MAAILMANKATSOMUSTEN RAKENTAMINEN JA ARVIOINTI SEKÄ ÄLYKKÄÄT SYSTEEMIT - MAAILMANKATSOMUS

Mikko Virtala

Pro gradu-tutkielma

Filosofia

Yhteiskuntatieteiden ja filosofian laitos

Jyväskylän Yliopisto

Kevät 2009

Ohjaajat: Mikko Yrjönsuuri ja Jussi Kotkavirta

Tiivistelmä

Tutkielman tavoitteena on muodostaa mahdollisimman hyvä maailmankatsomus. Tätä varten määritellään maailmankatsomuksen käsite, esitetään syitä miksi maailmankatsomuksia tarvitaan, esitetään kuinka maailmankatsomuksia rakennetaan, ja kuinka maailmankatsomuksia arvioidaan. Tämän jälkeen esitetään mahdollisimman hyvä maailmankatsomus. Maailmankatsomus on kokonaiskäsitys sen omaajan ympäristöstä, tavoitteista ja keinoista tavoitteiden saavuttamiseksi. Jokaisella elävällä oliolla - ja joillakin elottomilla kuten roboteilla - on maailmankatsomus, vähintään implisiittisenä ja tiedostamattomana. Tässä mielessä maailmankatsomus on niille välttämätön; ilman sitä ne eivät voisi toimia. Ihmiset voivat rakentaa, arvioida ja muuttaa tai vaihtaa maailmankatsomuksiaan. Ihmisillä on sekä tarve, että mahdollisuus valita parempi maailmankatsomus. Esitän kriteeristön, joiden avulla niitä voidaan arvioida, vertailla ja asettaa paremmuusjärjestykseen lähinnä Vidalin (2007) ja Apostelin ja kumppaneiden (2007) pohjalta. Nykyään on tarjolla erilaisia maailmankatsomuksia, tieteellisiä, uskonnollisia ja metafyysisiä, joista toiset vievät ääritapauksissa ihmiset ja ihmiskunnan tuhoon ja toiset menestykseen. Esitän ehdotuksen nykyiseksi maailmankatsomukseksi ihmisille. Tämä älykkäät systeemit -maailmankatsomus suureksi osaksi perustuu Walter Fritzin (2007) kirjassaan *Intelligent systems and their societies* esittämiin ajatuksiin. Pidän sitä hyvänä ehdokkaana maailmankatsomusten arviointikriteerien valossa. Se sisältää maailmankatsomusten välttämättömät osat, on yksinkertainen, lyhyt, sisäisesti ja ulkoisesti koherentti sekä kehityskelpoinen ja muokattavissa.

Sanastoa

Älykäs systeemi (engl. Intelligent system); Maailmankatsomus; Älykkyys; Rationaalisuus; Systeemiteoria; Tekoäly.

Sisällys

1 JOHDANTO

Työn motiivina on löytää mahdollisimman hyvä maailmankatsomus, joka yhdistäisi toimivaksi kokonaisuudeksi kaiken parhaan tiedon, ja olisi muuttuvien tarpeiden ja uuden tiedon perusteella edelleen kehitettävissä. Mahdollisimman hyvän maailmankatsomuksen löytäminen on filosofiassa yleinen tavoite, ja monien mielestä sen päätehtävä. Alun perin filosofia oli kaikenkattava yritys ymmärtää itseämme ja ympäristöämme. Myöhemmin tutkimus jakaantui erityistieteisiin. (Fritz 2007, The history of philosophy.) Filosofit Platonista Epikurokseen, Spinozasta Hegeliin ja Whiteheadista Deleuzeen ovat luoneet laajoja käsiterakennelmia, joiden avulla itseä, muita ja maailmaa voitaisiin ymmärtää yhtenäisellä ja hyödyllisellä tavalla. Näitä voidaan pitää maailmankatsomuksina.

Ensimmäiset tunnetut maailmankatsomukset ovat olleet biologisilla organismeilla. Ympäristön mallin rakentamisen välttämättömyys johtuu elävän organismin tarpeesta sopeutua siihen. Se johtaa myös maailmankatsomuksen ja sen kehittämisen tarpeisiin (Vidal 2007, 9). Tämän mallin rakenteeseen vaikuttavat sen tekijän tarpeet ja kyvyt (Aerts, Apostel ja muut 2007, 18-19.) sekä ympäristö. Myös mm. robotit ja ohjelma-agentit tarvitsevat maailmankatsomuksia, ja niiden rakentamisesta ja ohjelmoinnista saadut kokemukset ovat hyödyllisiä maailmankatsomusten tutkimukselle, koska niitä voidaan muutella, niiden osat tunnetaan ja rakenne voidaan pitää yksinkertaisena. Nykyään ihmisten kannalta on tärkeää kehittää maailmankatsomuksia laajan yhteistyön mahdollistamiseksi. Yksilöiden ja yhteisöjen elämänsuunnitelmat tai toimintasuunnitelmat edellyttävät maailmankatsomusta. Hyvien suunnitelmien tarve lisääntyy ympäristön monimutkaisuuden kasvaessa. Kulttuurienvälinen vuorovaikutus, ihmiskunnan suurempi yhdistyminen sekä tieteen ja teknologian edistyminen johtavat siihen, että yksilön elämänsuunnitelma ja aliyhteisöjen suunnitelmat ovat voimakkaammin yhteydessä globaaliin kokonaisuuteen. Elämänsuunnitelman muodostaminen on samalla tullut vaikeammaksi huomioonotettavan kokonaisuuden koon ja monimutkaisuuden lisääntymisen vuoksi. Esimerkiksi ihmiskunnan selviytymiseen tällä planeetalla liittyvät ekologiset ongelmat tulevat enenevästi olemaan jokaisen ongelma. (Aerts, Apostel et al. 2007, 8, 11.) Vastuullinen toiminta, ympäristön muuttaminen sekä yhteistyö edellyttävät maailmankatsomusta. Saadaksemme tietoa itsestämme, tarpeistamme,

päämääristämme ja arvoistamme, rakennamme malleja ympäristöstämme ja itsestämme. Meidän täytyy rakentaa malleja ihmisistä, historiasta, arvoista ja toimintastrategioista ja yhdistää kosmosta, maata ja biosfääriä koskevan tietomme kanssa. Ilman näiden käsitysten integraatiota vastuullinen toiminta näyttää olevan mahdotonta. Koska emme voi vain antaa asioiden kehittyä omalla painollaan, vaan on otettava vastuu maailmastamme, näiden elementtien uusien ja parempien integraatioiden kehittäminen on välttämätöntä. Eettisiä ja poliittisia valintoja valaiseva tieto ihmisistä, luonnosta, historiasta ja yhteiskunnista mahdollistaa kenties tulevaisuudessa ihmisten ottaa kohtalonsa omiin käsiinsä. (Aerts, Apostel et al. 2007, 11.) Tarvitaan viitekehys, joka sitoo kaiken yhteen ja mahdollistaa yhteiskunnan, maailman ja ihmisten paikan maailmassa ymmärtämisen, koska se auttaisi ihmisiä ja ihmiskuntaa tekemään tärkeitä päätöksiä, jotka muovaavat tulevaisuuttamme. (Heylighen 2000: "What is a world view?") Maailmankatsomuksen kehittäminen voi parantaa elämänhallintaa, ja mahdollistaa uusia kykyjä ja kokemuksia.

Tutkimuksen lähtökohtana on ns. agentti-ympäristö -viitekehys. Tämä laajasti käytetty ja joustava struktuuri, jota oletettavasti jokainen käyttää ja ymmärtää jossakin muodossa, vähintään intuitiivisesti, ei sinällään ole uusi keksintö. Tekoälytutkimuksessa sitä käytetään mm. vahvistusoppimisen viitekehyksenä (Sutton, R. ja Barto, A. 1998), ja eri tavalla nimettynä kontrolliteoriassa (Bertsekas, D. P. ja Tsitsiklis, J. N. 1996). Viitekehys tuli tunnetuksi, kun Norbert Wiener (1894- 1964) kollegoidensa Arturo Rosenbluethin ja Julia Bigelowin kanssa haastoivat behavioristisen ajattelutavan tarkastellen käyttäytymistä nykyisen ja tavoitetilan välistä virhettä säätelevänä systeeminä. Wienerin kirjasta Kybernetiikka (1948) tuli bestseller, joka myös kiinnitti yleisön huomion mahdollisuuteen luoda älykkäitä koneita (Russel ja Norvig 2003, 15). Viitekehys sisältää kolme olennaista komponenttia: agentin, ympäristön ja tavoitteen. Agentti on mikä tahansa, mitä voidaan tarkastella sensoreillaan ympäristöään havainnoivana ja siinä aktuaattoreillaan toimivana (Russell ja Norvig 2003, 32). Agentin ja ympäristön tulee voida olla vuorovaikutuksessa toistensa kanssa. Agentin tulee voida lähettää signaaleja ympäristöön ja ottaa vastaan ympäristöstä tulevia signaaleja ja ympäristön tulee voida saada signaaleja ja lähettää niitä agentille. Agentin ympäristöön lähettämiä signaaleja voidaan sanoa toiminnoiksi (engl. actions), ja sieltä vastaanottamia havainnoiksi (engl. perceptions). (Legg ja Hutter 2007b, 15- 16.) Agentilla täytyy olla ainakin

yksi tavoite. (Legg ja Hutter 2007b, 15- 16.) Tavoite on tietty tilanne, johon se pyrkii. (Fritz 2007, What is intelligence?, Acting on the environment, Objectives.) Agentti voi olla älykäs ilman tavoitetta, jonka saavuttamiseen se voi käyttää älykkyyttään, ja siinäkin tapauksessa, että agentti ei halua käyttää älyään tavalla joka vaikuttaa sen ympäristöön, mutta kummassakaan tapauksessa sen älykkyyttä ei voida havaita. Älykkyys voidaan havaita kun sillä on tavoite, jota se aktiivisesti ympäristöönsä vaikuttamalla yrittää saavuttaa. (Legg ja Hutter 2007b, 15- 16.) Ilmeisesti mikä tahansa peli, haaste, ongelma tai testi voidaan ilmaista tämän yksinkertaisen viitekehyksen avulla ilman suurta vaivaa. Viitekehys ei sano mitään siitä, miten agentti tai ympäristö todella toimii, vaan kuvaa niiden roolit. (Legg ja Hutter 2007b, 17.) Määritelmä esitetään formalisoidussa muodossa lähteessä: (Legg ja Hutter 2007b, 17- 24). Käytän viitekehystä, koska se on yleinen, eri aloilla menestyksellä käytetty, sekä täsmällisesti ja formaalisti määriteltävissä. Siten se on käyttökelpoinen perusta, jolle on suhteellisen turvallista rakentaa uusia käsitejärjestelmiä. Käsitykseni maailmankatsomuksista, niiden rakentamisesta ja arvioinnista sekä ehdotus nykyiseksi maailmankatsomukseksi perustuvat tähän viitekehykseen. Tutkielma on melko yleisluonteinen, ja joudun jättämään yksityiskohtaisemmat kehittelyt ja jotkin kokonaiset aihealueet myöhempiin tutkimuksiin.

Tätä viitekehystä käyttäen maailmankatsomusta voidaan pitää agentille kuuluvana informaatiojärjestelmänä, jonka tehtävänä on tuottaa oikea vaste syötteeseen. Oikea vaste määräytyy viime kädessä agentin tavoitteen, resurssien ja ympäristön mukaan. Oikea vaste on se, joka parhaiten auttaa agenttia saavuttamaan tavoitteensa. Tietty tavoite, tietyt resurssit ja tietty ympäristö muodostavat mahdollisuudet ja ehdot menestyksekkäälle käyttäytymiselle. Sopiva maailmankatsomus on funktio, jonka lähtöjoukkoon tavoite, resurssit ja ympäristö kuuluvat, ja jonka maalijoukkoon kuuluu menestyksekäs käyttäytyminen. Tässä merkityksessä kaikki agentit tarvitsevat maailmankatsomuksen vuorovaikutukseen ympäristönsä kanssa, ja jokainen agentti tahtoo mahdollisimman hyvän sellaisen.

Johdannon jälkeen

- Määritellään maailmankatsomuksen käsite (kappale 2.1 Maailmankatsomukset). Se on hyödyllistä määritellä mahdollisimman täsmällisesti, jotta sitä voitaisiin käyttää apuna, kun tutkitaan, miten maailmankatsomuksia voitaisiin rakentaa ja arvioida.

- Esitetään syitä, miksi maailmankatsomuksia tarvitaan (kappale 2.2 Maailmankatsomusten tarve). On hyödyllistä tietää, mihin maailmankatsomuksia tarvitaan, koska tämä ratkaisee, millä perusteilla niitä arvioidaan ja minkälaiseen käyttöön niitä rakennetaan.

- Esitetään, kuinka maailmankatsomuksia rakennetaan (kappale 2.3 Maailmankatsomusten rakentaminen). Maailmankatsomusten rakentaminen on haastavaa, joten on hyödyllistä selvittää, mikä taho tähän tehtävään sopii parhaiten, ja millä tavalla se tulisi suorittaa.

- Esitetään, kuinka maailmankatsomuksia arvioidaan (kappale 2.5 Maailmankatsomusten arviointi). Jotta voitaisiin tunnistaa ja erottaa hyvä maailmankatsomus huonosta, täytyy esittää kriteerit, joiden perusteella maailmankatsomuksia arvioidaan, hyväksytään ja hylätään.

- Esitetään mahdollisimman hyvä maailmankatsomus (nykyisille ihmisille) (kappale 3 Älykkäät systeemit -maailmankatsomus). Tämä on tutkielman päätavoite, jonka alitavoitteita edellisissä kappaleissa esitettävät määritelmät ovat. Nimitän tätä maailmankatsomusta Älykkäät systeemit -maailmankatsomukseksi. Se perustuu Walter Fritzin kirjassaan *Intelligent systems and their societies* (2007) esittämiin ajatuksiin. Sen peruskäsitteet tulevat eri tieteenaloilta, pääasiassa tekoälystä ja systeemiteoriasta. Fritzin (2007) lisäksi olen kerännyt mielestäni vakuuttavia ja yhteensopivia ajatuksia myös muista lähteistä.

2 MAAILMANKATSOMUSTEN RAKENTAMINEN JA ARVIOINTI

Tässä osassa määritellään maailmankatsomuksen käsite ja siihen liittyvä käsitteellinen viitekehys. Lisäksi esitetään syitä miksi maailmankatsomukset ovat agenteille tarpeellisia, miten maailmankatsomuksia rakennetaan ja kenen tehtävä sen tulisi olla, sekä esitetään ja perustellaan kriteerejä, joilla maailmankatsomuksia arvioidaan.

2.1 Maailmankatsomukset

Maailmankatsomuksen käsite (alun perin Immanuel Kantin käyttämä *Weltanschauung* (Kant 1987, 111- 112)) on määritelty monella eri tavalla (kts. Naugle 2002). Täysin tyydyttävää tai yhtä yleisesti hyväksyttyä ei ole olemassa. Urpo Harvan mukaan maailmakatsomus on kokonaiskäsitys maailmasta sekä ihmisen asemasta ja tehtävästä siinä. Maailmankatsomukseen kuuluu toisaalta maailmankuva, eli luontoa ja ihmistä koskeva tieto, joka saadaan omista havainnoista, tieteen tuloksista sekä filosofiasta, ja toisaalta näkemys siitä, mikä on hyvää ja pahaa sekä miten ihmisen pitää toimia ja elää. (Harva 1980, 9.) Oiva Ketosen mukaan ihmisellä on luonnostaan jonkinlainen "olevaista koskeva kokonaisnäkemys", johon kuuluvat hänen toiveensa, pelkonsa ja perimmäiset tavoitteensa. Se on tietoinen tulos olevaisen perimmäisiä salaisuuksia koskevasta omakohtaisesta kysymisestä. (Ketonen 1981, 2.) Nietzschen mukaan maailmankatsomukset ovat kulttuurisia entiteettejä, joiden tuotteita ja joista riippuvaisia ihmiset annetussa maantieteellisessä paikassaan ja historiallisessa kontekstissaan ovat. Ne ovat keinotekoisia ja idiosynkraattisia näkökulmia elämään, välttämättömiä ihmiselämän kannalta, viime kädessä kaoottisessa tai tuntemattomassa maailmassa navigoimiseen. Maailmankatsomus tuottaa välttämättömän, hyvin määritellyn rajan, joka strukturoi ajatuksia, ja käyttäytymistä ja tuottaa standardit, joilla kaikkia asioita mitataan. Kannattajiensa näkökulmasta maailmankatsomukset oovat vastaansanomattomia, toistensa kanssa yhteismitattomia konstruktioita, mikä voi tehdä kulttuurienvälisen kommunikaation vaikeaksi tai mahdottomaksi. (Naugle 2002, 101- 102, 106.) Toisia, hyvin erilaisiakin määritelmiä on esitetty (katso: Naugle 2002). Yleensä maailmankatsomuksia on pidetty enemmän tai vähemmän yhtenäisinä ajatuskokonaisuuksina, joiden avulla ihmiset ovat

yrittäneet ymmärtää itseään ja ympäristöään. Käsitettä käytetään paitsi filosofiassa, myös mm. teologiassa, antropologiassa ja kasvatustieteessä (Vidal 2007, 7).

