupasta tai matkusta lounaalle Berliiniin? Molemmat vaihtoehdon tyydyttäisivät hänen tarpeensa, joten syllogismi ei voi olla riittävä selitys Riitan toiminnalle. Selitystä vaatii, miksi juuri tietty uskomus tai toimintavaihtoehto on hänen toimintansa kannalta relevantti kaikkien mahdollisten vaihtoehtojen joukosta. Yleensä tämä kysymys ei herää, koska osaamme itse intuitiivisesti arvioida, mitkä seikat ovat toiminnan kannalta oleellisia ja mitkä toimet mielekkäitä. Tekoälytutkimuksen ehkä tärkein opetus on, että tämä intuitiivinen arviointikyky perustuu jonkinlaisille epätriviaaleille ja erinomaisen tärkeille psykologisille prosesseille, jotka tulee selittää, eikä arkipsykologisen selittämisen tapaan vain olettaa, jotta mielekkään toiminnan etiologiaa voisi ymmärtää ja mallintaa.

Jos kerran komputationalismi on epäilyttävä mielenteoria ja funktionalismi uhkaa upota sen mukana, jääkö kognitivismista oikeastaan mitään hedelmällistä käteen? On syytä huomata mihin tässä luvussa esitetty kognitivismin kritiikki tarkalleen ottaen kohdistuu. Tähtäimessä on ollut erityisesti klassinen komputationaalinen teoria, joka kattaa lähinnä ajattelun kielen teorian ja siihen sisältyvän käsityksen propositionaalisten representaatioiden ensisijaisuudesta sekä deduktiosta keskeisimpänä psykologisena mekanismina. Funktionalismissa puolestaan ongelmallista on oletus, jonka mukaan psykologiset mekanismit ja tilat ainakin suunnilleen vastaavat arkipsykologisia latteuksia, ja erityisesti homunkulaarisessa versiossa, että mieltä voidaan analysoida kerroksittain siten, ettei esimerkiksi organismeille ominaisia ruummillisia kykyjä ja tarpeita tarvitse huomioida uskomusten ja halujen vuorovaikutusten kuvaamisessa. Nämä teoreettiset sitoumukset eivät kuitenkaan ole välttämättömiä komputationalismille eivätkä myöskään funktionalismille.

Funktionalistisen teorian keskeinen tavoite on yrittää poimia mentaalisten olioiden luonnollinen luokka. Taustalla on käsitys, jonka mukaan pätevän mielenteorian tulee kyetä selittämään millä perusteella mentalistisia predikaatteja voidaan soveltaa erilaisiin olioihin, ja samalla luonnollisesti määritellä se olioiden joukko, joka näitä predikaatteja ei toteuta. Lähtökohtana tässä on, että vaikuttaisi olevan mielekästä pitää hyvin monenkaltaisia olioita uskovina ja haluavina, mutta teorian tulisi kyetä myös rajaamaan mentaalisten olioiden joukosta pois esimerkiksi yksinkertaiset tietokoneohjelmat, termostaatit ja muut ilmeisen mielettömät oliot. Toisin sanoen funktionalistisen teorian tavoite on erotella oliot, joilla oikeasti on intentionaalisia tiloja, olioista joilla niitä ei ole.

Toisaalta funktionalismissa on filosofisesti keskeistä myös, että mentaalisuudessa ei ole kysymys siitä, mitä olio on vaan mitä se tekee, missä tekeminen viittaa havaittavan käyttäytymisen lisäksi organismin sisäisiin prosesseihin. Sinänsä erityisesti psykofunktionalismi ei ole sitoutunut väittämään, että nämä prosessit vastaisivat arkipsykologisia ilmiöitä. Jos esimerkiksi dennettiläisen analyysin mukaisesti propositionaaliset asenteet eivät poimi varsinaisia kognitiivisia prosesseja, vaan olioiden käyttäytymistaipumuksia, niin funktionaalinen analyysi ei koske intentionaalista, vaan alipersoonallista psykologiaa. Tällainen alipersoonallinen funktionalismi säästää intuition, jonka mukaan hyvin monenlaiset oliot voivat toteuttaa samanlaisia uskomus–halu-yleistyksiä. Myös marsilaisesimerkkien opetus pätee yhä, koska alipersoonallinen funktionalismi ei erottele psykologisesti identtisiä olioita. Näin ollen merkittävä siivu funktionalismia voidaan säästää luopumalla arkipsykologian ja funktionaalisten analyysien samaistamisesta. Mitä käsittääkseni kuitenkin väistämättä menetetään, on funktionalismin mahdollisuus määritellä mentaalisten olioiden luonnollinen luokka. Tämä seuraa siitä, että mitä tahansa systeemiä voidaan kuvata tavalla tai toisella epämääräisen sopivasti abstraktina kausaalisena systeeminä, ja toiseksi samanlaisia uskomus–halu-yleistyksiä noudattavat oliot voivat olla alipersoonalliselta funktionaaliselta rakenteeltaan hyvinkin erilaisia. Näin ollen ei ole mitään syytä olettaa, että intentionaalisia systeemeitä yhdistäisi samankaltainen funktionaalinen organisaatio, eikä myöskään ole mielekästä tehdä heikompaa oletusta, että systeemi on intentionaalinen vain jos sen rakenne toteuttaa jonkin funktionaalisen kuvauksen.

Jääkö tällöin mitään vahvaa syytä kannattaa funktionalismia? Teorian mielekkyyden voi oikeuttaa vetoamalla sen hedelmällisyyteen kognitiivisen psykologian metodologiana, eli että paras tapa muotoilla alipersoonallisen kognitiivisen psykologian teoriat on sopivan abstrakti kausaalinen kuvaus systeemin sisäisistä prosesseista. Näin mielenteoria voi olla funktionalistista ilman pyrkimystä redusoida arkipsykologia kognitiivisen systeemin teoriaan. Huomautettakoon, että tämänlaisen analyysin ei tarvitse sitoutua väittämään, että ihmisten alipersoonallinen psykologia ei muistuta arkipsykologiaa. Ydinajatus on vain, että arkipsykologisten yleistysten soveltamisen kannalta tämä on samantekevää. Allekirjoittaneelle tämä vaihtokauppa kyllä käy. Luulisin mentaalisuuden olevan lähinnä esitieteellinen käsite, joka ei oikeastaan viittaa mihinkään yksiselitteiseen ilmiöön vaan olioihin, jotka toimivat jotenkin enemmän tai vähemmän joustavasti ja rationaalisesti monipuolisessa ympäristössä. Joidenkin mukaan mentaalisuus edellyttää rationaalisuutta, toisten mielestä kieltä, kolmannen tietoisuutta ja mitä nyt milloinkin. Tieteelliseen keskusteluun mentaalisuudesta puhuminen tuskin tuo mitään erityistä lisäarvoa, vaan on ehkä parempi yksinkertaisesti puhua tähän käsitteeseen liittyvistä ilmiöistä erikseen. En siis näe erityis-

tä tarvetta kysyä, mitkä oliot ovat mentaalisia ja mitkä eivät, vaan on mielekkäämpää tutkia minkälaisia mentaalisuuteen liittyviä ilmiöitä mikäkin olio ilmentää ja millä tavalla. Filosofisen ymmärryksen kannalta vastaavasti lienee valaisevampaa huomata tämän käsitteen olevan eräänlainen esitieteellisesti poimittu ryväs erinäköisiä ilmiöitä, kuin pyrkiä määrittelemään, mitkä mentaalisuuden riittävät ja välttämättömät ehdot ovat.