Määritelmien yleisimmät puutteet ovat epätäsmällisyys, ei-tieteellisyys ja määrittelyalueen kapeus (Useiden määritelmien mukaan maailmankatsomuksia voi olla vain ihmisillä) sekä liian vähäinen yksityiskohtaisuus. Aiheen kannalta ei ole tarpeen arvioida yksittäisiä maailmankatsomuksia tarkemmin, vaan riittää, että muodostetaan määritelmä, jolla ei ole mm. näitä puutteita. Esitän ja puolustan määritelmää, jonka tavoitteena on istuttaa maailmankatsomuksia koskeva (lähinnä filosofien harrastama) tutkimus täsmälliseen ja empiirisesti testattuun käsitteistöön, jotta tutkimusta voidaan tehdä eksaktimmin, viedä pidemmälle ja lopulta liittää tieteelliseen tietoon. Mielestäni tekoälytutkimuksen käsitteistö sopii tähän tehtävään hyvin. Se on täsmällinen, tieteellinen, formalisoitu, yleinen (toisin kuin psykologian joka keskittyy tiettyihin toteutuksiin, yleensä ihmisten tai eläinten aivoihin).

Käsitteitä kannattaa tehdä ja käyttää niiden hyödyllisyyden takia. Muut kriteerit voidaan oikeuttaa vain sillä perusteella, että niiden täyttäminen johtaa hyödyllisyyden lisääntymiseen. Ihmisen, kuten muidenkin tunnettujen agenttien älyllinen kapasiteetti on rajallinen, joten liian monimutkaiset määritelmät ovat käyttökelvottomia, ja epätäsmälliset eli liian monitulkintaiset, sumeat tai ristiriitaiset antavat liian monia tai ei yhtään mahdollista tulkintaa. Mitä yleisempi määritelmä on, sitä useammassa tapauksessa käsitettä voidaan käyttää ja sen tarjoamaa hyötyä, mm. laskennallisten resurssien säästöä tapahtuu. Määritelmien tulee olla yksinkertaisia, jotta niihin perustuva päättely olisi helppoa. Niiden tulee olla täsmällisiä ja ilmaistu aikaisemmin määritellyillä sanoilla, jotta päättely olisi eksaktia, ja jotta määritellyt käsitteet muodostaisivat yhtenäisen järjestelmän. Määritelmien tulee olla käyttökelpoisia, jotta niistä voidaan tehdä kiinnostavia ja käyttökelpoisia johtopäätöksiä.

Yleisen määritelmäni mukaan maailmankatsomus on kokonaiskäsitys (teknisesti: informaatiosysteemi) sen omaajan ympäristöstä, tavoitteista ja menetelmistä tavoitteiden saavuttamiseksi. Jokaisella agentilla ja siksi ainakin useimmilla elävillä oliolla ja joillakin elottomilla, kuten roboteilla, on maailmankatsomus, vähintään implisiittisenä ja

tiedostamattomana. Tämä on maailmankatsomuksen käsitteen yleinen määritelmä, jota täsmennän ja perustelen.

Lähdemme liikkeelle tutkimalla, mitä osia maailmankatsomuksiin kuuluu. Niiniluodon mukaan maailmankatsomuksilla tarkoitetaan enemmän tai vähemmän jäsenneltyä uskomusten ja arvostusten järjestelmää. Niiden osia ovat 1) maailmankuva, eli käsitys siitä, millainen maailma on, 2) tieto-oppi, eli käsitys siitä, miten maailmaa koskevaa tietoa hankitaan ja 3) arvot eli käsitys siitä, millainen maailman pitäisi olla ja mitkä ovat hyvän elämän tavoitteet. Uskomukset ja arvostukset yhdessä antavat perustan elämänkatsomukselle, henkilökohtaiselle käsitykselle elämän tarkoituksesta, omasta paikasta ja tehtävästä maailmassa. (Niiniluoto 1984, 87 ja 1994, 45.)

Maailmankuva on maailmaa koskevien, tavalla tai toisella perusteltujen väitteiden järjestelmällinen kokonaisuus. Maailmankuvia voivat kannattaa yksityiset ihmiset ja ryhmät. Niitä voidaan keksiä, omaksua ja puolustaa eri tavoilla ja luokitella näiden eri tapojen perusteella. (Niiniluoto 1984, 79.) Maailmankuva on maailmankatsomuksen osa, joka ei Niiniluodon mukaan yksinään riitä maailmankatsomukseksi, koska sen perustelemisessa ja käytössä joudutaan menemään sen itsensä ulkopuolelle. Esimerkiksi "jos joku henkilö ilmoittaa kannattavansa tieteellistä maailmankuvaa, hän ei voi muodostaa tätä maailmankuvaa ottamatta ensin kantaa tieteen määritelmään, ja tiedon hankkimiseen liittyviin tieto-opillisiin ongelmiin. Muodostettuaan maailmankuvansa hän ei voi perustaa toimintaansa sen varaan ottamatta kantaa moraalisia ja yhteiskunnallisia arvoja koskeviin filosofisiin ongelmiin". (Niiniluoto 1984, 86.)

Niiniluoto jakaa maailmankuvat kolmeen luokkaan:
1) Tieteelliseen maailmankuvaan kuuluvat väitteet ovat tieteellisin menetelmin hankittuja ja perusteltuja sekä tiedeyhteisön hyväksymiä. Esimerkiksi kvanttiteoria ja evoluutioteoria kuuluvat nykyiseen tieteelliseen maailmankuvaan, kun taas Raamatun luomiskertomus ei. Tieteellinen maailmankuva on historiallisesti kehittyvä, avoin ja itseään korjaava: Sen kaikki osat ovat periaatteessa arvosteltavissa ja muutettavissa uuden tosiasia-aineiston perusteella. Tieteellisille väitteille asetetaan julkisen perusteltavuuden ja testattavuuden vaatimus: niiden

on kestettävä vertailu todellisuuden kanssa ja läpäistä tiedeyhteisön sisällä käytävä kriittinen keskustelu. Uuden tieteellisen teorian tulee a) kyetä selittämään ainakin samat tosiseikat kuin nykyinen teoria sekä jotakin muuta mitä nykyinen teoria ei selitä. Lisäksi sen tulisi olla b) loogisesti ristiriidaton, c) yhteensopiva muun tieteellisen tiedon kanssa, d) periaatteessa kokemusperäisesti testattavissa ja myös läpäistä testit. Maailmankuva on epätieteellinen, jos se sisältää sellaisia väitteitä, jotka ovat ristiriidassa (tietyllä hetkellä vallitsevan) tieteellisen maailmankuvan kanssa. (Niiniluoto 1984, 79- 81.) Maailmankuva on ei-tieteellinen, jos se ei ole ristiriidassa tieteellisen maailmankuvan kanssa, ja jos se muuten koskee eri asioita kuin tiede. (Niiniluoto 1984, 82.)

2) Uskonnolliseen maailmankuvaan sisältyy väitteitä, joiden ainoana tukena on vetoaminen joihinkin uskonnollisiin auktoriteetteihin tai henkilökohtaisiin uskonnollisiin kokemuksiin tai elämyksiin. (Niiniluoto 1984, 81- 82.)

3) Metafyysinen maailmankuva sisältää maailmaa koskevia väitteitä, jotka on perusteltu tieteen kokemusperäisen metodin sijasta filosofisten argumenttien avulla. Metafyysiset väitteet eivät ole edellä mainitussa mielessä uskonnollisia, sillä ne nojautuvat ihmisen järkeen. Metafyysinen maailmankuva voi olla epätieteellinen, mutta tavallisemmin se on ei-tieteellinen. Metafyysistä maailmankuvaa voi kutsua tiedepohjaiseksi, jos se sisältää tieteellisen maailmankuvan ohella joitakin erityistieteen tuloksiin pohjautuvia filosofisia yleistyksiä maailman perusluonteesta. (Niiniluoto 1984, 82- 83.)

Tieteellinen maailmankatsomus olisi tietyssä mielessä ihanteellinen, koska se perustuisi tieteelliseen tietoon, jota usein pidetään varmimpana, perustetaanhan se yleensä empiirisiin faktoihin. Varma ja hyvin järjestelty tietämys on usein menestyksellisen toiminnan edellytys. Tieteellinen tietämys pyritään muotoilemaan mahdollisimman tiiviiksi, selkeäksi, yksiselitteiseksi ja aukottomaksi. Siksi maailmankatsomuksen tieteellisyys on pidettävä niin suurena kuin mahdollista. Niiniluodon mukaan ainakaan nykyään maailmankatsomus ei voi koostua pelkästään tieteellisestä tiedosta, koska tämän hetken tieteellisen tiedon kokonaisuus ei sisällä vastauksia kaikkiin kysymyksiin, joihin kokonaisen maailmankatsomusten yleensä vaaditaan vastaavan. Tässä tiukassa mielessä ns. tieteellinen maailmankatsomus ei ainakaan nykyään ole mahdollinen. Maailmankatsomuksen osat tietoteoria ja arvoteoria ovat nykyään filosofian osa-alueita, ja vaikka filosofia voikin parhaimmillaan olla yhtä järjestelmällistä ja

kriittistä kuin mikä tiede tahansa, on silti aiheellista tehdä ero filosofian ja erityistieteiden (kuten fysiikka, psykologia, jne.) välillä. Niiden tutkimuskohde ja menetelmät ovat erilaisia, eikä filosofian johtopäätösten pätevyydelle ole yhtä objektiivisia kriteerejä kuin reaalitieteissä. (Niiniluoto 1984, 87.)

Silti voidaan puhua tieteellisestä maailmankatsomuksesta hieman avarammassa mielessä: Niiniluodon mukaan tieteellisen maailmankatsomuksen on oltava kaikilta osiltaan tieteen ihanteita kunnioittava siinä mielessä, että se on filosofisilta osiltaankin avoin, kriittinen ja itseään korjaava. Tieteellistä maailmankatsomusta luonnehtii tietoteoreettinen näkemys, jonka mukaan tieteellinen metodi on paras ja luotettavin menetelmä maailmaa koskevan tiedon hankkimiseksi. Tieteen metodin tunnusmerkkejä ovat julkisesti koeteltavissa olevan aineisto, oletusten kriittinen arvioiminen ja pitäytyminen luonnollisiin selityksiin yliluonnollisten sijasta. Tieteellinen maailmankatsomus sisältää tieteellisen maailmankuvan sekä sen tietoteoreettisesta osasta riippuen mahdollisesti joitakin filosofisia yleistyksiä. (Niiniluoto 1984, 88.) Tieteellinen ja uskonnollinen maailmankatsomus ovat tietoteoreettiselta osaltaan ristiriidassa keskenään. Tieteellisessä maailmankatsomuksessa hyväksytään periaatteessa vain yksi maailmaa koskevan tiedon tavoittelun keino - tieteellinen menetelmä - kun taas uskonnollisissa katsomuksissa sitoudutaan sen lisäksi tai sijasta johonkin menetelmään, joka ei täytä tieteellisyyden kriteerejä. (Niiniluoto 1984, 88.)

Vaikka Niiniluodon määritelmä antaa joitakin keinoja maailmankatsomusten ymmärtämiseen, tarkempaa erittelyä, arviointia ja rakentamisen ohjetta varten tarvitaan yksityiskohtaisempi määritelmä. Clement Vidal (2007), Aerts ja kumppanit (2007) sekä Francis Heylighen (2000) tarjoavat yksityiskohtaisemman maailmankatsomuksen määritelmän. Ne ovat hyvin samankaltaisia ja suurelta osin lähtöisin Aertsin ja kumppaneiden (2007) tutkimuksesta. Kokonaiseen maailmankatsomukseen kuuluu ainakin seuraavat kuusi osaa: mallit maailmasta, menneisyydestä ja tulevaisuudesta sekä teoriat tiedosta, arvosta ja toiminnasta. Ne täytyy yhdistää koherentiksi ja käyttökelpoiseksi kokonaisuudeksi. Osien alaan kuuluvat kysymykset ovat usein ikivanhoja ja perinteisiä filosofian kysymyksiä. (Vidal 2007, 7).

1) Malli maailmasta/ympäristöstä (Ontologia): Mitä on olemassa? Miksi on olemassa jotain? Millainen maailma/universumi on? Mikä on maailman rakenne ja kuinka se toimii? (Vidal 2007, 8.), (Aerts, Apostel ja muut. 2007, 14.), (Heylighen 2000, What is a worldview?)

2) Malli menneisyydestä (Historia): Mistä kaikki on tullut? Miksi maailma on sellainen kuin on? Miksi olemme tällaisia kuin olemme emmekä erilaisia? Mikä on universumin alkuperä? Millaisia yleisiä selittäviä periaatteita voidaan käyttää?(Vidal 2007, 8.) Maailmankatsomuksen selitysvoima on sitä suurempi, mitä yleisempiä ja käyttökelpoisempia lakeja tai säännönmukaisuuksia se löytää todellisuudesta. Käytännössä selitys tarkoittaa vähemmän itsestään selvien tosiasioiden johtamista yleisistä ja hyväksytyistä laeista. (Aerts, Apostel et al. 2007, 14- 15.) Jos kykenemme selittämään miten ja miksi tietty ilmiö syntyy, voidaan usein helpommin ennustaa, kuinka se kehittyy. (Heylighen 2000, What is a worldview?)

3) Malli tulevaisuudesta (Futurologia): Minkälaisia tulevaisuuksia me ja lajimme voi saavuttaa? Millä kriteereillä valitsemme näistä mahdollisista tulevaisuuksista? Tulevaisuuden kehitystä koskevan mallin tulisi antaa lista enemmän tai vähemmän todennäköisistä tulevaisuuden kehityksistä. Näistä voitaisiin sitten valita, mihin tulisi pyrkiä ja mitä tulisi välttää. (Heylighen 2000, What is a worldview?). Mikä on elämän kohtalo universumissa? Minne olemme menossa? (Vidal 2007, 8.) Kuinka kulttuurit vuorovaikuttavat tulevaisuudessa? Mikä on tieteen ja talouden rooli tulevaisuudessa? Mikä taho tekee tai tulee tekemään ihmiskunnan tulevaisuuteen vaikuttavia päätöksiä? Leviävätkö ihmiset maapallon ulkopuolelle tulevaisuudessa? (Aerts, Apostel et al. 2007, 17.)

4) Teoria tiedosta (Epistemologia): Mitä tieto on ja kuinka sitä voidaan hankkia? Millä tavalla voidaan rakentaa käsitys maailmasta, jotta voitaisiin vastata maailmankatsomuksen osia koskeviin kysymyksiin? Kuinka voidaan hankkia tietoa? Mikä on totta ja mikä ei? (Vidal 2007, 9.) Suunnitelmat perustuvat tietoon, teorioihin ja malleihin, jotka kuvaavat ilmiöitä. Tämän takia täytyy ymmärtää, kuinka rakentaa luotettavia teorioita ja malleja, ja erottaa huonot teoriat hyvistä. (Heylighen 2000, What is a worldview?)

5) Teoria arvoista (Aksiologia): Mikä on arvokasta ja miten se saavutetaan? Mihin meidän tulisi pyrkiä? Mitä on hyvä ja paha, tai hyvä ja huono? Mikä on elämän tarkoitus? Miksi tunnemme niin kuin tunnemme ja miten suhtaudumme todellisuuteen ja mikä on roolimme siinä? (Vidal 2007, 8.) Miten muodostetaan moraali tai etiikka, sääntöjen joukko, joka kertoo millä tavalla tulisi tai ei tulisi käyttäytyä. (Heylighen 2000, What is a worldview?) Aksiologia tutkii mm. moraalia, etiikkaa ja estetiikkaa koskevia kysymyksiä. Ihmiset ihailevat, rakastavat tai arvostavat joitain asioita maailmassa, ovat välinpitämättömiä toisista ja vihaavat tai inhoavat joitain. Maailmankatsomus ei tule ainoastaan tehdä maailmasta ymmärrettävä, vaan tarjota myös keinoja arvioida sitä. Tämän osan tulee antaa tavoite, päämäärä, suunta elämälle. (Vidal 2007, 8.) Onko/voidaanko muodostaa arvojen ja päämäärien hierarkiaa? Onko yleisiä arvoja tietyille agenttien joukoille kuten kaikille ihmisille? Miten muodostetaan etiikka, tiettyjen agenttien vuorovaikutusta koskevien sääntöjen joukko, joka auttaa toteuttamaan tietyt arvot/tavoitteet?

6) Teoria toiminnasta (Prakseologia): Millä tavalla toimimalla tavoitteet voidaan saavuttaa? Mitkä ovat yleiset periaatteet, joiden mukaan toiminta pitäisi organisoida? Tällainen tieto auttaa suunnitelmien toteuttamista käytännön ongelmien ratkaisemiseksi. (Vidal 2007, 8.) Sopeutua voi monella tavalla. Voidaan muuttua itse tai muuttaa ympäristöä. Ihmislajin tyypillinen piirre on muuttaa ympäristöä toteuttamaan ihmisten tavoitteita. Maailmankatsomuksen tulee sisältää organisoitu näkemys tosiasiallisista ja mahdollisista vaikutuksista, joita ihmisillä voi olla ympäristöönsä. Esimerkiksi jos halutaan rakentaa globaali maailmankatsomus, sille on hyväksi yleinen prakseologia, yleinen päätöksenteko- ja ongelmanratkaisuteoria ja strategisen tutkimuksen ja suunnittelun yksikkö, jotta sen toiminta olisi mahdollisimman tehokasta. (Aerts, Apostel et al. 2007, 19- 20.) Prakseologia kertoo millä tavalla toimimalla tavoitteet voidaan saavuttaa. (Heylighen 2000, What is a worldview?)