Entä miltä näyttää komputationalismin kohtalo? Kuten aiemmin on jo esitetty, teorian ongelmat näyttävät liittyvän lähinnä oletukseen, että mentaaliset representaatiot ovat propositionaalisia ja ajattelu deduktiivista. Komputationalismin ydinoivallus ei kuitenkaan edellytä mitään tällaista. Logiikan kaltainen kalkyyli on melko luonteva ehdokas mentaalisen toiminnan perustaksi, jos taustalla on propositionaalisille asenteille perustuva mielenteoria sekä näkemys, jonka mukaan ajattelu on kielenkaltaisten symbolirakenteiden käsittelyä. Jos nämä teoreettiset taustaoletukset hylätään, on hankala keksiä komputationalismista sinänsä mitään vikaa. Tarjolla on kaikenlaisia laskennan malleja, jotka eivät ole luonteeltaan deduktiivisia, eivätkä edes kovin helposti tulkittavissa symbolienkäsittelyksi, kuten esimerkiksi neuroverkot, soluautomaatit ja analoginen laskenta. Toisaalta komputationalismi vaikuttaa varsin tyhjältä teorialta ilman minkäänlaista selontekoa mentaalisten prosessien luonteesta. Jos komputationaalisen teorian ainoa tarkoitus on tarjota jokin mekanismi toteuttamaan funktionalismin edellyttämä kausaalinen rakenne, miksei saman tien tehdä pienintä mahdollisinta oletusta, että mekanismi on suoraan fysikaalinen?

Vastaus löytyy taustalla olevan mielenteorian muotoilusta. Mikäli kognitiivisen psykologian teorianmuodostus perustuu mentaalisten ilmiöiden mallintamiseen jonkinlaisina representaatioiden käsittelyjärjestelminä, ovat nämä prosessit komputationaalisia suoraan määritelmän perusteella. Representaation ja laskennan käsitteet ovat sen verran epämääräisiä, että en nyt ryhdy käsittelemään, mitä tarkalleen ottaen on mielekästä pitää representaatioiden käsittelynä ja mitä ei, mutta komputationaalisen teorian ei joka tapauksessa tarvitse olla sidottu mihinkään tiettyyn representaatioformaattiin. On hyvin mahdollista, että kognitiivinen systeemi käyttää erilaisia representaatioita esimerkiksi näköjärjestelmän analogisista ja perspektiivisidonnaisista esityksistä kielellisen ajattelun propositionaalisiin representaatioihin. Luonnollisesti ajattelun kielen teorian hylkääminen ei tarkoita symbolisten representaatioiden hylkäämistä kaikkinensa, vaan ainoastaan sen kiistämistä, että ne olisivat keskeisin tai peräti ainoa mentaalisten representaatioiden järjestelmä. Kaikesta huolimatta komputationalismi lienee silti paras käsitteellinen teoria selittämään, miten rationaalisuus, mentaalisten representaatioiden sisällöille perustuva kausaatio ja mielekäs käyttäytyminen ylipäätään ovat mahdollisia fysikaalisen kausaation sulkemassa maailmassa. Toisin sanoen kaikista komputationaalisen teorian puitteissa tehdyistä harha-askelista huolimatta, teoria on silti käsittääkseni paras selitys eräälle mieli–ruumis-ongelman keskeisimmälle mysteerille.

Lukija saattaa olla nyt hieman hämmentynyt tästä edestakaisesta liikkeestä. Tämän työn keskeinen sanoma on ollut, että kognitivistinen mielenteoria on kahden toisistaan periaatteessa riippumattoman komponentin yhdistelmä, joista molemmat ovat jokseenkin epäilyttäviä, mutta toisaalta ytimeltään elinvoimaisia. Käsittääkseni funktionalismin ja komputationalismin filosofisista perusideoista ei ole mitään pakottavaa syytä luopua –

päin vastoin – mutta klassinen kognitivistinen teoria lienee tullut tiensä päähän. Klassinen teoria koskee korkeintaan pientä siivua koko inhimillisestä kognitiosta, ja mikäli kyky symbolirakenteiden käsittelyy perustuu pääpirteissään ulkoisten represenaatioiden kanssa toimimiseen, on tämänkin kyvyn etiologia klassisessa teoriassa ymmärretty väärin. Kognitiotiede on ollut käymistilassa sitten 80-luvun lopun, jolloin neuroverkkomallinnus teki läpimurron kognitiotieteen valtavirtaan. Tämä vapautti tieteenalan klassisen teorian kahleista, ja avasi kognitiotieteissä portit kaikenlaisille uudentyyppisille käsitteen- ja teorianmuodostustavoille. Murros aiheutti myös kasvavaa kiinnostusta vaihtoehtoisiin komputationaalisiin malleihin ja sitä myötä uudenlaisiin painotuksiin tutkimuksen kysymyksenasettelussa ja metodologiassa. Tällä hetkellä kehityspsykologia, aivotutkimus, kehollisuus, robotiikka ja muut enemmänkin biologiasta ja jopa fenomenologiasta, kuin tietojenkäsittelystä ja kielitieteestä, ammentavat tutkimussuuntaukset alkavat olla melko valtavirtaa. On vaikeaa sanoa miltä kognitiotiede näyttää viiden vuosikymmenen päästä, mutta varmasti hyvin erilaiselta kuin viisikymmentä vuotta sitten. Mikäli funktionalismin ja komputationalimin filosofiset ydinajatukset ovat vielä silloin voimissaan, kuten ne kaikesta huolimatta ovat vielä tällä hetkellä, kaiketi tutkimusta voi pitää edelleen eräänlaisena kognitivismina.

Otetaan vielä lopuksi pikainen katsaus kysymykseen, voiko kone ajatella. Kaiken edellä sanotun valossa tämä on edelleen yhtä huonosti määritelty kysymys kuin vuonna 1950, jolloin Turing koitti korvata sen kysymyksellä, voiko kone käyttäytyä inhimillisesti. Kuten sivulla 61 on mainittu, mikä tahansa olio voidaan kuvata koneena yleensä vieläpä monella tavalla, joten sikäli kysymykseen on triviaali vastaus: kyllä, koska kaikki ajattelevat oliot ovat koneita jossain mielessä. Varsinaisesti mielenkiintoinen kysymys on, voidaanko esimerkiksi ihminen kuvata koneena tavalla, joka selittää meidän mentaalisiksi mieltämämme piirteet. En keksi mitään syytä olettaa, ettei tällaisen kuvauksen löytyminen olisi mahdollista. Entä onko mahdollista luoda inhmimillisesti toimivia ja ajattelevia koneita, jotka eivät varsinaisesti ole ihmisiä, vai osoittavatko tekoälyn ongelmat tämän tavoitteen toivottomaksi? Tämä on hieman hankalampi kysymys, mutta en myöskään näe miksi keinotekoisen mekaanisen älyn luominen olisi periaatteessa mahdotonta. Komputationaalisten systeemien universaalisuuden ansiosta ainoa peruste tekoälyn periaatteelliselle mahdottomuudelle olisi, että ihmismieli jotenkin ylittäisi äärellisen mekaanisen laskettavuuden rajoitukset. En ole nähnyt kovin hyviä argumentteja sen puolesta, että näin olisi.[72] Mikäli ihmismieli voidaan kuvata jonkinlaisena tietojenkäsittelymekanismina, voidaan täsmälleen sama mekanismit ilmplementoida tietokoneella. Tämä ei kuitenkaan tarkoita, että paras tapa yrittää ymmärtää mieltä ja ihmisiä olisi tutkia tietokoneita.