Seuraavassa taulukossa on esimerkkejä maailmankatsomuksista eriteltynä maailmankatsomusten osiin. Tarkoituksena ei ole antaa täydellisiä kuvauksia niistä, vaan esittää, kuinka tarkastelutapa toimii.

Taulukko 1: Esimerkkejä maailmankatsomuksista, jotka on eritelty maailmankatsomusten osiin. (Lähde: Vidal 2008, 6).

	Tieteellinen	**Uskonnollinen**	**Bakteerin**	**Yhteisön**
1. Ontologia	Fysikalismi, ei jumalaa	Kaksi aspektia: mieli ja ruumis	Nykyiset aistimukset	Yhteinen kulttuurinen ontologia
2. Historia	Universumin ja sen kehittymisen tieteelliset mallit	Vastaukset pyhissä kirjoituksissa	Eräänlainen muisti, joka voi olla bakteerin nykyinen biokemiallinen tila	Nykyisen yhteisön selitys, historia, traditionaalinen tieto
3. Futurologia	Maailmaa ennustavia malleja	Elämä kuoleman jälkeen	Geeneihin perustuva palautesysteemi (feedback system)	Poliittiset suunnitelmat, ennustaminen
4. Epistemologia	Havaintojen ja teorioiden vuorovaikutus osien 1, 2, ja 3 rakentamiseksi	Tieto saadaan pääosaksi pyhistä kirjoituksista ja uskonnollisista kokemuksista	Joitakin perushavaintoja	Informaatio saadaan sosiaalisen kulttuurinsiirron kautta (mm. koulut, media)
5. Aksiologia	Vain tieteellistä tutkimusta koskevia arvoja	Konkreettiset ja muuttumattomat arvot pyhissä kirjoituksissa	Pääasiassa geneettisesti määräytyneet: etsi ravintoa, lisäänny	Utopia, poliittiset ja taloudelliset arvot
6. Prakseologia	Ei toiminnanohjeita	Joitakin toimintoja ehdotetaan,	Liiku, syö, lisäänny	Poliittiset toiminnot, yhteisön

		uskonnollisen elämän kirjoituksista ja muilta auktoriteeteilta		jäsenten toiminnot

Näyttää siltä, että kaikki tunnetut älykkäät systeemit tarvitsevat ainakin jonkinlaisen maailmankatsomuksen vuorovaikutukseen ympäristönsä kanssa. Jokainen sen osista ja niiden yhteensopivuus on välttämätön tähän, ja lisäksi näyttää siltä että enempiä osia ei välttämättä tarvita, ainakaan tuntemissamme tapauksissa. Niinpä Vidalin/Aertsin ja kumppaneiden/Heylighenin maailmankatsomuksen määritelmä on oikea. Tämän voi todeta keksimällä erilaisia vuorovaikutustilanteita agentin ja ympäristön välillä. Riittävätkö nämä osat mahdollisimman menestyksekkääseen vuorovaikutukseen? Ilmeisesti kyllä. Onko joku niistä liikaa? Ilmeisesti ei. Välttämättömiä osia voitaisiin myös etsiä historiallisen tutkimuksen avulla: Löytyykö osia, jotka löytyvät kaikista tunnetuista maailmankatsomuksista? Tutkimuksen lähtökohtien kannalta tämä ei ole niin relevantti lähestymistapa, koska historiallinen esiintyvyys ei ole yksinkertaisessa suhteessa osan yleisen hyödyllisyyden kanssa. Osien lisäksi tarvitaan täsmällinen ja uskottava käsitteellinen viitekehys, jotta maailmankatsomuksen tehtävät ja yhteys agentin kognitioon voidaan ymmärtää.

2.1.1 Maailmankatsomusten käsitteellinen viitekehys

Maailmankatsomus on perusrakenteeltaan seuraavanlainen informaatiosysteemi: Maailmankatsomukseen kuuluu 1) tavoitteet, 2) tieto, ja tavoitteiden ja tiedon perusteella muodostetut 3) toiminnot/toimintastrategiat tavoitteiden saavuttamiseksi:

1) Tavoite on tietty tilanne, johon systeemi pyrkii (Fritz 2007, Glossary). Usein tavoitteet voidaan sijoittaa preferenssijärjestelmään tai preferenssihierarkiaan. Ylimpänä hierarkiassa on perimmäinen tavoite/tavoitteet. Sen alapuolella on perimmäisen tavoitteen saavuttamiseen tähtääviä alitavoitteita. Alitavoitteen saavuttaminen auttaa seuraavaksi hierarkiassa ylempänä olevan tavoitteen saavuttamista, tai on osa sen saavuttamisen suunnitelmaa. Agentti yrittää

saavuttaa tavoitteitaan toimimalla ympäristössään. Tämä tarkoittaa sitä, että se yrittää löytää ja suorittaa toiminnon A tilanteessa B, joka siirtää ympäristön tavoitetilaan C. Tämän toiminnon agentti yrittää löytää ympäristöstään saamansa tiedon avulla.

2) Tiedosta muodostetaan tietokanta, jonka perusteella agentti rakentaa esityksen nykyisestä tilanteestaan, ja jonka perusteella se päättelemällä yrittää löytää toiminnon, jonka avulla se saavuttaa tavoitteensa. Tässä vaiheessa joudutaan määrittelemään joitakin ymmärtämisen kannalta olennaisia käsitteitä.

Agentti on mikä tahansa, mitä voidaan tarkastella sensoreillaan ympäristöään havainnoivana ja siinä aktuaattoreillaan toimivana (Russell ja Norvig 2003, 32). Älykäs systeemi on mikä tahansa systeemi, mikä voi muokata käyttäytymistään ympäristössään tietyn tavoitteen mukaisesti. Tässä tekstissä agentilla ja älykkäällä systeemillä tarkoitetaan olennaisesti samaa. Tarkempi erottelu on mahdollista (agentit, jotka eivät voi muokata käyttäytymistään tietyn tavoitteen mukaisesti, eivät ole älykkäitä systeemejä), mutta ei ole tässä välttämätöntä. Agentti tai älykäs systeemi on systeemi. Systeemi on osa universumia, jolla on rajoitettu olemassaolo ajassa ja tilassa. Enemmän tai vahvempia korrelaatioita on systeemin eri osien välillä, kuin systeemin osien ja systeemin ulkopuolisten universumien osien välillä. (Fritz 2007, Glossary.) Universumi on kaikki mitä on olemassa, kokonaisuutena tarkasteltuna (Fritz 2007, Glossary).

Agentti toimii ympäristössä. Systeemin ympäristöä on se osa universumista, joka on kommunikaatiossa systeemin kanssa, mutta ei ole osa sitä (Fritz 2007, Glossary). Kommunikaatio on materian tai energian liikettä kahden universumin osan välillä (Fritz 2007, Glossary). Agentti havainnoi ympäristöään, eli vastaanottaa ympäristöstä tulevia kommunikaatioita. Systeemin sensorit vastaanottavat kommunikaatioita ympäristöstä ja, joissakin systeemeissä, koodaavat ne, ja lähettävät tämän informaation edelleen keskuksiin edelleenprosessoitavaksi (Fritz 2007, Glossary). Sensori (aistinelin) on se osa systeemiä, joka voi vastaanottaa kommunikaatioita ympäristöstä (Fritz 2007, Glossary). Informaatio on käsitteiden ja vastesääntöjen summa, joka voidaan erottaa kommunikaatiosta (Fritz 2007, Glossary).

Agentti rakentaa käsitteistään esityksen nykyisestä tilanteesta havaintojensa perusteella. Se on käsitteiden ryhmä, joka ilmaisee agentin ja ympäristön tilan tiettynä (nykyisenä) ajanhetkenä. Käsite on informaation fysikaalinen tallenne (Fritz 2007, Glossary). Aivot (informaation käsittelyn keskus, jossa tapahtuu suuri osa tallentamisesta, hävittämisestä ja prosessoinnista) vastaanottavat sensoreilta saapuvaa informaatiota ja muodostavat niiden perusteella käsitteitä. Käsite on informaatiorakenne, joka liittää tiettyihin ominaisuuksiin tietyt muut ominaisuudet. Käsitteet sisältävät kohteitaan koskevaa, agentille hyödyllistä tietoa. Jos agentti on oppinut, miten tietty kohde jossakin suhteessa käyttäytyy, ja jos jotkin toiset kohteet muistuttavat jollakin tavalla tätä kohdetta, agentti voi koettaa olettaa, että näiden toisten kohteiden käyttäytyminen on jollakin tavalla samanlaista kuin ensimmäisen. Jos tämä oletus eri kokeilujen perusteella pitää paikkansa, agentti voi tästä eteenpäin olettaa, että mikä tahansa, mikä täyttää ainakin jossain määrin kohteen tuntomerkit, ainakin jossakin määrin käyttäytyy tällä tavalla. Tällainen tietorakenne, joka liittää tietyt ominaisuudet tiettyihin toisiin ominaisuuksiin, on käsite. Jos käsite ainakin yleensä pitää paikkansa (toimii käytännössä), se voi säästää agentin laskennallisia resursseja, koska tällöin joitakin kohteita ei tarvitse käsitellä uusina, vaan voidaan käyttää hyväksi jo ennestään hankittua tietoa. Käsite voidaan yksinkertaisimmillaan ilmaista seuraavana informaatiorakenteena: "Jos kohde x täyttää käsitteen Y tuntomerkit (käsitteen määrittelevät ominaisuudet) paremmin kuin muiden käsitteiden määrittelevät ominaisuudet, esitä kohde käsitteellä Y." Tällöin kohteen oletetaan käyttäytyvän niin kuin muidenkin käsitteen alaan kuuluvien kohteiden (eli instanssien).

Esitys nykyisestä tilanteesta voidaan rakentaa tietokannan (tietämyskannan, engl. knowledge base) tiedon perusteella. Tietokanta on lauseiden joukko, jossa jokainen lause on ilmaistu tiedonesittämiskielellä ja esittää jonkin väitteen jostakin kohteesta. Tietokantaan voidaan lisätä ja siitä voidaan poistaa lauseita. (Russell ja Norvig 2003, 195.) Tiedonesittämiskieliä on monia. Usein tekoälysovelluksissa käytetään predikaattilogiikkaa (Niemelä 1993, 118- 119). Tietokannan lauseet muodostetaan tiedonesittämiskielen syntaksin mukaan. Syntaksi määrittelee kaikki tiedonesittämiskielen hyvin muodostetut lauseet. (Russell ja Norvig 2003, 200.) Tiedonesittämiskielellä on myös semantiikka, joka määrittelee lauseiden totuuden. Esimerkiksi tavallisesti aritmetiikassa käytetyn semantiikan mukaan lause "x + y = 4" on tosi mahdollisessa maailmassa (tai ympäristössä), jossa x = 2 ja y = 2, mutta epätosi maailmassa,

jossa x = 1 ja y = 1. (Russell ja Norvig 2003, 201.) Tietokannan lauseista voidaan johtaa uusia lauseita joidenkin päättelysääntöjen mukaan. Yksinkertaisin esimerkki tästä on logiikassa looginen seuraamus (engl. entailment). Looginen seuraamus -suhde lauseiden A ja B välillä tarkoittaa, että jos lause A on tosi, tästä seuraa loogisesti, että myös lause B on tosi. (Russell ja Norvig 2003, 201- 202.) Usein tietokannan lauseista loogisesti seuraavien lauseiden (jotka siis implisiittisesti kuuluvat tietokantaan) joukko on hyvin suuri. Päättelyalgoritmi on jokin periaate, jonka tehtävänä on tietyn tietokantaan kuuluvan lauseen johtaminen. Päättelyalgoritmi on totuuden säilyttävä (engl. truth preserving, sound), jos se johtaa vain lauseita, jotka seuraavat tietokannasta loogisesti. Se on täydellinen (engl. complete), jos se voi johtaa minkä tahansa tietokannan lauseista loogisesti seuraavan lauseen. Molemmat ovat haluttavia päättelyalgoritmien ominaisuuksia. (Russell ja Norvig 2003, 203.) Lauseet ovat agentin fysikaalisia konfiguraatioita (myös ihmisillä; aivot ovat fysikaalinen tallennus- ja laskentalaite), ja päättely on prosessi, jossa johdetaan uusia fysikaalisia konfiguraatioita vanhoista. On toivottavaa, että päättelysäännöt ovat totuudensäilyttäviä, eli että niiden perustella tosista lauseista johdetut uudet lauseet ovat myös tosia. (Russell ja Norvig 2003, 203.) Konnektiivien avulla lauseista voidaan muodostaa monimutkaisempia lauseita (Russell ja Norvig 2003, 204), joiden avulla tuotettujen lauseiden totuus on määritelty. Nekin ovat siis eräänlaisia päättelysääntöjä.

Tiedonesittämiskielen merkit tai symbolit viittaavat käsitteisiin tai (muihin informaatiorakenteisiin (esim. aisti-informaatioon)). Käsitteen merkki on informaatiorakenne, jonka tietty tulkinta liittää tiettyyn käsitteeseen. Merkkejä käytetään, koska niitä on usein helpompi prosessoida ja kommunikoida kuin käsitteitä. Kun merkki kommunikoidaan esim. äänen (puhe, musiikki) tai valon (kuva, kirjoitus) välityksellä (koodaamalla ja dekoodaamalla se eri muotoihin), systeemille, jolla on oikea merkin tulkinta, merkki viittaa suunnilleen samanlaisiin käsitteisiin tai vastesääntöihin tämän toisen systeemin tietokannassa. Tulkinta määrittelee, mitkä merkit viittaavat mihinkin käsitteisiin. Käsitteet taas viittaavat sisäisesti tai ulkoisesti havaittaviin kohteisiin. Nykyinen tila voidaan esittää lauseiden konjunktiona: Nykyinen tila on tietty tila, jossa pätee: "A, B, C. . .", jossa A, B ja C ovat lauseita, jotka ovat tosia nykyisessä tilanteessa.

3) Toiminnot, suunnitelmat, keinot tai toimintastrategiat muodostetaan tietojen ja tavoitteiden perusteella. Vaste on se osa vastesääntöä, joka kertoo, minkä toiminnon älykäs systeemi suorittaa tietyssä tilanteessa. Se koostuu yhdestä tai useammasta käsitteestä, joiden suorittaminen systeemin aktuaattoreilla tuottaa muutoksen ympäristössä. (Fritz 2007, Glossary.) Vastesääntö on fysikaalinen tallenne, johon kuulu tilanne, siinä suoritettava vaste ja sen tulos (Fritz 2007, Glossary), eli ympäristön tilan muutos, ja arvio siitä, oliko se haluttava vai ei ja kuinka paljon. Vaste voidaan ilmaista esimerkiksi seuraavasti: "Tilanteessa D suorita vaste E." Vastesääntö voidaan ilmaista esimerkiksi näin: "Tilanteessa F suoritettu vaste G tuottaa tilanteen H, jonka haluttavuus on esimerkiksi (4) asteikolla (1-5)." Vastesäännöt on muodostettu aikaisemmista kokemuksista tai johdettu jo tietokannassa olevista vastesäännöistä (Fritz 2007, Glossary). Suunnitelmat ovat vasteiden tai vastesääntöjen sekvenssejä, jotka älykäs systeemi suorittaa yksi toisensa jälkeen (Fritz 2007, Glossary). Ne ovat siis yksittäisistä vasteista koottuja monimutkaisempia vasteita.

2.1.2 Maailmankatsomuksen yhteys kognitiiviseen arkkitehtuuriin

Esitän maailmankatsomuksen yhteyden kognitioon yleisesti riippumatta mistään tietystä kognitiivisesta arkkitehtuurista. Tarkoitus on selvittää tämän filosofiassa usein käytetyn hieman epäselvän käsitteen yhteys vakiintuneempaan ja testattavaan tekoälyn käsitteistöön. On esitetty monia kognitiivisia arkkitehtuureja, mm. SOAR (kotisivu: http://sitemaker.umich.edu/soar/home) ja OSCAR (Pollock 2008). Kognitiiviset arkkitehtuurit ovat agentin arkkitehtuurien osajoukko. Ne kuvaavat agentin kognitiossaan tarvitsemien prosessien vaatimia rakenteita. Maailmankatsomuksen osat ovat laskennallisia osarakenteita ja osaprosesseja, joita agentin kognitiivinen koneisto käy läpi kun se yrittää saavuttaa tavoitteensa. Koko prosessi on syötteeseen tavoitteen saavuttamisen kannalta mahdollisimman sopivan vasteen muodostaminen.