Muistettakoon, että jo tekoälyn filosofian klassikkoartikkelissaan Turing esitti kaksi mahdollista metodia inhimillisesti toimivan koneen ohjelmoimiseksi. Toinen oli niin sanottu ylhäältä alas -menetelmä, joka tarkoittaa, että ohjelmointi aloitetaan keskittymällä abstraktien ongelmien ratkontakykyihin, joita kasaamalla ohjelman toivotaan lopulta saa-

[72] Tämä ei tarkoita, etten olisi nähnyt asiaa koskevia huonoja argumentteja, esim. (Lucas, 1961, 1996) ja (Penrose, 1994), jotka kaikesta huolimatta ovat edustavia esimerkkejä asiasta kiinnostuneille. Lucas ja Penrose pyrkivät osoittamaan Gödelin epätäydellisyyslauseista seuraavan, ettei mieli ole komputationaalinen systeemi. Näiden argumenttien perusidean on aiemmin esittänyt jo itse Gödel (1951), joka kuitenkin lähinnä vain huomautti, että joko on olemassa matemaattisia ongelmia, joita ihmiset eivät voi periaatteessakaan ratkaista, tai sitten emme ole koneita.

vuttavan kyvyn yleisesti älykkääseen toimintaan. Näin tekoälyä on perinteisesti tehty, mikä parhaiten näkyy esimerkiksi Newellin ja Simonin töissä. Toinen Turingin ehdottama metodi oli lähestyä ongelmaa täysin toisesta suunnasta rakentamalla kone, joka voi vuorovaikuttaa ympäristönsä kanssa ja jota opettamalla ja parantamalla se saadaan oppimaan uusia kykyjä lopulta saavuttaen jotakuinkin inhimillisen älykkyyden. Sikäli kun tällä on jotain merkitystä, Turing itse katsoi tämän jälkimmäisen lähestymistavan olevan ensiksimainittua mielenkiintoisempi ja hedelmällisempi (Turing, 1951, s.473–474). Ei ole mitenkään ihmeellistä, että kuusi vuosikymmentä sitten tämä menetelmä ei ottanut tulta alleen. On huomattavasti vaikeampaa rakentaa todellisessa ympäristössä jotain mielekästä tekevä robotti, kuin simuloida abstraktien, rajoitettujen ja hyvin määriteltyjen ongelmien ratkaisua tietokoneella. On melko ilmeistä, että tekoälytutkimuksen suuntaa on ohjannut myös klassiseen kognitivismiin elimellisesti liittyvä käsitys, jonka mukaan mentaalinen toiminta on ennen kaikkea järjenkäyttöä tai ylipäätään propositioiden käsittelyä. Mikäli kuitenkin inhimillisen toiminnan perusteet löytyvät ruummillisista valmiuksistamme, alhaalta–ylös-menetelmä saattaa olla käytännössä ainoa toimiva tieteellinen strategia inhimillisen älyn keinotekoiseksi luomiseksi, ja onneksi kehityspsykologia, oppivat systeemit sekä robotiikka otetaan nykyään vakavasti myös tekoälynpiireissä.

Myöskään kognitiivisen antropologian mahdollisuuksia ei tulisi unohtaa. Mielekkyys ei ole puhtaasti formaali kysymys, vaan sidoksissa elämänmuotoomme ja -tapaamme, ja monissa asioissa ympäröivä kulttuurimme määrittää sen, mikä on oleellista ja rationaalista, ei niinkään biologinen olemuksemme. Tämä on melko ilmeistä esimerkiksi moraalin ja tapakulttuurin saralla, mutta sama on nähtävissä myös järjenkäytössä. Esimerkiksi koulutetut länsimaalaiset erottelevat työkalut työstettävästä materiaalista, koska ne kuuluvat käsitteellisesti eri kategoriaan, mutta klassisessa tutkimuksessaan neuvostoliittolainen neuro- ja kehityspykologi Alexander Luria osoitti, että kouluttamattomien Keski-Aasialaisten talonpoikien ajattelumallissa tällaista erottelua ei pidetä mielekkäänä. Tämä johtuu yksinkertaisesti siitä, että työstettävä ja työstön välineet kuuluvat toiminnassa erottamattomasti yhteen ja yksistään molemmat ovat tarpeettomia. Lurian haastattelemat henkilöt eivät tyypillisesti edes ymmärtäneet tämän erottelun merkitystä. (Luria, 1976, s.53–61) Sen lisäksi, että koneeseen yritetään saada ohjelmoitua jonkinlaista rationaalisuutta, olisi syytä miettiä täsmälleen ottaen minkälaista rationaalisuutta tällöin pyritään replikoimaan. Mahdollisesti eräs tekoälytutkimuksen ongelmista onkin tiedostamaton oletus, että meidän – siis tässä tapauksessa oikeastaan matemaattista logiikkaa ja analyyttistä filosofiaa taitavan länsimaalaisen tiedemiehen – tapa hahmottaa maailmaa on jollain tavalla universaali ja vapaa mielivaltaisista oletuksista. Ehkä tästä syystä tätä – tietyin varauksin – meidän rationaalisuutemme luonnetta ei ole tekoälypiireissä ymmärretty tarkemmin analysoida ja yrittää mallintaa, vaan on oletettu, että jotenkin automaattisesti kumpuaa universaaliseksi ymmärretystä logiikasta. Sen lisäksi, että kulttuurisidonnaisten ajatteluja käsitteellistystapojen tutkimus saattaisi tuoda jotain valoa kehysongelmaan, olisi myös hyvin mielenkiintoista, mikäli tulevaisuuden tekoälytutkimus voisi valaista jotenkin täysveristä inhimillisyyttämme, siis ei pelkästään mielen mekanismeja ja ruumillista olemustamme, vaan myös ihmisenä olemisen kulttuurista ulottuvuutta.

Viitteet

Aristoteles 1989: *Nikomakhoksen etiikka.* Suomentanut Simo Knuuttila. Helsinki: Gaudeamus.

Armstrong, S.L.; Gleitman, L.R. & Gleitman, H. 1983: "What Some Concepts Might not Be." *Cognition*, 13, s.263–308. (Teoksesta Margolis, E. & Laurence, S. (toim.) 1999: *Concepts: Core readings.* Cambridge, MA: The MIT Press, s.225–260.)

Baars, B. 1988: *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.

Barnes, J. 2001: *Early Greek Philosophy (2nd ed.)* Lontoo: Penguin.

Bechtel, W. 2001: "Representations: From Neural Systems to Cognitive Systems." (Teoksesta Bechtel et al., 2001, s.332–348.)

Bechtel, W. & Abrahamsen, A. 2002: *Connectionism and the Mind (2nd ed.): Parallel Processing, Dynamics, and Evolution in Networks.* Malden, MA: Blackwell.

Bechtel W., Abrahamsen, A. & Graham, G. 1998: "The Life of Cognitive Science." Teoksesta Bechtel, W. & Graham, G. (toim.) 1998: *A Companion to Cognitive Science.* Malden, MA: Blackwell, s.1–104.

Bechtel, W.; Mandik, P.; Mundale, J. & Stufflebeam R.S. (toim.) 2001: *Philosophy and the Neurosciences: A Reader.* Malden, MA: Blackwell.

Block, N. 1978: "Troubles with Functionalism." Teoksesta Savage, C.W. (toim.) 1978: *Minnesota Studies in the Philosophy of Science, volume IX: Perception & Cognition, Issues in the foundations of psychology.* Minneapolis, MN: University of Minnesota Press, s.261–325.

Block, N. 1980a (toim.): *Readings in Philosophy of Psychology, volume 1.* Lontoo: Methuen.

Block, N. 1980b: "What is Functionalism?" (Teoksesta Block, 1980a, s.171–184.)

Block, N. 1986: "Advertisement for a Semantics for Psychology." *Midwets Studies in Philosophy*, 10, 1, s.615–678. (Teoksesta Stich, S. & Warfield, T. (toim.) 1994: *Mental Representation: A Reader.* Cambridge, MA: Blacwell, s.81–141.)

Block, N. 1995: "An argument for holism." *Proceedings of the Aristotelian Society*, 95, s.151-69.

Block, N. & Fodor, J. 1972: "What Psychological States are Not." *The Philosophical Review*, 81, 2, s.159–181.

Boden, M.A. (toim.) 1990: *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

Boden, M.A. 1991: "Horses of a Different Color?" (Teoksesta Ramsey et al., 1991, s.3–19.)

Boden, M.A. 2006: *Mind as Machine: A history of cognitive science.* Oxford: Oxford University Press.

Boring, E.G. 1933: *The Physical Dimensions of Consciousness.* New York, NY: The Century Co.

Brooks, R.A. 1991: "Intelligence Without Representation." *Artificial Intelligence*, 47, s.139–160. (Teoksesta Brooks, R.A. 1999: *Cambrian Intelligence: The Early History of the New AI.* Cambridge, MA: The MIT Press, s.79–101.)

Brush, S.G. 1976: "Statistical Mechanics and the Philosophy of Science: Some Historical Notes." *Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol. 2: Symposia and Invited Papers.* s.551–584.

Carnap, R. 1956: "The Methodological Character of Theoretical Concepts." (Teoksesta Feigl & Scriven, 1956, s.38–76.)

Ceruzzi, P.E. 2003: *A History of Modern Computing.* Cambridge, MA: The MIT Press.

Chomsky, N. 1956: *Three Models for the Description of Language. IRE Transactions on Information Theory*, 2, 3, 113–124.

Chomsky, N. 1957: *Syntactic Structures.* Haag: Mouton.

Chomsky, N. 1959: "Review of Verbal Behavior by B. F. Skinner." *Language*, 35, 1, s.26–58. (Arvostelu on itse asiassa otsikoimaton, mutta kulkee kirjallisuudessa muun muassa mainitulla nimellä.)

Chomsky, N. 1965: *Aspects of the Theory of Syntax.* Cambridge, MA: The MIT Press.

Chomsky, N. 2006: *Language and Mind (3rd. ed.)* Cambridge: Cambridge University Press. (Teoksen ensimmäinen laitos julkaistiin vuonna 1968.)

Chrisley, R.L. 1990: "Cognitive Map Construction and Use: A Parallel Distributed Processing Approach." Teoksesta Touretzky, D.S.; Elman, J.L.; Hinton, G.E., & Sejnowski, T.J. (toim.) *Connectionist Models: Proceedings of the 1990 Summer School.* San Mateo, CA: Morgan Kaufman, s.287–302.

Chrisley, R.L. & Holland, A. 1995: "Connectionist Synthetic Epistemology: Requirements for the Development of Objectivity." Teoksesta Niklasson, L. & Boden, M. (toim.) 1995: *Current Trends in Connectionism: Proceedings of the 1995 Swedish Conference on Connectionism.* Hillsdale, NJ: Lawrence Erlbaum, s.283–309.

Church, A. 1936a: "An Unsolvable Problem of Elementary Number Theory." *American Journal of Mathematics*, 58, 2, s.345–363.

Church, A. 1936b: "A Note on the Entscheidungsproblem." *The Journal of Symbolic Logic*, 1, 1, s.40–41.

Churchland, P.M. 1981: "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy*, 78, 2, s.67–90. (Teoksesta Churchland, 1989, s.1–22.)

Churchland, P.M. 1985: "Reduction, Qualia, and the Direct Introspection of Brain States." *Journal of Philosophy*, 82, 1, s.8–28. (Teoksesta Churchland, 1989, s.47–66.)

Churchland, P.M. 1988: "Folk Psychology and the Explanation of Human Behaviour." *Proceedings of the Aristotelian Society*, 62, s.209–221. (Teoksesta Churchland, 1989, s.111–128)

Churchland, P.M. 1989: *A Neurocomputational Perspective.* Cambridge, MA: The MIT Press.

Churchland, P.M. 2005: "Functionalism at Forty: A Critical Retrospective." *Journal of Philosophy*, 102, 1, 33–50. (Teoksesta Churchland, P.M. 2007: *Neurophilosophy at Work.* Cambridge: Cambridge University Press, s.18–36.)

Clark, A. 1989: *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* Cambridge, MA: The MIT Press.

Clark, A. 2008: *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.* Oxford: Oxford University Press.

Clark, A. 1986: "Psychofunctionalism and Chauvinism." *Philosophy of Science*, 53, 4, s.535–559.

Cohen, I.B. 1988: "Babbage and Aiken." *IEEE Annals of the history of Computing*, 10, 3, s.171–193.

Cooper, J. 2007: *Cognitive Dissonance: Fifty Years of a Classic Theory.* Los Angeles, CA: SAGE Publications.

Copeland, J. 2004: *The Essential Turing: The Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life* plus *The Secrets of Enigma.* Oxford: Oxford University Press.

Copleston, S.J. 1959: *A History of Philosophy, Volume V: The British Philosophers from Hobbes to Hume.* New York: Doubleday.

Cummins, R. 1975: "Functional Analysis." *The Journal of Philosophy*, 72, 20, s.741–765.

Cussins, A. 1990: "Connectionist Construction of Concepts." (Teoksesta Boden, 1990, s.368–440.)

Cutland, N.J. 1980: *Computability: An introduction to recursive function theory.* Cambridge: Cambridge Univeristy Press.

Damasio, A. 2006: *Descartes' Error (revised edition).* Lontoo: Vintage Books. (Teoksen ensimmäinen laitos julkaistiin vuonna 1994, New York, NY: G.P. Putnam's Sons.)