1) Malli maailmasta (ympäristöstä) ja 2) malli menneisyydestä (Historia) muodostavat agentin tietokannan. Agentin perseptiosekvenssissä voi jo sinällään olla hyödyllistä tietoa. Kaikkia agentin kokemuksia ei yleensä voida säilyttää muistitilan puutteen vuoksi. Siksi niillä täytyy olla jokin mekanismi, jolla muistojen säilytyksen pituus päätetään. Lisäksi monet agentit

kykenevät etsimään persptisekvenssistään säännönmukaisuuksia, jotka mahdollistavat käsitteiden ja niistä ontologian rakentamisen. Malli maailmasta on ontologia, jonka avulla agentti hahmottaa ympäristöään. Ontologia tarkoittaa teoriaa siitä, mitä ja millaisia asioita on olemassa (Russell ja Norvig 2003, 261). Ontologia on käsitteistö, johon kaikki ympäristön kohteet luokitellaan. Ontologiat koostuvat kategorioista, jotka ovat tietyn ontologian kannalta olennaisia käsitteitä. Ontologiat voidaan jakaa erityisiin ja yleisiin ontologioihin. Erityiset ontologiat (engl. special-purpose ontology) ovat vain tietyn, esimerkiksi tietyn tehtävän suorittamiseen tarvittavan tiedon esittämiseen luotuja käsitteellisiä viitekehyksiä, kun taas yleiset ontologiat (engl. general-purpose ontology) sopivat minkä tahansa tiedon esittämiseen. Kaksi tunnuspiirrettä erottaa yleisiä ontologioita erityisistä. Yleisen ontologian pitää soveltua käytettäväksi enemmän tai vähemmän millä tahansa erityisalueella. Toiseksi millä tahansa riittävän vaativalla määrittelyalueella (engl. domain), eri tiedon alueet on esitettävä yhtenäisesti, koska järkeilyyn ja ongelmanratkaisuun saatetaan käyttää monen eri alueen tietoa samanaikaisesti. (Russell ja Norvig 2003, 320, 321, 322.) Ontologian tehtävänä on mahdollistaa kohteiden luokittelu, jonka ansiosta niiden ominaisuuksia voidaan päätellä. Melkein kaikilla yleistyksillä on poikkeuksia. Esimerkiksi vaikka "tomaatit ovat punaisia" on hyödyllinen sääntö, jotkut tomaatit ovat vihreitä, keltaisia tai oransseja. Se, miten poikkeuksia tai epävarmaa tietoa käsitellään, on hyvin tärkeää. Relevanttien kohteiden luonne, agentin resurssit sekä se, millaisia suunnitelmia halutaan luoda (tavoitteet), määräävät sen, millainen menestystä edistävän ontologian täytyy olla. Juuri tämän enempää ei voida sanoa yleisellä tasolla.

3) Malli tulevaisuudesta (Futurologia) ja 6) teoria toiminnasta (Prakseologia) ovat suunnitelmia, joita agentti muodostaa tietokannan ja päättelyn avulla, tavoitteensa mukaisesti. Tavoitteen saavuttamiseen tarkoitetun toimintosekvenssin luomista sanotaan suunnitteluksi. (Russell ja Norvig 2003, 375.) Suunnittelun lähtökohta on maailman tila toimintosarjan alussa, jota sanotaan alkutilaksi tai lähtötilaksi. Tämä on tieto siitä, mitkä ongelman kannalta relevantit väitteet ovat tosia. Tavoitteet ja alkutila muodostavat yhdessä ongelman. Lisäksi suunnittelijalla on käytössään joukko toimenpiteitä eli toimintoja, joilla se voi muuttaa maailman tilaa tavoitteiden saavuttamiseksi. (Karanta 1993, 183.) Suunnittelutehtävän mielekkyyden edellytyksenä pidetään yleensä, että: 1) Maailma, jossa suunnittelu tapahtuu, on

ainakin jossain määrin ennustettavissa. Kaoottisissa tai hyvin satunnaisissa ympäristöissä suunnittelu on tehotonta, ja agentti voi vain reagoida tapahtumiin. Lopputulos on sama, jos ympäristö muuttuu niin nopeasti, ettei käytettävissä oleva laskentakapasiteetti riitä suunnitelmien tekemiseen ja muuttamiseen tarpeeksi nopeasti. 2) Käsiteltävä ongelma on mielekkäästi paloiteltavissa lähes erillisiin aliongelmiin, joilla on vain vähän vuorovaikutusta keskenään. (Karanta 1993, 183, 184.) Korkeammilla luonnollisilla älykkäillä systeemeillä ja monilla keinotekoisilla älykkäillä systeemeillä on mielikuvitus. Mielikuvitus on työkalu, jolla voi kokeilla turvallisesti eri toimintoja ennen toimintojen käyttämistä ympäristössä. Mielikuvitus tarkoittaa kykyä esittää nykyinen tilanne mielessä, käyttää siihen sopivaa vastesääntöä ja esittää näin syntyvä uusi nykyinen tilanne. Älykäs systeemi toistaa tätä prosessia, kunnes tavoitetilanne on saavutettu. Kaiken tämän se tekee mielessään, ilman mitään ulkoista toimintaa. (Fritz 2007, Mental methods and chains of response rules: habits.)

4) Teoria tiedosta (Epistemologia) sisältää keinot joilla agentti voi hankkia tietoa ympäristöstään. Tieto voidaan määritellä informaatioksi, jonka avulla voidaan tehdä oikeaan osuvia ennusteita. Yleensä agenteilla on jonkinlainen tietokanta, jossa on agentin tieto. Useissa tapauksissa tietokantaan voidaan lisätä tai ottaa pois tietoa. Agentin toteutuksessa täytyy ratkaista mm. se, millä tavalla tietoa hankitaan, millä periaatteilla perseptisekvenssistä etsitään säännönmukaisuuksia, mitkä muistot säilytetään. Nämä seikat vaikuttavat tiedon määrään ja esitysmuotoon, jotka taas vaikuttavat agentin menestykseen. Agentin epistemologian täytyy olla yhteensopiva sen resursseihin, ympäristöön ja tavoitteisiin.

5) Teoria arvoista (Aksiologia) sisältää agentin tavoitteen ja menestyksen määritelmän (ja mitan). Agentilla täytyy olla jokin mekanismi, jolla se voi päättää, onko tietty tilanne tavoitetila. Kaikilla älykkäillä systeemeillä on tavoite (engl. main objective). Monet voivat myös oppia luomaan ja käyttämään alitavoitteita (engl. sub objective). Alitavoitteet ovat alemman tason ja/tai väliaikaisia tavoitteita. Alitavoitteen saavuttamalla älykäs systeemi lähenee tavoitteensa saavuttamista. (Fritz 2007, Objectives.) Tavoite on usein määritelty ns. menestyksen kriteeristöllä (engl. performance criteria) jonka täyttämällä agentti saavuttaa tavoitteensa. Suoritusarvo (engl. performance measure) kuvaa, kuinka hyvin toiminta täyttää menestyksen kriteerit. Tavoitteen saavuttaakseen agentti suorittaa toimintoja usein

ympäristöstään saamansa palautteen perusteella. Menestyksellinen toiminta vastaa ympäristön tilan muuttumista tavoitetilaksi toiminnan avulla. Ei ole olemassa yhtä ja pysyvää menestyksen kriteeristöä, joka sopisi kaikille agenteille, vaan agenteilla on eri tavoitteita, ja menestys määräytyy niiden saavuttamisen mukaan. (Russell ja Norvig 2003, 35.) Menestyksen tunnistamisen menetelmät eivät useinkaan ole täydelliset. Tämä tieto on välttämätöntä, jotta se voisi tunnistaa menestyksen ja muokata käyttäytymistään sen perusteella. (Russell ja Norvig 2003, 33–35.) Tavoitetesti-funktio (engl. goal test) ratkaisee, onko tietty tila tavoitetila. Joskus mahdollisten tavoitetilojen joukko voidaan esittää eksplisiittisesti, jolloin tavoitetesti yksinkertaisesti selvittää onko annettu tila yksi niistä. Joskus taas tavoite on määritelty abstraktien ominaisuuksien avulla. (Russell ja Norvig 2003, 62.)

2.2 Maailmankatsomusten tarve

1) Maailmankatsomus tarvitaan vuorovaikutukseen ympäristön kanssa: Kaikki älykkäät systeemit tarvitsevat maailmankatsomuksen, vaikka tiedostamattoman, vuorovaikutukseen maailman kanssa. Tämä selvitetään koettamalla, voisiko älykäs systeemi toimia ilman jotakin edellä mainittua maailmankatsomuksen osaa ja lisäksi, tarvitseeko se välttämättä jotain näiden lisäksi toimiakseen. Vaikuttaa siltä, että osat riittävät ja muita ei tarvita. On olemassa käytännöllinen tarve omata edes implisiittinen ja hyvin naiivi vastaus maailmankatsomuksen osia koskeviin kysymyksiin. Niinkin yksinkertainen olio kuin bakteeri tarvitsee jonkinlaisen maailmankatsomuksen vuorovaikutukseen ympäristönsä kanssa (Vidal 2007, 9). Maailmankatsomus, niin kuin se edellä hyvin yleisessä mielessä on määritelty, on jokaisen agentin välttämätön osa.

2) Biologiseen evoluutioon perustuva tarve: Ensimmäiset tunnetut maailmankatsomukset ovat olleet biologisilla organismeilla. Ympäristön mallin rakentamisen välttämättömyys johtuu elävän organismin tarpeesta sopeutua siihen. Tämä johtaa myös maailmankatsomuksen ja sen kehittämisen tarpeisiin (Vidal 2007, 9). Tämän mallin rakenteeseen vaikuttavat sen tekijän tarpeet ja kyvyt (Aerts, Apostel ja muut 2007, 18- 19.) sekä ympäristö. Kaikki eliöt ovat jossakin määrin älykkäitä, koska ne hyödyntävät muistia ja ennustamista lisääntyäkseen tehokkaammin. Kaikki käyttäytyminen, olipa se sitten ihmisen, etanan, yksisoluisen eliön tai

puun käyttäytymistä, on keino hyödyntää maailman struktuuria lisääntymiseen. Jokainen laji käyttäytyy hieman eri tavalla, ja hyödyntää maailman struktuurista hieman eri osia. (Hawkins 2005, 181- 183.) Biologisen evoluution näkökulmasta ihmisen vahvuuksia ovat tiedon hankkiminen ja sen käyttö. Ihmiset eivät ole yhtä monilukuisia ja nopeita lisääntyjiä kuin muurahaiset, eivätkä kestäviä ja aggressiivisia kuten hait. Heikkoina ja helposti haavoittuvina ihmiset selviävät muita eläimiä suuremman älykkyytensä avulla. (Rescher 2001, 6-10). Älykkyys on evolutiivinen etu, koska sen avulla voidaan mallintaa, ennustaa ja manipuloida ympäristöä (Yudkowsky 2002, 1). Tämä koskee myös sosiaalista ympäristöä. Mahdollisuus menestyä maailmankatsomustaan kehittämällä on erityisesti ihmisellä. Kokonainen, koherentti, tarpeiden mukaan rakennettu muokattava maailmankatsomus on tämän suuntaisen kehityksen huippu. Ihannetapauksessa se kiteyttää kaiken olennaisen tiedon ja tiedonkäsittelyn, jonka älykäs systeemi tarvitsee.

Muitakin kuin välittömästi biologiseen evoluutioon liittyviä syitä maailmankatsomuksen tarpeeseen voidaan mainita, vaikka ne voidaan ymmärtää sen pohjalta.

3) Tunteisiin ja elämänlaatuun liittyvät tarpeet: Eheä maailmankatsomus voi auttaa yksilöiden ja yhteisöjen selviytymistä nopeasti muuttuvassa yhteiskunnassa. Yksi suurimpia ongelmia nyky-yhteiskunnassa on muutoksen ja kehityksen aiheuttama psykologinen paine. Tieteellinen ja teknologinen edistys on johtanut monien prosessien nopeutumiseen. Jos uusi menetelmä tai teknologia voi tuottaa saman tuloksen vähemmällä vaivalla, joku tulee ottamaan sen käyttöönsä. Se joka ottaa uuden teknologian käyttöönsä saa kilpailuedun tuottaessaan enemmän pienemmällä sijoituksella. Tämä johtaa tarpeeseen kehittää yhä uusia optimoinnin ja rationalisoinnin menetelmiä. (Heylighen 1998, "Technological acceleration".) Yksilöt ja yhteisöt eivät näytä kykenevän selviytymään jatkuvassa ennustamattomassa muutoksessa ja kasvavassa monimutkaisuudessa. Stressi, epävarmuus ja turhautuminen lisääntyvät kun mielet ylikuormittuvat informaatiosta, tieto on pirstaleista, arvot murenevat ja negatiivisia kehityksiä ylipainotetaan ja positiiviset jätetään huomiotta. Tästä voi seurata nihilismin, ahdistuneisuuden ja epätoivon ilmapiiri (Heylighen 2000: "What is a world view?") Sosiologinen tutkimus näyttää osoittavan, että epävarmuuden ja epäluottamuksen tunteet ovat vahvempia ihmisissä, joilla on vähiten uskonnollisia tai filosofisia maailmankatsomuksia (Elchardus, 1998).

Toiseksi maailmankatsomuksia tutkineet psykologit ovat huomanneet, että tällaiset uskomukset lisäävät hyvinvointia tuottamalla elämään merkityksen tunnetta, toivon ja luottamuksen tunteita, pitkän aikavälin perspektiivin elämän vastoinkäymisissä ja tunteen johonkin suurempaan kuulumisesta (Myers, 1993). (Vidal 2007, 9.)

4) Yksilöllisten ja yhteiskunnallisten toimintasuunnitelmien luomiseen liittyvä tarve: Yksilöiden ja yhteisöjen elämänsuunnitelmat tai toimintasuunnitelmat edellyttävät maailmankatsomusta. Hyvien suunnitelmien tarve lisääntyy ympäristön monimutkaisuuden kasvaessa. Kulttuurienvälinen vuorovaikutus, ihmiskunnan suurempi yhdistyminen sekä tieteen ja teknologian edistyminen johtavat siihen, että yksilön elämänsuunnitelma ja aliyhteisöjen suunnitelmat ovat voimakkaammin yhteydessä globaaliin kokonaisuuteen. Elämänsuunnitelman muodostaminen on samalla tullut vaikeammaksi huomioonotettavan kokonaisuuden koon ja monimutkaisuuden lisääntymisen vuoksi. Esimerkiksi ihmiskunnan selviytymiseen tällä planeetalla liittyvät ekologiset ongelmat tulevat enenevästi olemaan jokaisen ongelma. (Aerts, Apostel et al. 2007, 8, 11.)

Vastuullinen toiminta, ympäristön muuttaminen sekä yhteistyö edellyttävät maailmankatsomusta. Saadaksemme tietoa itsestämme, tarpeistamme, päämääristämme ja arvoistamme, rakennamme malleja ympäristöstämme ja itsestämme. Meidän täytyy rakentaa malleja ihmisistä, historiasta, arvoista ja toimintastrategioista ja yhdistää kosmosta, maata ja biosfääriä koskevan tietomme kanssa. Ilman näiden käsitysten integraatiota vastuullinen toiminta näyttää olevan mahdotonta. Koska emme voi vain antaa asioiden kehittyä omalla painollaan, vaan on otettava vastuu maailmastamme, uusi näiden elementtien integraatio on välttämätön. Tämän yrityksen tulee olla kollektiivinen, koordinoitu ja tietoinen. Eettisiä ja poliittisia valintoja valaiseva tieto ihmisistä, luonnosta, historiasta ja yhteiskunnista mahdollistaa kenties tulevaisuudessa ihmisten ottaa kohtalonsa omiin käsiinsä. (Aerts, Apostel et al. 2007, 11.) Tarvitaan viitekehys, joka sitoo kaiken yhteen ja mahdollistaa yhteiskunnan, maailman ja ihmisten paikan maailmassa ymmärtämisen, koska se auttaisi ihmisiä ja ihmiskuntaa tekemään tärkeitä päätöksiä, jotka muovaavat tulevaisuuttamme. (Heylighen 2000: "What is a world view?")

2.3 Maailmankatsomusten rakentaminen

Usein agentit käyttävät maailmankatsomustaan sitä elämänsä aikana muuttamatta, mutta jotkut voivat muuttaa sitä spontaanisti tai suunnitelmallisesti. Yhden asian tai ilmiön muutosta sanotaan sen kehittymiseksi, ja jos se sen ansiosta myös auttaa enemmän jonkin tavoitteen saavuttamisessa, voidaan puhua tältä kannalta edistymisestä. Maailmankatsomuksen muutos on usein osittainen ja koskee useimmiten ympäristön mallia. Ympäristön muutoksesta tai samanlaisena pysyvää ympäristöä koskevien uusien havaintojen perusteella agentti voi muodostaa mielessään uusia käsitteitä ja vastesääntöjä, jolloin maailmankatsomus joiltain osin voi muuttua. Yleensä sitä yritetään muuttaa paremmaksi ympäristöön sopeutumisen eli siinä agentin tavoitteiden saavuttamisen kannalta, jolloin maailmankatsomus edistyy. Usein jotkin osat pysyvät muuttumattomina koko elämän ajan. Eliöillä perimmäiset tavoitteet, selviytyminen ja lisääntyminen eivät luultavasti voi muuttua. Jos käyttäytyminen poikkeaa paljon jostain näiden kannalta paremmasta ja mahdollisesta käyttäytymisestä, voidaan puhua maladaptaatiosta (Ks. Richerson ja Boyd 2006 ja 2006b, 16). Toisaalta esimerkiksi robotti voidaan ohjelmoida kokonaan uudestaan perimmäisiä tavoitteita myöten.

Joillakin, yleensä hyvin yksinkertaisilla agenteilla maailmankatsomus on synnynnäinen ja muuttumaton. Esimerkiksi bakteeri ei muuta maailmankatsomustaan juuri miltään osin riippuen siitä, kykenevätkö bakteerit elämänaikaiseen oppimiseen. Muutos tapahtuu uuden hieman erilaisen sukupolven syntyessä.