Darrach, B. 1970: "Meet Shaky, the first electronic person." *Life*, 69, 24, s.58B–68.

Davidson, D. 2001: "What Thought Requires." Teoksesta Branquinho, J. (toim.) 2001: *The Foundations of Cognitive Science.* Oxford: Oxford University Press, s.121–132.

Davies, M. 1991: "Concepts, Connectionism, and the Language of Thought." (Teoksesta Ramsey et al., 1991, s.229–257.)

Dennett, D.C. 1971: "Intentional Systems." *The Journal of Philosophy*, 68, 4, s.87–106.

Dennett, D.C. 1975: "Why the Law of Effect Will Not Go Away." *Journal for the Theory of Social Behavior*, 5, 2, s.169–187. (Teoksesta Dennett, 1978c, 71–89.)

Dennett, D.C. 1978b: "Skinner Skinned." (Teoksesta Dennett, 1978c, s.53–70.)

Dennett, D.C. 1978c: *Brainstorms: Philosophical Essays on Mind and Psychology.* Montgomery, VT: Bradford Books.

Dennett, D.C. 1981: "Three Kinds of Intentional Psychology." (Teoksesta Dennett, 1987, s.41–68.) Alunperin artikkeli on julkaistu teoksessa Healy, R. (toim.) 1981: *Reduction, Time and Reality.* Cambridge: Campridge University Press.

Dennett, D.C. 1982: "Beyond Belief." (Teoksesta Dennett, 1987, s.117–202) Alunperin artikkeli on julkaistu teoksessa Woodfield, A. (toim.) 1982: *Thought and Object.* Oxford: Clarendon Press.

Dennett, D.C. 1987: *The Intentional Stance.* Cambridge, MA: The MIT Press.

Dennett, D.C. 1991: *Consciousness Explained.* Boston, MA: Little, Brown & Company.

Descartes, R. 2001: *Teokset I: Yksityisiä ajatelmia, Järjen käyttöohjeet, Metodin esitys, Optiikka ja Kirjeitä 1619–1640.* Suom. Sami Jansson, Helsinki: Gaudeamus. (Viitattu kohta on vuonna 1637 julkaistusta ranskankielisestä teoksesta *Discours de la méthode pour bien conduire sa raison, et chercher la vérité dans les sciences.*)

Descartes, R. 2002: *Teokset II: Mietiskelyjä ensimmäisestä filosofiasta – Kirjeitä 1640–1641.* Suom. Tuomo Aho ja Mikko Yrjönsuuri, Helsinki: Gaudeamus. (Viitattu kohta on vuonna 1641 julkaistusta latinankielisestä teoksesta *Meditationes de Prima Philosophia in qua Dei existentia et animae immortalitas demonstratur.*)

Dretkse, F. 1981: *Knowledge and the Flow of Information.* Cambridge, MA: The MIT Press.

Dreyfus, H.L. 1965: *Alchemy and Artificial Intelligence.* Santa Monica, CA: The RAND Corporation.

Dreyfus, H.L. 1992: *What Computers Still Can't Do: A Critique of artificial reason.* Cambridge, MA: The MIT Press.

Dreyfus, H.L. & Dreyfus, S.E. 1986: *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* New York, NY: The Free Press.

Ernst, G.W. & Newell, A. 1969: *GPS: A Case Study in Generality and Problem Solving.* New York, NY: Academic Press.

Evans, J.St.B.T. 2003: "In two minds: dual-processing accounts of reasoning." *Trends in Cognitive Sciences*, 7, 10, s.454–459.

Evans, J.St.B.T.; Newstead, S.E. & Byrne, R.M.J. 1993: *Human Reasoning: The Psychology of Deduction.* Hove: Lawrence Erlbaum Associates.

Eysenck, M.W. & Keane, M. 2000: *Cognitive Psychology: A Student's Handbook (4th ed.)* Hove: Psychology Press.

Feigl, H. 1958: "The 'Mental' and the 'Physical'." (Teoksesta Feigl et al., 1958, s.370–497.)

Feigl, H. & Scriven, M. (toim.) 1956: *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis.* Minneapolis, MN: University of Minneosta Press.

Feigl, H.; Scriven, M. & Maxwell, G. 1958 (toim.): *Minnesota Studies in the Philosophy of Science, Volume II: Concepts, Theories, and the Mind-Body Problem.* Minneapolis, MN: University of Minneapolis Press.

Ferreirós, J. 1999: *Labyrinth of Thought: A History of Set Theory and Its Role in Modern Mathematics.* Berliini: Springer.

Field, H.H. 1978: "Mental Representation." *Erkenntnis*, 13, 1, s.9–61. (Teoksesta Block, N. (toim.) 1981: *Readings in Philosophy of Psychology, vol. 2.* Lontoo: Methuen, s.78–115.)

Fodor, J.A. 1968a: *Psychological Explanation: An Introduction to the Philosophy of Psychology.* New York, NY: Random House.

Fodor, J.A. 1968b: "The Appeal to Tacit Knowledge in Psychological Explanation." *The Journal of Philosophy*, 65, 20, s.627–640.

Fodor, J.A. 1974: "Special Sciences. (Or: the disunity of science as a working hypothesis.)" *Synthese*, 28, s.97–115.

Fodor, J.A. 1975: *The Language of Thought.* Cambridge, MA: Harvard University Press.

Fodor, J.A. 1981: "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology." *Beahvioral and Brain Sciences*, 3, s.63–73. (Teoksesta Fodor, J.A. 1981: *Representations: Philosophical Essays on the Foundations of Cognitive Science.* Cambridge, MA: The MIT Press, s.225–253.)

Fodor, J.A. 1983: *The Modularity of Mind.* Cambridge, MA: The MIT Press.

Fodor, J.A. 1987: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, MA: The MIT Press.

Fodor, J.A. 1990: *A Theory of Content: And Other Essays.* Cambridge, MA: The MIT Press.

Fodor, J.A. 1992: "A Theory of the Child's Theory of Mind." *Cognition* 44, 3, s.283–296.

Fodor, J.A. 1997: "Special Sciences: Still Autonomous After All These Years." *Nôus* 31, Supplement: Philosophical Perspectives 11, s.149–163.

Fodor, J.A. 2000: *The Mind Doesn't Work That Way.* Cambridge, MA: The MIT Press.

Fodor, J.A. 2008: *LOT 2: The Language of Thought Revisited.* Oxford: Clarendon.

Fodor, J.A. & Pylyshyn, Z. 1988: "Connectionism and cognitive architecture: A critical analysis." *Cognition* 28, 1–2, s.3–71.

Frege, G. 1970: *Begriffsschrift (Chapter 1).* Teoksesta Geach, P. & Black, M. (toim.) 1970: *Translations from the Philosophical Writings of Gottlob Frege.* Malden, MA: Blackwell, s.1–20. Englanniksi kääntänyt Peter Geach. Saksankielinen alkuteos *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* on alunperin julkasitu Hallessa 1879.

Gillett, G. 1997: "Husserl, Wittgenstein and the Snark: Intentionality and Social Naturalism." *Philosophy and Phenomenological Research*, 57, 2, s.331–349.