Suuri osa maailmankatsomuksista ihmisilläkin rakentuu osittain spontaanisti (eli siihen tietoisesti puuttumatta). Kun aivomme ovat kehittyneet riittävästi, saamme kokemuksia ympäristöstä, havaitsemme suhteita, kategorisoimme, erottelemme ja yleistämme sitä, mitä aistimme havaitsevat. Korvaamme aistikokemukset ja muistot abstrakteilla yleistetyillä käsitteillä ja niiden rakennelmilla. Sovitamme monia käsitteitä yhtenäisiin skeemoihin, ja rakennamme näistä skeemoista käsitteellisiä viitekehyksiä. Uusien käsitteiden muodostaminen ja käsitteellisten viitekehysten muovautuminen hidastuu ikääntymisen myötä (Project Worldview 2007, Worldviews-An Introduction), mutta monissa tapauksissa jatkuu koko elämän ajan Elämänsä aikana ihminen rakentaa hitaasti elementaarikäsitteistä alkaen

korkeamman tason käsitteitä ja vastesääntöjä. Vastesäännöt muodostavat hierarkian konkreeteimmista abstrakteimpiin, jotka vaikuttavat hyvin moniin vasteisiin. (Fritz 2007, Mindscapes.) Käyttäjä testaa ja kehittää sitä jatkuvasti ympäristöstään saamansa palautteen perusteella. Yksilö- ja aliyhteisökohtaisten kokemusten ohella monet tieteen, tekniikan ja etiikan löydöt ja keksinnöt sekä yhteiskuntien ja luonnon tapahtumat sekä kulloinkin vallitseva kulttuuri (yhteisön jäsenten yhteinen tietämys ja tavat (Fritz 2007, members)) vaikuttavat ihmisten maailmankatsomusten muotoutumiseen

Maailmankatsomuksia voidaan myös rakentaa suunnitelmallisesti. Suunnitelma edellyttää jotain tavoitetta, jonka perusteella maailmankatsomuksen suoritusarvoa voidaan arvioida tai mitata. Jotkin maailmankatsomusten rakentamiseen erikoistuneet yksilöt tai ryhmät voivat tarjota maailmankatsomuksia halukkaille. Mahdollisesti rakennettavien maailmankatsomusten laatua voidaan ja tulee arvioida ja valvoa joidenkin haittojen vähentämiseksi.

Kenen tai minkä tahon tehtävänä on rakentaa maailmankatsomuksia? Periaatteessa kuka tahansa kykenevä voi sen tehdä. Mille taholle tämä tehtävä on ensisijainen, ja/tai mikä tuottaa parhaan tuloksen? Tiede ja filosofia ovat hyviä ehdokkaita tähän, koska niissä ylläpidetään kokeellisen tutkimuksen ja objektiivisen, sekä hyvin järjestetyn tiedon ihanteita. Nämä ihanteet eivät ole satunnaisesti eivätkä tietyn edun takia valittuja, vaan perustuvat niiden avulla tuotetun tiedon ylivertaisen tehokkaaseen ja laajaan käyttökelpoisuuteen. Tällainen tieto auttaa mahdollisimman hyvin pääsemään tavoitteeseen, ja se on monikäyttöistä.

Maailmankatsomusten kannalta tieto on tärkeää, koska sen avulla voidaan rakentaa parempia maailmankatsomuksia. Tieteen ja filosofian tavoite on sama: hankkia tietoa. Perimmältään tietoa hankitaan jotta voitaisiin tehdä parempia ennusteita. Ennusteita tehdään jotta voitaisiin selviytyä, ja ylipäänsä menestyä (saavuttaa tavoitteet). Tiedon pätevyyden kannalta varmin tapa hankkia tietoa on kokeellinen tutkimus, koska tällöin tieto on jo aluksi ainakin jossain yhteydessä havaintoaineistoon. Aina ja kaikissa tapauksissa tämä ei ole mahdollista. Tällöin täytyy turvautua muihin menetelmiin. Toiseksi paras tapa on kokeellisen tutkimuksen tuloksiin ja hallittuun ajatteluun (joihinkin ajattelua koskeviin sääntöihin, esim. johonkin monista nykyisistä logiikoista) perustuva spekulaatio. Suuri osa filosofiasta käyttää tätä

menetelmää. Kolmanneksi paras on hallittu ajattelu, joka ei perustu kokeellisen tutkimuksen tuloksiin. Usein filosofit sortuvat tähän. Huonoin on perusteettomien ja koettelemattomien uskomusten ylläpitäminen. Uskonto kuuluu tähän. Tästäkin voi joskus olla hyötyä, erityisesti jos uskomukset ovat jostain syystä sopivia ja ympäristö ei muutu.

Tällä hetkellä ei ole tieteenalaa, jonka tehtävänä on muodostaa näin laajoja tietojärjestelmiä, mutta asiantila voi muuttua ajan kuluessa. Suurta osaa maailmankatsomuksen rakentamiseen tarvittavasta tiedosta ei tällä hetkellä voida hankkia pelkästään tieteellisen tutkimuksen menetelmällä, mikä kuitenkin on ihanne, ja joudutaan käyttämään myös toiseksi parasta eli filosofian menetelmää. Siksi nykyään maailmankatsomusten rakentaminen on ennen kaikkea tieteen ja filosofian yhteinen tehtävä.

Jos tieteilijät ja filosofit eivät rakenna maailmankatsomuksia, muut kulttuurin tekijät ottavat hyödyn tästä tilanteesta ja tuottavat sekä tarjoavat omia vastauksiaan. Näitä voivat olla esimerkiksi uskonnot, vaarallisemmassa muodossa kultit, äärimmäisiä mielipiteitä kannattavat ideologiat tai fundamentalistisia tulkintoja uskonnoista levittävät tahot (Vidal 2007, 9.) Mitä ovat tiede ja filosofia? Mihin niiden suurempi luotettavuus perustuu?

2.3.1 Tiede

Tiede on muuttuva ja kehittyvä instituutio, ja käsitykset sen menetelmistä ja tavoitteista ovat historian aikana vaihdelleet (Niiniluoto ja Haaparanta 1998, 8) ja vaihtelevat edelleen. Antiikissa ja keskiajalla tiede oli saavutettujen totuuksien järjestämistä, opettamista ja puolustamista. Tieteen menetelmällä tarkoitettiin tätä toimintaa auttavaa tiedon esittämisen menetelmää. Renessanssiaikana tieteen tehtäväksi asetettiin lisäksi myös uuden tiedon tavoittelu tutkimustyön avulla, ja tieteen metodissa tehtiin ero "tutkimisen menetelmän" ja "esittämisen menetelmän" välillä. (Niiniluoto ja Haaparanta 1998, 1.) Nykyään tieteellä tarkoitetaan *tieteellisen tutkimuksen tuloksia*, eli luontoa, ihmistä ja yhteiskuntaa koskevien tietojen systemaattista kokonaisuutta, ja *tieteellistä tutkimusprosessia*, eli tällaisten tietojen systemaattista tavoittelua. (Niiniluoto 1980, 13.) Tutkimus- ja kehittämistoiminta voi olla:

1*)* Perustutkimusta, joka on uuden tieteellisen tiedon etsintää ilman ensisijaista pyrkimystä nimenomaisiin käytännöllisiin tavoitteisiin tai sovellutuksiin.
2) Soveltavaa tutkimusta, joka on tiettyyn käytännön tavoitteeseen tai sovellutukseen tähtäävää perustutkimuksen tuloksille perustuvaa tiedon etsintää. Kehittämistyö on toimintaa, jonka tavoitteena on tutkimustulosten avulla saavuttaa uusia tai paranneltuja tuotteita, tuotantovälineitä tai -menetelmiä ja palveluja. (Niiniluoto 1980, 13.)

Niiniluodon ja Haaparannan määritelmän mukaan tiede on järjestelmällistä ja järkiperäistä uuden tiedon hankintaa. Tiede eroaa muista uskomusten muodostamisen tavoista järjestelmällisyytensä ansiosta. Tieteellä on menetelmä. Tieteellinen tiedonhankinta on organisoitu yhteiskunnassa erityisten instituutioiden tehtäväksi (yliopistot, korkeakoulut, tutkimuslaitokset), ja tutkimustoiminnan tulokset kootaan tietojärjestelmiksi. Tieteellisen tiedon perustelu ei voi nojautua yksilölliseen vaistoon tai intuitioon, suostutteluun, propagandaan, tai jumalalliseen ilmoitukseen, vaan sen on tapahduttava tiedeyhteisön hyväksymän tutkimusmenetelmän avulla. Vaatimus uudesta tiedosta sulkee tieteen käsitteen ulkopuolelle toiminnot, joissa on kyse vain aikaisemmin hankittujen tietojen omaksumisesta (esim. opiskelu) ja järjestämisestä (esim. komitea- ja selvitystyö). Tiede on toimintaa, jonka tavoitteena on tiedon hankkiminen. Tämä erottaa tieteen muista toiminnan muodoista, joilla on erilainen tavoite: urheilu, taide, tekniikka, maanviljely, kaupankäynti, politiikka. (Niiniluoto ja Haaparanta 1998, 7-8.)

Toisaalta Niiniluodon mukaan on epäilyttävää, voidaanko tieteen "olemusta" kiteyttää johonkin määritelmään, mutta on kuitenkin ominaisuuksia, jotka näyttävät erottamattomasti kuuluvan järkevään tiedekäsitykseen. Näitä ovat 1) objektiivisuus, 2) kriittisyys, 3) autonomisuus ja 4) edistyvyys:
1) Tieteellisten teorioiden tulee olla intersubjektiivisesti testattavissa, ts. niistä tulee voida johtaa seurauksia, joiden totuus tai epätotuus voidaan julkisesti tarkistaa. Ollakseen tieteellistä tutkimuksen tulee olla objektiivista ainakin kahdessa mielessä. Tutkimuskohteena olevan todellisuuden osan olemassaolon ja ominaisuuksien on oltava riippumattomia tutkijan mielipiteistä ja toiveista. Toiseksi tutkimuskohteen on myös annettava vaikuttaa

tutkimustuloksen muotoutumiseen, ja tämän vaikutuksen tulee olla intersubjektiivisesti todettavissa.
2) Tieteen tuloksiksi voidaan hyväksyä vasta tieteellisessä yhteisössä käydyn kriittisen keskustelun tulokset. Tieteellinen tulos voidaan aina ja vapaasti asettaa kyseenalaiseksi.
3) Tiede on itseään korjaava siinä mielessä, että tieteen tulosten korjaaminen on yhteisön oma asia, johon tieteenulkoiset ryhmät eivät saa vaikuttaa. Tieteellisten väitteiden perusteleminen ja testaaminen perustuu kriteereihin, jotka koskevat väitteiden tiedollista pätevyyttä, ja vetoaminen siihen, että väitteen totuus olisi esimerkiksi poliittisesti, uskonnollisesti tai moraalisesti toivottavaa, ei ole tieteen kannalta asianmukainen peruste.
4) Tieteen edistyminen on sitä, että virheellisiä teorioita korvataan uusilla, jotka ovat tosia tai ainakin vähemmän virheellisiä kuin aikaisemmat. Tässä mielessä tieteellisen tutkimuksen tulosten - silloinkin kun ne ovat puutteellisia tai epätäsmällisiä - voidaan sanoa ainakin lähestyvän totuutta. (Niiniluoto 1984, 23- 29.)

Agentti-ympäristö -viitekehyksen näkökulmasta *tiede on oikeita ennusteita tuottavien käyttökelpoisten teorioiden rakentamista, joka perustuu kokeelliseen tutkimusmenetelmään.* Tiedon tarkoitus on auttaa älykästä systeemiä ennustamaan ympäristön ilmiöitä, ja tähän perustuen käyttäytymään tavoitteidensa kannalta mahdollisimman menestyksekkäästi. Tieteellinen menetelmä (engl. scientific method) on sarja toimenpiteitä, joiden tavoitteena on tuottaa yksimielisyys havaitsijoiden joukossa, esimerkiksi: 1. Määritellään ongelma, 2. kerätään asiaankuuluva data, 3. muodostetaan työhypoteesi tai selitys, 4. tehdään kokeita hypoteesien testaamiseksi, 5. tulkitaan tulokset, 6. vedetään johtopäätökset ja muokataan hypoteesia tarvittaessa (Heylighen 2008, scientific method). Tiedon pätevyyttä voidaan arvioida sen perusteella, kuinka menestyksekkäästi sen avulla voidaan tehdä ennusteita sen kohteena olevasta ilmiöstä. Tiedon määrää voidaan arvioida objektiivisesti sen perusteella, paljonko merkkejä sen ilmaiseminen jollakin standardikielellä vie, suhteutettuna ilmaisun tehokkuuteen, eli siihen, mikä on ilmaisun pituus suhteutettuna kohteen laajuuteen ja monimutkaisuuteen. Pätevyyttä voidaan mitata sen perusteella, kuinka monessa prosentissa tapauksista tiedon avulla muodostettu ennuste osuu oikeaan. Tiedon pätevyyttä kuvataan joskus käsitteellä "totuus". Haluttaessa se ehkä voitaisiin määritellä tämän käsityksen mukaan teorian tai mallin täydelliseksi pätevyydeksi. Tällöin tieteellinen teoria on tosi, jos se ennustaa

kohteensa käyttäytymisen aina oikein. Tämän testaaminen ei kai ole periaatteessa mahdollista, joten "totuus" on aika abstrakti käsite. Se on myös epäkäytännöllinen käsite vaikean määriteltävyytensä takia, ja kun jonkinlainen määritelmä on onnistuttu tekemään, kuten edellä, se ei siltikään ole kovin käytännöllinen, koska suuri osa käytettävästä tiedosta vaatii hienostuneempaa pätevyyden kuvausta. Lisäksi haittana voi olla sen mahdollisesti mukanaan tuomien turhien metafyysisten sitoumusten painolasti.

Tieteen yhtenäisyys, eli hankitun tiedon mahdollisimman tiivis, yhtenäinen esitys on luonteva tieteen tavoite. Tieteessä pyritään ilmaisun tehokkuuteen ja jo pelkästään tämä voi johtaa tieteellisen ilmaisun mahdollisimman yhtenäiseen esitykseen. Tieteiden ykseyden eli konsilienssin tavoite tarkoittaa yritystä koota tieteen tulokset yhtenäiseksi järjestelmäksi. Sen kannattajia ovat olleet mm. 1800-luvun lopun naturalistit, joita innoitti Charles Darwinin evoluutioteoria, John Deweyn johtamat pragmatistit, jotka halusivat kumota jyrkän erottelun luonnon ja kulttuurin välillä, 1800- ja 1900-lukujen vaihteessa kansainvälinen monistiliike, joka sai vaikutteita comtelaisesta positivismista ja darwinistisesta evoluutioteoriasta. Lähempänä nykyaikaa Wienin piiri julistautui 1929 "tieteellisen maailmankäsityksen" kannattajaksi. Vuoden 1931 jälkeen suosiota sai Otto Neurathin muotoilema fysikalismi, joka pyrki palauttamaan kaiken mielekkään kielenkäytön havaittavista fysikaalisista ilmiöistä puhuvaan yhtenäiskieleen. Neurathin ja Rudolf Carnapin toimesta vuodesta 1935 ryhdyttiin järjestämään Tieteen Ykseyden kansainvälisiä konferensseja, ja 1938 aloitettiin Neurathin toimittama kirjasarja International Encyclopedia of Unified Science, jossa julkaistiin 20 nidettä kunnes Yhdysvaltoihin siirtyneen liikkeen toiminta kuihtui. Lähempänä nykyaikaa mm. Edward O. Wilson puhuu tieteen ykseyden puolesta teoksessaan *Consilience: The unity of knowledge* (1998), Baconin ajatuksia, darwinismia ja nykytieteen kehitystrendejä yhdistellen. (Niiniluoto 2003, 117- 119.)

Tieteiden ykseydestä voidaan puhua monella eri tasolla: historiallisella, ontologisella, kielellisellä, metodologisella tai yhteistyön tasolla ja vahvimmat ykseydet yhdistävät kaikki nämä tasot toisiinsa (Niiniluoto 2003, 117). Esimerkiksi teesin perustana voi olla ontologinen ajatus, että kaikki on yhtä ja samaa perusainesta. Tällaisia ns. monistisia oppeja ovat materialismi, jonka mukaan todellisuus koostuu aineellisista hiukkasista tai prosesseista, ja

idealismi, jonka mukaan todellisuus on henkeä tai ajattelua. Myös naturalistit voidaan lukea monisteihin, sillä heidän mukaansa kaikki on luontoa, eikä ole mitään jyrkkää rajaa aineellisen luonnon ja kulttuurin, ruumiin ja hengen, eläinten ja ihmisten välillä. Tieteiden ykseyttä kannattavia naturalisteja ovat mm. Comte, Mill Spencer, Mach, Haeckel ja Ostwald. Nykyisin tieteiden ykseyden puolustajat ovat yleensä erään materialismin muodon, fysikalismin kannattajia. Kari Enqvist määrittelee fysikalismin kannaksi, jonka mukaan "kaikki on viime kädessä fysiikkaa" (Enqvist 1998). (Niiniluoto 2003, 121- 124.) Perustana voi olla myös ajatus universaalikielestä, jonka avulla mikä tahansa asia voidaan ilmaista symbolien avulla, ja siihen liittyvät ongelmat ovat ratkaistavissa näitä symboleja manipuloivan mekaanisen päättelyn avulla. Raimundus Lullus esitti 1400-luvulla ajatuksen universaalikielestä, sen muotoilivat eri tavoilla 1600-luvulla Comenius ja Leibniz. Moderni logiikka ja digitaalisten tietokoneiden binäärikoodi ovat tavallaan toteuttaneet universaalikielen ja siihen liittyvän kalkyylin tavoitteen, mutta nykyään tiedämme, että loogiselle totuudelle ei ole olemassa algoritmista ratkaisumenetelmää edes elementaarisen logiikan tasolla, ja että formaalikielen looginen totuus on eri asia kuin reaalitieteiden tavoitteena oleva faktuaalinen totuus. (Niiniluoto 2003, 124.)