Godfrey, M.D. & Hendry, D.F. 1993: "The Computer as von Neumann Planned It." *IEEE Annals of the History of Computing*, 15, 1, s.11–21.

Goldberg, R.P. 1974: "Survey of Virtual Machine Research." *Computer*, 7, 6, s.34–45.

Gregory, R. L. (toim.) 1987: *The Oxford Companion to the Mind.* Oxford: Oxford University Press.

Gödel, K. 1951: "Some basic theorems on the foundations of mathematics and their implications." Teoksesta Gödel, K. 1995: *Collected Works, vol. III: Unpublished Essays and Lectures.* Oxford: Oxford University Press, s.304–323.

Hardcastle, V. 2001: "The Nature of Pain." (Bechtel et al., 2001, s.295–311.)

Harman, G. 1987: "(Nonsolipsistic) Conceptual Role Semantics." (Teoksesta Harman, G. 1999: *Reasoning, Meaning, and Mind.* Oxford: Oxford University Press, s.206–231.) Alunperin artikkeli on julkaistu teoksessa Lepore, E. (toim.) 1987: *New Directions in Semantics.* Lontoo: Academic Press.

Haugeland, J. 1978: "The Nature and Plausibility of Cognitivism." *Behavioral and Brain Sciences* 1, 2, s.215–226. (Teoksesta Haugeland, 1981b, s.243–281.)

Haugeland, J. 1981a: "Semantic Engines: An Introduction to Mind Design." (Teoksesta Haugeland, 1981b, s.1–34.)

Haugeland, J. (toim.) 1981b: *Mind Design.* Cambridge, MA: The MIT Press.

Haugeland, J. 1985: *Artificial Intelligence: The Very Idea.* Cambridge, MA: The MIT Press.

Hayes, P.J. 1979: "The Naïve Physics Manifesto." (Teoksesta Boden, 1990, s.171–205.) Alunperin artikkeli on julkaistu teoksessa Michie, D. (toim.) 1979: *Expert Systems in the Microelectronic Age.* Edinburgh: Edinburgh University Press, s.242–270.

Heil, J. (toim) 2004: *Philosophy of Mind: A Guide and Anthology.* Oxford: Oxford University Press.

Hempel, C.G. 1935: "The Logical Analysis of Psychology." (Teoksesta Heil, 2004, s.85–95.) Englanniksi kääntänyt Wilfrid Sellars. Alunperin artikkeli on ilmestynyt ranskaksi kausijulkaisussa *Revues de synthèse* 1935.

Hempel, C.G. 1966: *Philosophy of Natural Science.* Englewood Cliffs, NJ: Prentice-Hall.

Hobbes, T. 1651: *Leviathan, or The Matter, Forme and Power of a Common Wealth Ecclesiasticall and Civil.* Lähteenä on käytetty G.A.J. Rogersin ja K. Schumannin toimittamaa, vuonna 2003 julkaistua kriittistä laitosta, Lontoo: Thoemmes Continuum.

Hopcroft, J.E.; Motwani, R. & Ullman, J.D. 2001: *Introduction to Automata Theory, Languages, and Computation (2nd. ed.)* Upper Saddle River, NJ: Pearson.

James, W. 1890: *The Principles of Psychology.* New York, NY: Henry Holt and Company.

Kim, J. 1989: "The Myth of the Nonreductive Materialism." *Proceedings and Addresses of the American Philosophical Association*, 63, 3, s.31–47.

Kim, J. 1992: "Realization and the Metaphysics of Reduction." *Philosophy and Phenomenological Research*, 52, 1, s.1–26.

Kim, J. 2006: *Philosophy of Mind (2nd ed.)* Cambridge, MA: Westview.

King, P. 2005: "William of Ockham: Summa Logicae." Teoksesta Shand, J. (toim.) 2005: *Central Works of Philosophy, vol. 1: Ancient and Medieval.* Montreal: McGill-Queen's University Press, s.242–270.

Kneale, W. & Kneale, M. 1962: *The Development of Logic.* Oxford: Oxford Unversity Press.

Kosslyn, S.M. & Smith, E.E. 2000: *Chapter VIII, Higher Cognitive Functions: Introduction.* Teoksesta Gazzaniga, M.S. (toim.) 2000: *The New Cognitive Neurosciences.* Cambridge, MA: The MIT Press, s.961–963.

Knuth, D.E. 1970: "Von Neumann's First Computer Program." *Computing Surveys*, 2, 4, s.247–260.

de La Mettrie, J.O. 2003: *Ihmiskone.* Tampere: Eurooppalaisen filosofian seura. Suomentanut Tapani Kilpeläinen. Ranskankielinen alkuteos *L'Homme machine* julkaistiin vuonna 1747.

Lakoff, G. & Johnson, M. 1999: *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought.* New York, NY: Basic Books.

Lenat, D.B. & Feigenbaum, E.A. 1987: "On the Tresholds of Knowledge." *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, s.1173–1182.

Lewis, D. 1966: "An Argument for the Identity Theory." *The Journal of Philosophy*, 63, 1, s.17–25.

Lewis, D. 1972: "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, 50, s.249–258. (Teoksesta Block, 1980a, s.207–215.)

Lifschitz, V. (toim.) 1990: *Formalizing Common Sense: Papers by John McCarthy.* Norwood, NJ: Ablex Publishing Corporation.

Lucas, J.R. 1961: "Minds, Machines, and Gödel." *Philosophy*, 36, s.112–127.

Lucas, J.R. 1996: "Minds, Machines, and Gödel: A Retrospect." Teoksesta Millican, P. & Clark, A. 1996 (toim.): *Machines and Thought: The Legacy of Alan Turing, vol. I.* Oxford: Oxford University Press, s.103–124.

Luria, A.R. 1976: *Cognitive Development: Its Cultural and Social Foundations.* Harvard: Harvard University Press. Englanniksi kääntäneet Martin Lopez-Morillas ja Lynn Solotaroff. Alunperin teos on julkaistu venäjäksi vuonna 1974.

Lycan, W.G. 1981: "Form, Function, and Feel." *The Journal of Philosophy*, 78, 1, s.24–50.

Mander, J.M. 2004: *The Foundations of Mind.* Oxford: Oxford University Press.

Marcus, G.F. 1998: "Rethinking Eliminative Connectionism." *Cognitive Psychology*, 37, s.243–282.

Marr, D. 1982: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York: W. F. Freeman and Company.

McCarthy, J. 1959: "Programs with Common Sense." *Proceedings of the Teddington Conference on the Mechanization of Thought Processes.* Lontoo: Her Majesty's Stationery Office, s.75–91. (Teoksesta Lifschitz, 1990, s.9–20.)

McCarthy, J. 1986: "Applications of Circumscription to Formalizing Common Sense Knowledge." *Artificial Intelligence*, 28, 1, s.89–116. (Teoksesta Lifschitz, 1990, s.198–225.)

McCarthy, J. 2005: "The Future of AI – A Manifesto." *AI Magazine*, 26, 4, s.39.