Nykyään yleinen systeemiteoria ja kybernetiikka tähtäävät tieteen universaaliin kieleen. Esimerkiksi jotkut yleiset käsitteet kuten palaute tai itseorganisaatio soveltuvat yhtä hyvin fysiikkaan, kemiaan, biologiaan, psykologiaan, sosiologiaan, jne... (Vidal 2007, 24.) Recher argumentoi tämän pyrkimyksen puolesta parhaimpana strategiana ja kognitiivisena menetelmänä. Yksinkertaisuuteen ja systemaattisuuteen ei tähdätä siksi, että oletettaisiin luonnon olevan sellainen, vaan koska tämä on tehokkain tapa tehdä tutkimusta. (Rescher 2001, 202) Vaikka tiedon määrän räjähdysmäinen lisääntyminen on tosiasia, joitakin tapoja vastustaa tätä tilannetta voidaan esittää. Ensinnäkin systeemiteoria tarjoaa yleisiä käsitteitä, joita voidaan soveltaa kaikilla tieteenaloilla. Tämän käsitteistön käyttö opittuaan meillä voi olla pääsy kaikkeen tieteelliseen tietoon, vaikkakaan niiden spesifeihin tiivistämättömiin yksityiskohtiin, niin ainakin pääperiaatteisiin. Toiseksi maailmankatsomuksen rakentaminen on iso projekti, ja tarvitaan yhteistyötä tutkijoiden välillä. Niinpä sekä yksilöllisesti että kollektiivisesti tarvitaan informaation valtavan määrän vuoksi parempaa organisointia. Yleiset ja yhteiset käsitteet auttavat tässä. (Vidal 2007, 25.) Tieteiden ykseys voi auttaa

muodostamaan laajoja yhtenäisiä tietämyskantoja, mikä on hyödyllistä maailmankatsomusten kannalta.

Luultavasti tiede jo olemuksestaan johtuen lähestyy yhtenäistä kaikenkattavaa mallia tai teoriaa maailmasta, riippumatta siitä, tavoitellaanko tätä. Muutamat tieteen ihanteet tai usein hyvänä pidetyt ominaisuudet ohjaavat käytännössä tieteellisen tiedon esitystä tiiviiseen ja yhtenäiseen muotoon. Eri variaatioissaan usein ajatellaan, että teoria T2 on parempi kuin T1, jos siitä voidaan päätellä useampia päteviä havaintolauseita (Lakatos ja Musgrave 1970, Kuipers 2000). Tämä tavoite tuntuu johtavan laajempiin teorioihin ja koska pyritään yhtenäisyyteen, niin myös laajempiin tietokantoihin. Tieteellinen selittäminen on eri ilmiöiden välisten yhteyksien kuvailua. Menestyksekkäät selitykset paljastavat ennen erillisinä pidettyjen ilmiöiden välisen yhteyden. Tiede edistää ymmärrystämme luonnosta näyttämällä, kuinka johtaa monien ilmiöiden kuvauksia käyttämällä samoja päättelysääntöjä uudestaan ja uudestaan, ja vähentää näin perustavimmiksi hyväksyttävien faktojen määrää (Kitcher 1989, 423). Joidenkin mukaan tieteen edistys tarkoittaa lisääntyvää ongelmanratkaisua ja ilmiöiden kontrollia tieteen mahdollistamien sovellusten kautta (Recher 1977). Usein ongelmien ratkaisemiseksi tarvitaan tiedon yhdistämistä. Tiede edistyy ja karkeasti tämä voidaan todeta siitä, että sen avulla voidaan nykyään tehdä enemmän asioita kuin ennen. Tieteessä jos teoriat ovat havaintoaineiston alideterminoimia, tällöin usein kehoitetaan valitsemaan yksinkertaisin evidenssin kanssa yhteensopiva teoria (Foster ja Martin 1966). Yksinkertaisuus voi olla paitsi esteettinen kriteeri, myös kognitiivinen; se voi auttaa ymmärtämään maailmaa kognitiivisesti ekonomisesti (Niiniluoto 2008). Käsitykset tieteen yhtenäisyydestä ovat merkityksellisiä tieteessä ja filosofiassa. Tieteessä ne tuottavat heuristista ja metodologisia opastusta ja oikeutusta hypoteeseille, projekteille, ja tavoitteille sekä rahoitukselle ja koulutukselle. Filosofiassa oletukset yhdistämisestä auttavat valitsemaan kysymyksiä ja tutkimusaiheita. (Cat 2008.)

2.3.2 Filosofia

Alun perin filosofia oli kaikenkattava yritys ymmärtää itseämme ja ympäristöämme. Myöhemmin tutkimus jakaantui erityistieteisiin. (Fritz 2007, The history of philosophy.)

Nykyään työnjako tieteen kanssa on seuraava: tavoite on sama kuin tieteenkin: tiedon hankkiminen, mutta filosofiassa tutkitaan ongelmia, joita ei tähän aikaan voida tutkia tieteen eli kokeellisen menetelmän keinoin. Tieteessä (teorian) väitteet perustellaan viime kädessä kokeilla. Filosofiassa väitteet perustellaan järjellä. Järkevä ajattelu perustuu kokeellisen tutkimuksen tuloksiin (yleisemmin: parhaaseen käytössä olevaan tietoon) ja käyttökelpoisiin, yleisesti hyväksyttyihin päättelysääntöihin. Päättelysäännöt ovat myös eräänlaisen kokeellisen tutkimuksen tuloksia; ne ovat sellaisia kuin ovat, koska tähän mennessä parempiakaan ei ole kyetty löytämään. Ihmisen kognitiiviset perustoiminnot (luultavasti myös yleisimmät päättelyprosessit) ovat evoluution tuotetta. Mikään ei takaa, että ne olisivat parhaita mahdollisia. Filosofia ei ole tiedettä, koska sen tutkimus ei perustu tieteen menetelmään. Se ei myöskään ole uskontoa eli perusteettomista uskomuksista kiinni pitävää ajattelua, koska filosofiset väitteet täytyy perustella. Filosofia perustuu tieteen tuloksiin, mutta jatkaa niistä spekulaation ja kriittisen ajattelun keinoin - tällä hetkellä - tieteen menetelmällä tutkimattomien ongelmien ratkaisuun. Tulostensa luotettavuuden kannalta filosofia sijoittuu tieteen ja uskonnon väliin, tieteen ollessa luotettavin.

Filosofia on kirjaimellisesti ottaen maailman ensimmäinen akateeminen oppiala. Sen asema vuosisatojen varrella on vaihdellut "tieteiden kuningattaresta teologian palkkapiiaksi ja tieteiden äidistä tieteiden aputyöläiseksi". Nykyisin filosofialla on kaikki erityistieteen ulkoiset tuntomerkit: vakiintunut asema akateemisena oppiaineena, omat julkaisusarjat, tieteelliset seurat, kongressit ja kansainvälisen yhteistyön organisaatiot. Tässä suhteessa filosofit muodostavat tiedeyhteisön siinä missä muidenkin alojen tutkijat. (Niiniluoto 1984, 59, 60.)

Filosofian tavoite:
1) Filosofian tehtäviin kuuluu tarkastella sellaisia kysymyksiä, joita mikään erityistiede ei - ainakaan vielä - tutki. Se pyrkii hahmottamaan ja jäsentämään sellaisia todellisuuden alueita, joihin järjestelmällinen tutkimus ei ole vielä saanut pysyvää jalansijaa. Tällä tavoin se on 2500 vuoden ajan toiminut kantatieteenä, joka on synnyttänyt uusia tieteitä, ja tämä prosessi jatkuu yhä, nykyään mm. logiikan ja kielen tutkimuksen alueilla. Filosofian tehtävänä on ollut kulkea muiden tieteiden edellä ja ulottaa rationaalista ajattelua aikaisemmin koskemattomille alueille. (Niiniluoto 1984, 60, 61.)

2) Toiseksi filosofia ylittää yksityiset tieteenalat siinä mielessä, että se pyrkii rakentamaan niiden tulosten pohjalle kokonaisvaltaisia maailmankatsomusten järjestelmiä. Erityistieteiden tulosten summa jollain hetkellä ei vielä riitä muodostamaan maailmankatsomusta, vaan niiden yhteensitomiseksi tarvitaan kokonaiskuvaa todellisuuden luonteesta, tiedon hankkimisen edellytyksistä sekä ihmisen tehtävistä maailmassa. (Niiniluoto 1984, 61.)
3) Kolmanneksi filosofian kriittisenä tehtävänä on inhimillisen kulttuurin kaikkien muotojen ennakkoedellytysten, sisällön ja vaikutusten erittely ja arvostelu. Näihin kulttuurin muotoihin kuuluvat mm. kielet ja käsitejärjestelmät, aatteet ja uskonnot, tieteet ja taiteet, ja itse filosofiakin. (Niiniluoto 1984, 61.)

Filosofian menetelmä:
Filosofian menetelmiä näyttää olevan yhtä paljon kuin filosofioitakin (Passmore 1967). Karkeasti sanoen esimerkiksi Platonille ja Hegelille filosofian menetelmä on dialektiikka, Bergsonille intuitio, Wittgensteinille merkityksettömien lauseiden paljastaminen, Husserlille fenomenologinen kuvailu, Humelle kokeellisen tutkimusmenetelmän seuraaminen, Spinozalle geometristen menetelmien käyttö, jne. (Vidal 2007, 5). Miksi näin on? Körner (1969, 20) ehdottaa, että löytäessään hedelmällisen menetelmän filosofeilla on taipumus laajentaa sen käyttöä, väittää sen olevan ainut oikea filosofian menetelmä, ja jopa määritellä filosofia tämän menetelmän käytön perusteella. Koska filosofit käyttävät eri menetelmiä ja sulkevat toisia pois filosofian piiristä, on vaikea muodostaa yhtenäistä käsitystä filosofian menetelmistä. Joillakin filosofian alueilla käytetään hyväksi joidenkin tieteenalojen menetelmiä: loogikot soveltavat eksakteja matemaattisia käsitteenmuodostustapoja ja menetelmiä, ja filosofian historian tutkijat käyttävät lähdekritiikin ja tekstien tulkinnan menetelmiä. Varsinaisiin uusiin filosofisiin tuloksiin pyrittäessä sovelletaan tavallisesti menetelmää, jonka vaiheita voidaan luonnehtia nimillä problematisointi, eksplikointi ja argumentaatio:
1) Problemaatio. Lähtökohtana voi olla ongelma, joka ei ole tällä hetkellä tieteellinen ongelma.
2) Eksplikaatio, filosofisen ongelman vastausyritysten muotoileminen. Tähän kuuluu kysymysten täsmentäminen, käsitteiden analysointi ja määritteleminen ja uusien ajatusten esittäminen.
3) Argumentaatio, vastausyritysten vertailu ja arvostelu. Relevantteja kysymyksiä ovat tässä yhteydessä mm.: Onko esitetty ratkaisu ristiriidaton, selkeä ja yhteensopiva muiden pätevien

filosofisten teesien kanssa? Onko se vastaus esitettyyn ongelmaan? Kykeneekö se mielenkiintoisella tavalla käsittelemään muita samankaltaisia ongelmia? Esitetyt argumentit voivat puolestaan johtaa uusien filosofisten ongelmien heräämiseen, näiden ratkaisuyrityksiin jne. (Niiniluoto 1984, 62- 63.)

Filosofisen työn arviointi:
Filosofista työtä voidaan arvioida mm. sen uutuuden, selkeyden, tarkkuuden, järjestelmällisyyden, syvällisyyden ja hedelmällisyyden perusteella. Nämä kriteerit ovat julkisia ja objektiivisia (Niiniluoto 1984, 65- 67). Teoria, riippumatta siitä, onko se tieteellinen tai filosofinen, on rationaalinen siinä määrin kuin se ratkaisee tiettyjä ongelmia. Jos teoriaa tarkastellaan ratkaisuehdotuksena tiettyihin ongelmiin, sitä voidaan tarkastella kriittisesti, vaikka se ei olisikaan testattavissa ja falsifioitavissa: Voidaan kysyä: Ratkaiseeko se ongelman? Ratkaiseeko teoria sen paremmin kuin muut? Onko se ehkä muuttanut ongelman? Onko ratkaisu yksinkertainen? Onko se hedelmällinen? Onko se ristiriidassa toisten filosofisten teorioiden kanssa, joita tarvitaan toisten ongelmien ratkaisuun? Tällaiset kysymykset osoittavat, että kriittinen keskustelu on mahdollista, vaikka teoriat eivät olisikaan niiden esityshetkellä kokeellisesti testattavissa. (Popper 1958, 269.)

Filosofian edistyminen:
Filosofiassa tapahtuu aitoa edistystä: Siinä voidaan saavuttaa pysyvää arvoa omaavia tuloksia, jotka voivat olla hedelmällisiä kysymyksenasetteluja, ongelmien jakoja osaongelmiksi, selventäviä käsitteellisiä erotteluja, vakuuttavia päätelmiä sekä sisällöllisesti uusia ajatuksia. Filosofian edistyvyyttä voidaan puolustaa lisäksi mm. seuraavasti: 1) Filosofiassa voidaan saavuttaa varsin lopullisia negatiivisia tuloksia, esimerkiksi osoittamalla jokin intuitiivinen käsitys tai oletusten joukko ristiriitaiseksi. 2) Filosofiassa voidaan saavuttaa pysyviä tuloksia, jotka ovat ehdollista muotoa. Filosofiset johtopäätökset esitetään joskus aukottomassa muodossa näyttämällä, että joistakin oletuksista P seuraa loogisesti johtopäätös Q. Johtopäätös Q voidaan kuitenkin kiistää asettamalla oletukset P kritiikin kohteeksi, sillä mitään oletuksia ei voi dogmaattisesti hyväksyä filosofisen ajattelun lähtökohdiksi. Todistetuksi voi silti tulla ehdollinen väite "jos P niin Q". (Niiniluoto 1984, 67, 68.) Mielestäni filosofian yhteydessä voidaan puhua kolmenlaisesta edistymisestä: Joku filosofi voi edistyä omassa tutkimuksessaan.

Filosofilla on tietyt tavoitteet ja hän edistyy niissä ainakin niiltä osin, kuin se koskee tätä tutkimusta. Tätä tapahtuu nykyään ja on tapahtunut myös historiassa. Filosofia alana voi edistyä, jos tutkijoilla on yhteinen tavoite ja he saavuttavat tätä tutkimuksensa avulla. Koko filosofia tutkimuksen alana on laaja ja vaikeasti määriteltävä (ei ole selvää, onko tai pitäisikö filosofialla alana tai sen tutkijoilla olla yhteinen tavoite). Kapeammassa merkityksessä jotkut filosofit voivat muodostaa tutkijayhteisön (esim. koulukunnan, tutkimussuunnan), jolla on ainakin osittain yhteinen tavoite. Tällöin filosofian osa-alue tai tutkimussuunta edistyy. Kolmanneksi tieto voi lisääntyä. Edellä esitettiin, että filosofiankin tavoite perimmiltään on tiedon hankkiminen. Luotettavin tiedon lähde on tieteellinen tutkimus, joten filosofit voivat yrittää jouduttaa prosessia, jossa filosofian alaan kuuluva tutkimuskohde mahdollisesti siirtyy tieteelliseksi ongelmaksi. Tätä voidaan tehdä selventämällä ja kehittelemällä kohteeseen liittyviä käsityksiä, muodostamalla alustavia teorioita ja hypoteeseja sekä kehittelemällä ajatuksia siitä, kuinka kohdetta voidaan tutkia.

Usein käy myös niin, että filosofiset ongelmat eivät ratkea ollenkaan, ainakaan aivan siinä muodossa kuin ne on esitetty, vaan muuttuvat epärelevanteiksi. Esimerkiksi ongelma: ”Onko jumalaa olemassa?” ei ole ratkaistu, mutta verrattuna keskiaikaan, sen merkittävyys on vähentynyt. Ilmeisesti koska on ymmärretty, että ongelmaa ei edes periaatteessa voida omin keinoin ratkaista (periaatteellinen falsifioimattomuus), ja elämä sujuu ilman tätä ratkaisua, on keskitytty muihin asioihin. Tietoteoriassa kiistattomasti parasta menetelmää tiedon hankkimiseksi ei ole löydetty. Tieteellinen menetelmä on eräs vaihtoehdoista. Se on menestynyt hyvin ja jos menestys jatkuu edelleen, ongelma menettää merkittävyyttään, paitsi siinä mielesssä, että aina on hyödyllistä löytää jokin vielä parempi menetelmä. Kiistatonta perustetta jonkin eettisen järjestelmän paremmuudesta ei ole löydetty. Kuitenkin jotain eettistä järjestelmää joudutaan käyttämään käytännön pakosta. Ajan mittaan huomataan, että jotkin vaihtoehdot tuottavat muita parempaa elämää, ja niitä aletaan suosia. Vaikka lopullista perustetta näiden puolesta ei löytyisikään, niitä käytetään ja muut vaihtoehdot menettävät merkitystään, kuten myös koko ongelma.