McCarthy, J. & Hayes, P. 1969: "Some Philosophical Problems from the Standpoint of Artificial Intelligence." (Teoksesta Lifschitz, 1990, s.21–63.) Alunperin julkaistu teoksessa: Meltzer, B. & Michie, D. (toim.) 1969: *Machine Intelligence 4*, Edinburgh: Edinburgh University Press, s.463–502.

McCauley, R.N. 1996: "Explanatory Pluralism and the Co-evolution in theories of Science." (Teoksesta Bechtel et al., 2001, s.429–456) Aluperin julkaistu teoksessa Robert McCaulry (toim.) 1996: *The Churchlands and Their Critics.* Oxford: Blackwell, s.17–47.

McClelland, J.L.; Rumelhart, D.E. & The PDP Research Group 1986a: *Parallel Distributed Processing, vol. 1: Foundations.* Cambridge, MA: The MIT Press.

Rumelhart, D.E.; Smolensky, P.; McClelland, J.L. & Hinton, G.E. 1986: "Schemata and Sequential Thought Processes in PDP Models." Teoksesta McClelland, J.L.; Rumelhart, D.E. & The PDP Research Group 1986: *Parallel Distributed Processing, vol. 2: Psychological and Biological Models.* Cambridge, MA: The MIT Press, s.7–57.

McCulloch, W.S. & Pitts, W. 1943: "A logical calculus of the ideas immanent in nervous activity." *Bulletin of Mathematical Biology*, 5, 4, s.115–133. (Teoksesta Boden, 1990, s.22-39.)

McGhan, H. & O'Connor, M 1998: "PicoJava: A Direct Execution Engine for Java Bytecode." *Computer* 31, 10, s.22–30.

Miller, G.A. 1956: "The Magical Number Seven, Plus or Minus Two: Some limits on our capacity of processing information." *Psychological Review*, 63, 2, s.81–97.

Minsky, M. 1974: "A Framework for Representing Knowledge." MIT-AI Laboratory Memo 306. (Viitatut kohdat ovat lyhennelmästä, joka on julkaistu teoksessa Haugeland, 1981b, s.95–128.)

Minsky, M. 1985: *The Society of Mind.* New York, NY: Simon and Schuster.

Moraveck, H. 1988: Mind Children: The Future of Robot and Human Intelligence. Cambridge, MA: Harvard University Press.

Nagel, E. 1961: *The Structure of Science.* Lontoo: Routledge.

Nagel, E. & Newman, J.R. 2001: *Gödel's Proof (revised edition.)* New York, NY: New York University Press.

Neisser, U. 1967: *Cognitive Psychology.* Englewood Cliffs, NJ: Prentice-Hall.

von Neumann, J. 1945: "First Draft of a Report on the EDVAC." *IEEE Annals of the History of Computing*, 15, 4 (1993), s.28–77.

Newell, A. 1962: *Some Problems of Basic Organization in Problem-Solving Programs.* Santa Monica, CA: The RAND Corporation.

Newell, A. 1980: "Physical Symbol Systems." *Cognitive Science*, 4, 2, s.135–183.

Newell, A. & Simon, H.A. 1956: *The Logic Theory Machine: A Complex Information Processing System.* Santa Monica, CA: The RAND Corporation.

Newell, A. & Simon, H.A. 1961: "GPS, A Program that Simulates Human Thought." (Teoksesta Feigenbaum, E.A. & Feldman, J. (toim.) 1963: *Computers and Thought.* New York, NY: McGraw-Hill, s.279–293.) Alunperin artikkeli on julkaistu teoksessa Billing, H. (toim.) 1961 : *Lernende Automaten.* München, Oldenbourg, s.109–124.

Newell, A. & Simon, H.A. 1972: *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

Newell, A. & Simon, H.A. 1976: "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the ACM*, 19, 3, s.113–126.

Nisbett, R.E. & Wilson, T.D. 1977: "Telling more than we can know: Verbal reports on mental processes." *Psychological Review*, 84, 3, s.231–259.

Oppenheim, P. & Putnam, H. 1958: "The Unity of Science as a Working Hypothesis." (Teoksesta Feigl et al., 1958, s.3–35.)

Penrose, R. 1994: *The Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford: Oxford University Press.

Picard, R.W. 1997: *Affective Computing.* Cambridge, MA: The MIT Press.

Piccinini, G. 2004: "The First Computational Theory of Mind and Brain: A close look at McCulloch and Pitts's "Logical calculus of ideas immanent in nervous activity." *Synthese*, 141, s.175–215.

Piccinini, G. 2007: "Computational Modelling vs. Computational Explanation: Is everything a Turing machine, and does it matter to the philosophy of mind?" *Australasian Journal of Philosophy*, 85, 1, s.93–115.

Pinker, S. 1997: *How the Mind Works.* Lontoo: Penguin Book.

Place, U.T. 1956: "Is Consciousness a Brain Process?" *British Journal of Psychology,* 47, 1, s.44–50.

Popek, G.J. & Goldberg, R.P. 1974: "Formal Requirements for Virtualizable Third Generation Architectures." *Communications of the ACM*, 17, 7, s.412–421.

Putnam, H. 1960: "Minds and Machines." Teoksesta Hook, S. 1960 (toim.): *Dimensions of Mind: A Symposium.* New York: New York University Press, s.148–179.

Putnam, H. 1967a: "Psychological Predicates." (Teoksesta Heil, 2004, s.158–167.) Alunperin artikkeli on julkaistu teoksessa Capitan, W.H. & Merrill, D.D. (toim.) 1967:*Art, Mind and Religion.* Pittsbugh: University of Pittsbugh Press, s.37–48.

Putnam, H. 1967b: "Mental Life of Some Machines." (Teoksesta Putnam, 1975c, s.408–428.) Alunperin artikkeli on julkaistu teoksessa Castañeda, H. (toim.) 1967: *Intentionality, Minds and Perception.* Detroit, MI: Wayne State University Press, s.177–200.

Putnam, H. 1975a: "Philosophy and Our Mental Life." (Teoksesta Putnam, 1975c, s.291–303)

Putnam, H. 1975b: "The Meaning of "Meaning"." Teoksesta Gunderson, G. (toim) 1975: *Minnesota Studies in the Philosophy of Science, Volume VII: Language, Mind, and Knowledge.* Minneapolis, MN: University of Minnesota Press, s.131–193.

Putnam, H. 1975c: *Philosophical Papers, vol. 2: Mind, Language, and Reality.* Cambridge: Cambridge University Press.

Pylyshyn, Z.W. 1984: *Computation and Cognition.* Cambridge, MA: The MIT Press.

Ramsey, W.; Stich, S.P. & Rumelhart, D.E. (toim.) 1991: *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rapaport, W.J. 1995: "Understanding Understanding: Syntactic Semantics and Computational Cognition." *Philosophical Perspectives, vol. 9: AI, Connectionism and Philosophical Psychology*, s.49–88.

Rosch, E. 1978: "Principles of Categorization." Teoksesta Rosch, E. & Lloyd, B.B. 1978: *Cognition and Categorization.* Hillsdale, NJ: Lawrence Erlbaum, s.27–48.