2.3.3 Maailmankatsomuksen rakentaminen tieteen ja filosofian tehtävänä

Perinteisen käsityksen mukaan filosofia on "maailmankatsomusoppia". Wilhelm Jerusalemia lainaten: "Filosofia on ajatustyö, johon ryhdytään tarkoituksessa yhdistää jokapäiväisen elämän kokemukset ja tieteellisen tutkimuksen tulokset yhtenäiseksi ja ristiriidattomaksi maailmankatsomukseksi, joka on omansa tyydyttämään ymmärryksen tarpeet ja järjen vaatimukset." Niiniluodon mukaan Jerusalemin tehtävänasettelu vaikuttaa nykyisin kohtuuttomalta, sillä "kukapa voisi enää hallita kaikkien tieteiden tuloksia". Hänen mielestään on turha odottaa, että filosofia tarjoaisi yhtenäistä ja ristiriidatonta maailmankatsomusta valmiina siistinä pakettina, jonka käyttäjä voisi poimia koriinsa niin kuin tuotteen valintamyymälän hyllyltä. Se "pikemminkin antaa työpöydän, liimaa ja valmistusohjeita niille, jotka haluavat harjoittaa älyllistä askartelutoimintaa tieteiden tarjoamien palikoiden avulla." (Niiniluoto 1984, 5.)

Joidenkin mukaan filosofista tutkimusta vaivaa pirstaloituminen eli yhtenäisyyden puute. Nykyään ns. analyyttinen ja mannermainen filosofia näyttävät olevan filosofian päätrendit, joista analyyttinen toi filosofiaan tehokkaat analyysin ja kritiikin välineet, mutta siltä puuttuu yleinen suunta ja yhdistävä ajatus, ja mannermaisella filosofialla on paljon kiinnostavia lähestymistapoja filosofisiin ongelmiin, mutta eräs sen heikkouksista on kunnollisten menetelmien puuttuminen. Lisäksi nykyään filosofiassa tapahtuu spesialisaatiota; Agenda on levinnyt räjähdysmäisesti ja tässä suuren diversiteetin tilanteessa on olemassa niinkin spesiaaleja aloja kuin esimerkiksi liikunnan ja huumorin filosofiaa. Tutkijat keskittyvät eri ongelmiin, näiden aliongelmiin, alialiongelmiin jne. Niiden ratkaisu on hyödyllistä, mutta yhteys laajempiin kokonaisuuksiin on vaikea selvittää. Ongelmana on myös se, että nykyfilosofia on liian kriittistä spekulaation kustannuksella. (Vidal 2007, 3- 5.) Filosofia, kuten tiedekään, ei ole pelkkää spekulaatiota tai pelkkää kritiikkiä, vaan spekulaatiota, jota kontrolloi kritiikki. Tiede on valtaamassa perinteisen filosofian alaa, ja huomioonottaen tieteiden tähänastisen kehityksen, tämä prosessi ei luultavasti tule hidastumaan (Vidal 2007, 23). Tiede kykenee tutkimaan yhä useampia ongelmia, joita perinteisesti pidetty filosofian alaan kuuluvina. Kehitys voi tuottaa lisää materiaalia ja välineitä filosofian käyttöön. Tieteen

tutkimusalan laajentuminen ei välttämättä tarkoita, että filosofiaa ei tarvita enää. Filosofiaa tarvitaan aina kun on vielä olemassa filosofisia ongelmia, ja niitä, jotka haluavat ratkaista niitä. Tieteen edistyminen johtaa siihen, että myös filosofian tulee määritellä alansa uudelleen alansa ja suhteensa tieteeseen. Filosofien tulisi käyttää hyväkseen uutta tietoa ja tarkastella sen filosofisia seurauksia. (Vidal 2007, 3.) Vaikka filosofia on ei-tiedettä, tämä ei tarkoita, että se olisi epätiedettä. (Broad 1958, 103)

Laajassa mielessä, viitattaessa esimerkiksi inuiittien tai mayojen filosofiaan, filosofia tarkoittaa maailmankatsomusta (saksaksi *Weltanschauung)*. Käsitteellä on pitkä historia (katso: Naugle 2002). Sitä käytetään paitsi filosofiassa, myös mm. teologiassa, antropologiassa ja kasvatustieteessä. (Vidal 2007, 7.) Käsitys filosofiasta maailmankatsomusten rakentamisena tarjoaa selkeän tavoitteen tai tehtävän filosofeille, ja on harmoniassa filosofian alkuperän kanssa. (Vidal 2007, 26.) Wolters (1989) tiivisti maailmankatsomuksen ja filosofian suhteen. Hänen mukaansa "maailmankatsomus kruunaa filosofian", eli maailmankatsomuksen rakentaminen on filosofian päätavoite. Kunnianhimon puutteessa nykyisessä filosofiassa harvoin ehdotetaan tai edes tähdätään koherenttiin ja kokonaiseen maailmankatsomukseen. Kuitenkin maailmankatsomusten rakentaminen on filosofian päätehtävä, joten filosofien täytyy luopua vaatimattomuudestaan, vaikka varovaisuus onkin tarpeellista. (Vidal 2007, 26.)

Broad erotti kolmenlaista filosofista toimintaa: analyysin, synopsiksen ja synteesin. Analyysi on tutkimusta käsitteistä ja niiden välisistä suhteista. Synopsis on yritys tarkastella yleensä erillisinä pidettyjä ihmiskokemuksen aspekteja yhdessä ja nähdä, kuinka ne liittyvät toisiinsa. (Broad 1958, 116.) Synteesin tarkoituksena on tarjota käsitteiden ja periaatteiden joukko, joka tyydyttävästi kattaa eri alueet, joita tarkastellaan synoptisesti (Broad 1958, 126). Filosofinen työ voi olla eri tavoin painottunut näiden suhteen. Humen työ on niin analyyttistä, että sen voidaan kieltää olevan synoptista, ja Hegelin niin synoptista, että sen voidaan kieltää olevan analyyttistä. Broadin mukaan molempia on kuitenkin aina läsnä ja jokaiseen kuluu jonkin verran toisiaan. Synopsiksen ja synteesin välillä on vahva positiivinen korrelaatio. Synteesi edellyttää synopsista, ja laajoja synopseja tekevät yleensä ihmiset, joiden päätavoitteena on synteesi. (Broad 1947.) Filosofit voivat painottaa mitä tahansa näistä kolmesta toiminnasta, mutta Vidal pitää synteesin tekoa suuren filosofian ehtona. Suuret filosofit rakentavat

ajatussysteemejä, jotka vastaavat kaikkiin filosofisiin kysymyksiin. (Vidal 2007, 6.) Filosofia tarvitsee sekä kriittistä tai analyyttistä että spekulatiivista tai synteettistä lähestymistapaa, koska muuten se ei voi edistyä (Vidal 2007, 3-4). Filosofian pitää perustua tieteen tuloksiin ja maailmankatsomuksia luotaessa täytyy löytää tapoja yhdistää tieteellisen tutkimuksen tuloksia laajemmiksi kokonaisuuksiksi. (Vidal 2007, 24.) Tällaisen maailmankatsomuksen täytyy olla muutettavissa tieteen edistyessä. Jos tieteellinen teoria kumotaan, sen filosofiset seuraukset asettuvat kyseenalaisiksi. Tällainen lähestymistapa rajoittaa puhtaasti intellektuaalisia filosofisia konstruktioita pitämällä filosofiset teoriat ajan tasalla. Yleinen ongelma filosofiassa on käsitteellisen systeemin rakentaminen ilman riittävää yhteyttä todellisuuteen, eli liiallinen sisäisen koherenssin painottaminen. Maailmankatsomuksen täytyy olla yhteydessä myös ulkoiseen todellisuuteen, eli olla myös ulkoisesti koherentti. Tällöin filosofiset väitteet olisivat eksplisiittisesti yhteydessä enemmän tai vähemmän pysyviin faktoihin, usein tieteellisten teorioiden kautta. Hyvän systemaattisen filosofian kriteerinä olisi tällöin yhteys tieteelliseen tietoon, ja parhaassa tapauksessa koko inhimilliseen tietämykseen. (Vidal 2007, 24.)

Tietenkään filosofin ei tarvitse sitoutua mihinkään muka filosofiaan kuuluvaan projektiin, vaan hän voi tutkia mitä vain. Filosofiassa täytyy aina olla mahdollisuus kyseenalaistaa kaikki oletukset. Mahdollisimman hyvän maailmankatsomuksen muodostaminen on kuitenkin vanha ja edelleen arvokas projekti filosofiassa.

2.4 Ohje maailmankatsomusten rakentamiseksi

Yleinen ohje maailmankatsomuksen rakentamiseen:

1. Tee synopsis kaikesta, mikä saattaa olla hyödyllistä maailmankatsomuksen kysymyksiin vastaamiseen.
2. Valitse valmiita tai tee itse parhaat käsitteet synteesin luomiseksi tästä synopsiksesta.
3. Ehdota synteesiä systemaattisen filosofian muodossa.
4. Vertaile tuloksena olevaa maailmankatsomusta toisiin maailmankatsomuksiin edellä esitettyjen arviointikriteerien avulla, ja osoita, miksi se on parempi kuin muut.
5. Osoita, kuinka se voi ratkaista aikamme ongelmat.
6. Levitä tätä maailmankatsomusta. (Vidal 2007, 26.)

2.5 Maailmankatsomusten arviointi

Monista eri maailmankatsomuksista voidaan valita paras tai parhaat erilaisten kriteerien perusteella. Kohdetta voidaan arvioida vain jonkin kriteerin tai kriteeristön perusteella, ja jos halutaan arvioida maailmankatsomuksia, on siis valittava arvioinnin kriteeristö, joka on myös perusteltava. On esitettävä, mitä kriteereillä yritetään saavuttaa, miten ne sen tekevät, ja miksi ja kenelle tämä on tavoittelemisen arvoista.

Vidal ehdottaa seuraavaa Rescheriltä (2001, 31) ja Heylighenilta (1997) vaikutteita saanutta kriteeristöä. Myös Herrick ("Appendix: Evaluating Worldviews", 791- 792) ja Totten (2004, Worldview test site) esittävät hyvin samanlaiset listat. Lista on järjestetty. Objektiiviset kriteerit ovat tärkeämpiä kuin subjektiiviset tai intersubjektiiviset. Maailmankatsomus on toista parempi, jos se:

- 1 Objektiiviset kriteerit - valinta tiedon kohteen yhteensopivuuden perusteella.
    - Sopii paremmin yhteen nykytieteen kanssa. Tämä ns. ulkoisen koherenssin vaatimus, joka tarkoittaa, että maailmankatsomus ei saa olla ristiriidassa "tosiasioiden", "todellisuuden" tai empiirisen aineiston kanssa (Vidal 2007, 14).
    - Käsittelee ja ratkaisee laajemman joukon maailmankatsomuksen osiin liittyviä ongelmia. Maailmankatsomuksen tulisi olla kokonainen siinä mielessä, että se sisältää kaikki maailmankatsomuksiin kuuluvat kuusi osaa (Vidal 2007, 14).
    - On sisäisesti koherentimpi. Sisäisen koherenssin vaatimus tarkoittaa, että maailmankatsomuksen osat eivät saa olla ristiriidassa keskenään (Vidal 2007, 14).
- 2 Intersubjektiiviset kriteerit - valinta monien subjektien hyväksynnän perusteella.
    - On helpompi soveltaa ja esittää muille.
    - Kannustaa sosiaalisesti hyödyllisempään elämänkatsomukseen.
- 3 Subjektiiviset kriteerit - valinta yksilöllisen hyväksynnän perusteella.
    - On yksinkertaisempi. Vähemmän erotteluja ja vaatii vähemmän vaivalloisia selityksiä. Periaatteet ovat vähemmän keinotekoisia.
    - Vastaa paremmin tervettä arkijärkeä.

- Kannustaa henkilökohtaisesti palkitsevampaan elämänkatsomukseen. (Vidal 2007, 19, 20.)

Tällä hetkellä mikään tunnettu maailmankatsomus ihmisille ei täytä kriteerejä täydellisesti. Nykyinen tieteellinen maailmankatsomus ei ole kokonainen, koska se ei ainakaan nykyään kykene antamaan vastauksia arvoja ja toimintaa koskeviin kysymyksiin. Esimerkiksi Carvalho (2006, 122) väittää, että tieteellinen maailmankatsomus ei edes voi olla kokonainen. (Vidal 2007, 15.) Toisaalta empiiriset tutkimukset näyttävät osoittavan, että esimerkiksi murha, varkaus, raiskaus, valehtelu, jne... ovat negatiivisia arvoja kaikissa yhteisöissä, ja terveys, varakkuus, ystävyys, rehellisyys, turvallisuus, vapaus, tasa-arvoisuus, jne... ovat positiivisia (Heylighen, Bernheim 2000a). Tällä tavalla tieteessä voidaan selvittää arvoja ja lisäksi soveltavassa tutkimuksessa voitaisiin tutkia, miten arvojen perusteella muodostettujen tavoitteiden saavuttamiseksi voidaan toimia. Tässä mielessä myös tieteellinen maailmankatsomus voisi antaa vastauksia arvoja koskeviin kysymyksiin. Jotkut uskonnolliset maailmankatsomukset taas ovat kokonaisia mutta eivät ole koherentteja (Vidal 2007, 15). Lisäksi uskonnollisilla maailmankatsomuksilla ei ole rationaalista mekanismia ongelmallisten aiheiden käsittelyyn tai erimielisyyksien selvittämiseen. Niillä on vähän vastauksia nykyajan kehitykseen, ja ovat siten ei-sopeutuvia; ongelmallisissa tilanteissa ne taantuvat usein fundamentalismiksi, eli vanhojen tekstien kirjaimelliseksi tulkinnaksi. (Vidal 2007, 23.) Holistiset maailmankatsomukset, esimerkiksi "new age", ovat sumeita, irrationaalisia ja epäkäytännöllisiä. Humanistinen maailmankatsomus on liian antroposentrinen, ihmiskeskeinen; sen tulisi tarkastella ihmiskuntaa laajemmin, myös evolutiivisesta, ekologisesta, kosmologisesta, jne... näkökulmasta. Individualismi on laajalle levittäytynyt arvo, ja se voidaan tulkita maailmankatsomukseksi. Sitä pidetään usein monien ongelmien aiheuttajana yhteisöissä. Toisaalta se voi tarkoittaa, että jokaisella on eri tai oma maailmankatsomuksensa. Äärimmäistapauksessa tämä johtaa siihen, että ei ole yhteisiä arvoja eikä päämääriä. Toisaalta se voi tarkoittaa hyvin kapea-alaista maailmankatsomusta: kuinka tehdä omasta elämästä mahdollisimman hyvä, riippumatta siitä, miten tämä vaikuttaa muihin. (Vidal 2007, 23.)

Mihin maailmankatsomusten arviointikriteerit perimmiltään ja yksityisistä toteutuksista riippumatta perustuvat? Maailmankatsomukset ovat agenttien informaatiosysteemejä, jolla ne käsittelevät ympäristöstään saamaansa tietoa ja yrittävät muodostaa sen ja tavoitteensa perusteella toimintoja joiden suorittaminen auttaa niitä saavuttamaan perimmäisen tavoitteensa. Jos maailmankatsomusten tehtävä on tällainen, maailmankatsomusten hyvyyttä voidaan arvioida sen perusteella, täyttävätkö ne tämän tehtävän, eli auttavatko ne agenttia saavuttamaan tavoitteensa. Siten tavoitteellisen agentin kannalta maailmankatsomus on hyvä, jos se auttaa sitä saavuttamaan perimmäisen tavoitteen. Maailmankatsomus on paras, jos se tekee tämän paremmin kuin muut, ja sen hyvyys on yhtä kuin sen hyödyllisyys käyttäjälleen.

Lisää kriteerejä voidaan muodostaa sillä perusteella, että niiden täyttäminen auttaa maailmankatsomuksen perimmäisen tavoitteen saavuttamisessa. Agentin menestyksen kannalta tärkeitä tekijöitä ovat sen resurssit, ympäristö ja tavoitteet. Ilman tietoa näistä tekijöistä oikeaa toimintoa ei ole mahdollista määrittää. Jos ei tiedetä agentin resursseja, ei tiedetä, mitä toimintoja agentti voi suorittaa, mitä se voi havaita ja miten paljon ja miten monimutkaista tietoa se voi käsitellä. Ilman tietoa ympäristöstä ei tiedetä, mihin tilaan ympäristö siirtyy tietyn toiminnon suorittamisen seurauksena, joten ei voida selvittää, mikä toiminto siirtää ympäristön tavoitetilaan. Ilman tietoa tavoitteista ei voida tietää, mikä ympäristön tila on tavoitetila, joten sitä ei voida löytääkään. Yleisesti agentin resursseja ovat kaikki sen tavoitteen saavuttamisen kannalta hyödylliset asiat. Agentin resurssit voidaan luokitella sisäisiin, välittäviin ja ulkoisiin resursseihin. Sisäiset resurssit ovat hyödyllisiä agentin sisäisiä rakenteita tai prosesseja. Välittävät resurssit ovat agentin ja ympäristön kommunikaatioita välittävien agentin osien (sensorien ja aktuaattorien) hyödyllisiä kykyjä. Ulkoiset resurssit ovat hyödyllisiä ympäristön osia. Agentin maailmankatsomus kuuluu sen sisäisiin resursseihin. Parempi maailmankatsomus on suurempi/arvokkaampi resurssi. Mitkä maailmankatsomuksen ominaisuudet tekevät hyödyllisemmän, eli paremman/arvokkaamman resurssin? Useimmiten nämä:

- Yksinkertaisuus/lyhyys säästää laskennallisia resursseja, kuten prosessointitehoa, ja muistitilaa.
- Sisäinen, ja ulkoinen koherenssi mahdollistavat päättelyn, suunnittelun ym.