Rowlands, M. 1994: "Connectionism and the Language of Thought." *The British Journal for the Philosophy of Science*, 45, 2, s.485–503.

Ryle, G. 1949: *The Concept of Mind.* Lontoo: Hutchinson's University Library.

Schank, R.C. & Abelson, R.P. 1977: *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwitzgebel, E. 2002: "How Well Do We Know Our Own Conscious Exprerience? The Case of Visual Imagery." *Journal of Consciousness Studies*, 9, 5, 35–53.

Searle, J. 1980: ”Minds, Brains, and Programs.” *Behavioral and Brain Sciences*, 3, s.417–457.

Searle, J. 1984: *Minds, Brains and Science: The 1984 Reith lectures.* Lontoo: BBC.

Scriven, M. 1956: ”A Study of Radical Behaviorism.” (Feigl & Scriven, 1956, s.88–130)

Sellars, W. 1956: ”Empiricism and the Philosophy of Mind.” (Feigl & Scriven, 1956, s.253–329.)

Shoemaker, S. 1981: ”Some varieties of functionalism.” *Philosophical Topics*, 12, 2, s.93–119. (Teoksesta Shoemaker, S. 2003: *Identity, Cause, and Mind: Philosophical Essays (Expanded ed.)* Oxford: Oxford University Press, s.261–286.)

Shurkin, J. 1996: *Engines of the Mind: The Evolution of the Computer from Mainframes to Microprocessors.* New York. NY: W. W. Norton & Company.

Siegelmann, H.T. & Sontag, E.D. 1991: ”Turing Computability With Neural Nets.” *Applied Mathematics Letters*, 4, 6, s.77–80.

Simon, H.A. & Newell, A. 1958: ”Heuristic Problem Solving: The next Advance in Operations Research.” *Operations Research* 6, 1, s.1–10.

Skinner, B.F. 1945: ”The Operational Analysis of Psychological Terms.” *Psychological Review*, 52, s.270–277.

Skinner, B.F. 1965: *Science and Human Behavior (New impression edition).* New York, NY: Free Press. Teoksen ensimmäinen laitos julkaistiin vuonna 1953.

Smart, J.J.C. 1959: ”Sensations and Brain Processes.” *The Philosophical Review*, 68, 2, s.141–156.

Smith, J.E. & Nair, R. 2005: ”The Architecture of Virtual Machines.” *Computer*, 38, 5, s.32–38.

Smolensky, P. 1988: ”On the Proper Treatment of Connectionism.” *Behavioral and Brain Sciences*, 11, s.1–74. (Teoksesta Cole, D.J.; Fetzer, J.H. & Rankin, T.R. (toim.) 1990: *Philosophy, Mind, and Cognitive Inquiry.* Dordrecht: Kluwer, s.145–206.)

Sternberg, S. 1966: ”High-Speed Scanning in Human Memory.” *Science*, 153, s.652–654.

Sternberg, S. 1969: ”Memory-Scanning: Mental processes revealed by reaction-time experiments.” *American Scientist* 57, 4, s.421–457.

Stich, S.P. 1983: *From Folk Pschology to Cognitive Science: The Case Against Belief.* Cambridge, MA: The MIT Press.

Stich, S.P. 1996: *Deconstructing the Mind.* Oxford: Oxford University Press.

Stich, S.P. & Nichols, S. 1992: ”Folk Psychology: Simulation or Tacit Theory?” *Mind & Language*, 7, 1–2, s.35–71.

Tarski, A. 1956a: ”The Concept of Truth in Formalized Languages.” (Teoksesta Tarski, 1988, s.152–278.) Englanniksi kääntänyt J.H. Woodger 1956. Alunperin ”Pojęcie prawdy w językach nauk dedukcyjnych” on julkaistu kausijulaisussa *Travaux de la Société des Sciences et des letters de Varsovie* vuonna 1933.

Tarski, A. 1956b: ”On the Concept of Logical Consequence.” (Teoksesta Tarski, 1988, s.409–420.) Englanniksi kääntänyt J.H. Woodger 1956. Alunperin artikkeli on julkaistu Puolaksi ja Saksaksi vuonna 1936.

Tarski, A. 1944: "The Semantic Conception of Truth: and the Foundations of Semantics." *Philosophical and Phenomenological Research*, 4, 3, s.341–376.

Tarski, A. 1988: *Logic, Semantics, Metamathematics (2nd ed.).* Indianapolis: Hackett. Teoksen ensimmäinen laitos julkaistiin vuonna 1956.

Tarski, A. & Vaught, R.L. 1956–1958: "Arithmetical extensions of relational systems." *Compositio Mathematica*, 13, s.81-102.

Turing, A.M. 1936: "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of London Mathematical Society*, 2, 42, s.230–265.

Turing, A.M. 1946: "Proposed Electronic Calculator." (Teoksesta Copeland, J. 2005: *Alan Turing's Automatic Computing Engine* Oxford: Oxford University Press, s.369–454.)

Turing, A.M. 1950: "Computing Machinery and Intelligence." *Mind*, 59, 236, s.433–460.

Turing, A.M. 1951: "Intelligent Machinery, A Heretical Theory." (Teoksesta Copeland, 2004, s.465–475.)

van Gelder, T. 1995: "What Might Cognition Be, If Not Computation?" *The Journal of Philosophy*, 92, 7, s.345–381.

Vilkko, R. 2005: "Mitä uutta Frege löysi?" *Niin & näin*, 45, 2, s.27–30.

Vygotski, L. 1982: *Ajattelu ja kieli.* Espoo: Weilin+Göös. Suomentaneet Klaus Helkama ja Anja Koski-Jännes. Venäjänkielinen alkuteos ilmestyi vuonna 1931.

Watson, J.B. 1913: "Psychology as the Behaviorist Views it." *Psychological Review*, 20, 158–177.

Watson, J.B. 1930: *Behaviorism (Revised edition).* Chicago, IL: Univerity of Chicago Press. Teoksen ensimmäinen laitos julkaistiin vuonna 1924.

Wilson, T.D. 2002: *Strangers to Ourselves: Discovering the Adaptive Unconsciousness.* Cambridge, MA: Harvard University Press.

Winograd, T. 1971: *Procedures as a representation for data in a computer program for understanding natural language.* MIT AI Technical Report 235. (Tutkielma on julkaistu myös kausijulkaisun *Cognitive Psychology* vuoden 1972 numerona vol.3, no.1, ja samana vuonna kirjana nimellä *Understanding Natural Language*, Lontoo: Academic Press.)

von Wright, G.H. 1968: *Logiikka, filosofia ja kieli (2. laitos).* Helsinki: Otava. Suomentaneet Jaakko Hintikka ja Tauno Nyberg. Teoksen ensimmäinen ruotsinkielinen laitos *Logik, filosofi och språk* julkaistiin vuonna 1957.

Wulf, G. 2007: *Attention and motor skill learning.* Champaign, IL: Human Kinetics.

Zajonc, R.B. 1968: "Attitudinal Effects of Mere Exposure." *Journal of Personality and Social Psychology*, 9, 2, s.1–27.

Zajonc, R.B. 1980: "Feeling and Thinking: Preferences need no Inferences." *American Psychologist*, 35, 2, s.151–175.