- Kehityskelpoisuus/muokattavuus mahdollistaa oppimisen, eli maailmankatsomuksen osien muokkaamisen uusien kokemusten perusteella.

Lisäksi voi olla lukematon määrä erilaisia tapauskohtaisia kriteerejä, jotka määräytyvät tavoitteen, käytössä olevien resurssien sekä ympäristön ominaisuuksien mukaan. Maailmankatsomuksen hyvyys tai arvo on usein tapauskohtaista ja siten sitä voidaan luonnehtia vain hyvin rajoitetussa määrin. Tämä johtuu siitä, että voi olla olemassa hyvin suuri joukko hyvin erilaisia tavoitteita, resursseja ja ympäristöjä. Ehkä voitaisiin jakaa arvo yleiseen ja tapauskohtaiseen, jolloin yleinen arvo viittaisi niihin ominaisuuksiin, jotka ovat kaikissa tapauksissa (tai ainakin suuressa osassa tapauksista) hyödyllisiä, ja tapauskohtainen vain tietyissä. Useimmiten on ilmeisesti käytännössä kokeiltava, millainen toimii milloinkin. Koska maailmankatsomuksia voidaan arvioida ja muuttaa, niitä voidaan optimoida.

3 ÄLYKKÄÄT SYSTEEMIT -MAAILMANKATSOMUS

Älykkäät systeemit -maailmankatsomus perustuu Walter Fritzin kirjassaan *Intelligent systems and their societies* (2007) esittämiin ajatuksiin. Sen peruskäsitteet tulevat eri tieteenaloilta, pääasiassa tekoälystä ja systeemiteoriasta. Fritzin (2007) lisäksi olen kerännyt mielestäni vakuuttavia ja yhteensopivia ajatuksia muista lähteistä, joista tärkeimmät ovat:

- Russell ja Norvig (2003), tekoälyn oppikirja, jonka merkitys on paljon laajempi kuin voisi olettaa. Se antaa tieteellisesti testatun ja formalisoidun käsitteistön kaikkien älykkäiden systeemien älykkyyteen liittyvien prosessien analysointiin. Tekoälytutkimus on nuori tutkimuksen ala, eivätkä tulokset ole lopullisia tai kattavia.
- Shane Leggin ja Marcus Hutterin (2005, 2006, 2007a, 2007b) älykkyyttä käsittelevät kirjoitukset esittävät universaalin, tekoälyn käsitteistölle perustuvan älykkyyden määritelmän. Niiden lähtökohtana on ns. agentti-ympäristö -viitekehys, joka on myös älykkäät systeemit -maailmankatsomuksen perusta.
- Jeff Hawkinsin (2005) aivojen toimintaa käsittelevä kirja, joka antaa uutta ja kattavampaa tietoa verkkoihin perustuvien aivojen (mm. ihmisaivojen) toiminnasta.
- Richersonin ja Boydin (2006a, 2006b) kulttuurievoluutiota käsittelevät kirjoitukset antavat tietoa yhteistyön, yhteisöllisyyden, etiikan, tapojen ja instituutioiden synnystä, entistä tieteellisemmin ja uskottavammin.

Se on tarkoitettu nykyisille ihmisille. Sen hyvyys perustuu sen omaamiin maailmankatsomuksille toivottaviin ominaisuuksiin. Toisin kuin monet muut maailmankatsomukset, se vastaa kaikkiin maailmankatsomuksen osia koskeviin kysymyksiin, ei kuitenkaan täydellisesti tai lopullisesti. Se on kokonainen (vastaa kaikkiin maailmankatsomuksen osia koskeviin kysymyksiin), koherentti (tieteellinen, systeemiteoreettinen kieli), ja kehityskelpoinen (avoin ehkä kieltä, tutkimuksen menetelmää sekä joitakin peruskäsitteitä lukuun ottamatta ja uuden tiedon hankintaa hyödyllisenä pitävä).

3.1 Älykkäät systeemit -maailmankatsomuksen taustaa

Johdantona tähän maailmankatsomukseen esittelen sen tärkeän osan, ns. agentti-ympäristö -viitekehyksen sekä tiiviisti eniten maailmankatsomukseen antaneet tutkimusalat, ja joitakin sen kehitykseen vaikuttaneita tai muuten sille läheisiä ongelmia.

3.1.1 Agentti-ympäristö -viitekehys

Maailmankatsomuksen keskeinen osa on ns. agentti-ympäristö -viitekehys. Tämä laajasti käytetty ja joustava struktuuri, jota oletettavasti jokainen käyttää ja ymmärtää jossakin muodossa, vähintään intuitiivisesti, ei sinällään ole uusi keksintö. Tekoälytutkimuksessa sitä käytetään mm. vahvistusoppimisen viitekehyksenä (Sutton, R. ja Barto, A. 1998), ja eri tavoin nimettynä kontrolliteoriassa (Bertsekas, D. P. ja Tsitsiklis, J. N. 1996). Viitekehys tuli tunnetuksi, kun Norbert Wiener (1894- 1964) kollegoidensa Arturo Rosenbluethin ja Julia Bigelowin kanssa haastoivat behavioristisen ajattelutavan tarkastellen käyttäytymistä nykyisen ja tavoitetilan välistä virhettä säätelevänä systeeminä. Wienerin kirjasta Kybernetiikka (1948) tuli bestseller, joka myös kiinnitti yleisön huomion mahdollisuuteen luoda älykkäitä koneita (Russel ja Norvig 2003, 15). Viitekehys sisältää kolme olennaista komponenttia: agentin, ympäristön ja tavoitteen. Agentti on mikä tahansa, mitä voidaan tarkastella sensoreillaan ympäristöään havainnoivana ja siinä aktuaattoreillaan toimivana (Russell ja Norvig 2003, 32). Agentin ja ympäristön tulee voida olla vuorovaikutuksessa toistensa kanssa. Agentin tulee voida lähettää signaaleja ympäristöön ja ottaa vastaan ympäristöstä tulevia signaaleja ja ympäristön tulee voida saada signaaleja ja lähettää niitä agentille. Agentin ympäristöön lähettämiä signaaleja voidaan sanoa toiminnoiksi (engl. actions), ja sieltä vastaanottamia havainnoiksi (engl. perceptions). (Legg ja Hutter 2007b, 15- 16.) Agentilla täytyy olla ainakin yksi tavoite. (Legg ja Hutter 2007b, 15- 16.) Tavoite on tietty tilanne, johon se pyrkii. (Fritz 2007, What is intelligence?, Acting on the environment, Objectives.) Agentti voi olla älykäs ilman tavoitetta, jonka saavuttamiseen se voi käyttää älykkyyttään, ja siinäkin tapauksessa, että agentti ei halua käyttää älyään tavalla joka vaikuttaa sen ympäristöön, mutta kummassakaan tapauksessa sen älykkyyttä ei voida havaita. Älykkyys voidaan havaita kun sillä on tavoite, jota se aktiivisesti ympäristöönsä vaikuttamalla yrittää saavuttaa. (Legg ja Hutter 2007b, 15- 16.) Ilmeisesti mikä tahansa peli, haaste, ongelma tai testi voidaan ilmaista tämän yksinkertaisen viitekehyksen avulla ilman suurta vaivaa. Viitekehys ei sano mitään siitä,

miten agentti tai ympäristö todella toimii, vaan kuvaa niiden roolit. (Legg ja Hutter 2007b, 17.) Määritelmä esitetään formalisoidussa muodossa lähteessä: (Legg ja Hutter 2007b, 17- 24).

Käytän viitekehystä, koska se on yleinen, eri aloilla menestyksellä käytetty, sekä täsmällisesti ja formaalisti määriteltävissä. Siten se on käyttökelpoinen perusta, jolle on suhteellisen turvallista rakentaa uusia käsitejärjestelmiä.

3.1.2 Tekoäly

Tekoälytutkimuksessa tavoitteena on hyödyllisten keinotekoisten älykkäiden systeemien rakentaminen. Tekoälytutkimus on antanut paljon tietoa ja välineitä älykkäiden systeemien yleiseen ymmärtämiseen. Varsinkin se valaisee kysymyksiä kuten: Mitä älykkyys on? Mitä laskennallisia prosesseja älykkääseen käyttäytymiseen tarvitaan? Miten näitä prosesseja tuottavia struktuureja rakennetaan?

Älykkäät systeemit -maailmankatsomuksen kannalta tekoälytutkimuksen merkitys on suuri. Tekoälytutkimus on tuottanut sen lähtökohdat. Walter Fritz (Fritz 2007) on luonut monia maailmankatsomuksen kannalta keskeisiä käsitteitä. Hän on tekoälytutkija, ja useat käsitteet ovat syntyneet käytännössä tekoälylaitteiden rakentamisessa saatujen kokemusten perusteella. Tekoälytutkimusta tehdään näiden tai niitä hyvin paljon muistuttavien käsitteiden avulla ja niitä kehittäen. Jos tutkimus menestyy, tämä tarkoittaa, että käsitteet ovat käyttökelpoisia, ja puhuu myös niiden pätevyyden puolesta.

Tekoälytutkimus hahmottui omaksi alakseen vuosina 1943- 1955. Warren McCulloch ja Walter Pitts tekivät ensimmäisinä työtä, jota pidetään tekoälytutkimuksena. He ehdottivat keinotekoisten neuronien mallia, jossa jokainen neuroni on joko "päällä", jos riittävän monta naapurineuronia stimuloi sitä, tai muuten "pois päältä", ja osoittivat, että mikä tahansa laskettava funktio voidaan laskea jollakin tällaisten neuronien verkostolla, ja että kaikki loogiset konnektiivit voidaan toteuttaa yksinkertaisilla neuronien verkoilla. He myös ehdottivat, että sopivalla tavalla määritelty verkko voi oppia. Donald Hebb esitteli yksinkertaisen päivityssäännön muuttamaan neuronien välisten yhteyksien vahvuuksia. Alan Turing artikuloi ensimmäisenä kokonaisen näkemyksen tekoälytutkimuksesta (1950)

artikkelissaan *Computing machinery and intelligence*, jossa hän esitteli turingin testin, koneoppimisen, geneettiset algoritmit ja vahvistusoppimisen. (Russel ja Norvig 2003, 16- 17.)

Kesällä 1956 John McCarthy, Marvin Minsky, Claude Shannon ja Nathaniel Rochester organisoivat workshopin automaatioteoriasta, hermoverkoista ja älykkyydestä kiinnostuneille tutkijoille, joista kollegoineen ja oppilaineen tuli tekoälytutkimuksen merkittävimmät hahmot seuraavaksi 20 vuodeksi. Tekoälytutkimuksesta tuli siellä itsenäinen ala, jonka tavoitteena oli tuottaa koneelle ominaisuuksia kuten luovuus, oppiminen ja kielenkäyttö sekä kyky toimia itsenäisesti monimutkaisissa muuttuvissa ympäristöissä. (Russel ja Norvig 2003, 17- 18.)

Tekoälytutkimuksen alkuvuodet olivat menestyksekkäitä rajoitetussa mielessä. Newellin ja Simonin General problem solver (GPS) suunniteltiin alusta alkaen imitoimaan ihmisen ongelmanratkaisua, ja rajoitetulla toiminta-alallaan sen havaittiin arvioivan alitavoitteita ja mahdollisia toimintoja samassa järjestyksessä kuin ihmistenkin. Herbert Gelernter (1959) rakensi Geometry theorem proverin (GTP), joka todisti monien matematiikan opiskelijoiden haastavina pitämiä teoreemoja. Arthur Samuelin shakkia pelaava ohjelma kehittyi nopeasti tekijäänsä paremmaksi. (Russell ja Norvig 2003, 18- 21.) Melkein kaikissa tapauksissa systeemit kuitenkin epäonnistuivat surkeasti, kun niitä testattiin laajemmilla ongelma-alueilla ja vaikeammilla ongelmilla. Ensimmäisen tyypin vaikeus, josta esimerkiksi kielenkääntäjä-ohjelmat kärsivät, johtui siitä, että aikaiset ohjelmat eivät käyttäneet ongelma-aluettaan koskevaa tietoa. Kielenkääntäminen vaatii aihealueeseen liittyvää yleistä tietämystä, jotta voidaan ratkaista ilmaisun epäselvyys ja monimerkityksellisyys. Toinen vaikeus oli moniin ongelmiin liittyvä tutkittavien tilojen määrän eksponentiaalinen kasvu, eli ns. kombinatorinen räjähdys. Jo sellaisten ongelmien, joihin liittyi yli muutama tusinaa faktaa ratkaiseminen tällä tavalla oli käytännössä mahdotonta, mistä huomattiin, että ohjelman periaatteellinen kyky löytää ratkaisu ongelmaan ei tarkoita, että se voisi käytännössä ratkaista sen. (Russell ja Norvig 2003, 21- 22.)

Viime vuosina (1995-nykyaika) sekä tekoälytutkimuksen menetelmät että sisältö ovat kokeneet vallankumouksen. Tekoälytutkimusta tehdään nykyään tieteellisellä menetelmällä, jossa hypoteesien hyväksyminen perustuu toistettaviin empiirisiin kokeisiin. Symbolien

manipuloinnista on siirrytty enemmän konnektionistiseen lähestymistapaan ja hermoverkkojen käyttöön. Tekoälyn osaongelmien ratkaisussa saadun menestyksen myötä tutkijat ovat alkaneet tarkastella koko agentin rakentamisen ongelmaa, minkä johdosta on alettu yhdistellä monia kauan erillisinä pidettyjä tutkimusaloja. Myös esimerkiksi ohjausteoria ja taloustiede käsittelevät agentteja, ja niistä on saatu vaikutteita tekoälytutkimukseen. (Russell ja Norvig 2003, 24- 27.) Tekoälyn nykyisistä kehityksistä kts. myös (López 2005 ja Fulcher 2006).

3.1.3 Systeemiteoria

Nykyään luonnon struktuurit ja lait selitetään usein monimutkaisten systeemien dynamiikan käsitteistön avulla: atomien ja molekyylien systeemeistä fysiikassa ja kemiassa solullisten organismien ja ekologisiin systeemeihin biologiassa, neuraalisista ja kognitiivisista systeemeistä aivotutkimuksessa ja kognitiotieteessä yhteiskuntiin ja talous/markkinasysteemeihin sosiologiassa ja taloustieteessä. (Mainzer 2004, 28.) Systeemiteorian ja kybernetiikan kieltä voidaan pitää metakielenä, jonka käsitteitä ja malleja voidaan käyttää monilla eri aloilla (François, 1999, 1), myös maailmankatsomusten rakentamisessa.

Historiaa

Systeemiteoria on kehittynyt eri aloilta tulevien käsitysten yhdistyttyä. Tämä kehitys alkoi 1948 Wienerin, von Neumannin, von Bertalanffyn, von Försterin ja Ashbyn pioneeritöistä, ja jatkuu edelleen. (François, 1999, 1.) Kreikan sana "sustema" tarkoittaa jälleennäkemistä, kokoonpanoa tai yhdistämistä. Sanaa "Kubernetes" (perämies, ruorimies) käytti jo Platon abstraktissa merkityksessä poliittisen entiteetin "pilottina", ohjaajana. (François, 1999, 1-2.) Systeemiteoria ja kybernetiikka tai kontrolliteoria liittyvät toisiinsa siten, että kontrolli viittaa systeemin käyttäytymisen joko sisäiseen tai ulkoiseen kontrollointiin.

Wienerin rooli kybernetiikan luojana ja kehittäjänä on tärkeä. Hän tutki ennustamisen, kontrolloimisen ja ohjauksen ongelmia ja keksi, että onnistuneen ohjauksen perusehto on ns.

palautteenkorjaus (engl. corrective feedback), mikä tarkoittaa, että itseohjautuva systeemi pyrkii tavoitetilaan korjaamalla ympäristöstään saamansa palautetta. Toinen Wienerin kontribuutio oli systeemin osien välillä ja osien ja ympäristön välillä tapahtuvan kommunikaatioon liittyvän käsitteistön sekä monimutkaisuuden tutkimuksen aloittaminen. Systeemin informaation määrä on sen järjestyksen ja entropia sen epäjärjestyksen aste. (Wiener, 1948.) Shannon ja Weaver (1949) selvensivät kommunikaation käsitettä käyttämällä lähteen, koodin, viestin, lähettäjän, signaalin, kanavan ja vastaanottajan käsitteitä. (François, 1999, 6.) Kaikki nämä käsitteet ovat tärkeitä mille tahansa systeemien luokalle, koska ne kaikki koostuvat elementeistä, jotka kommunikoivat keskenään. Ilman näitä käsitteellisiä perustyökaluja systeemiteoria ei olisi ollut mahdollista. (François, 1999, 7.)

1949 ja 1950 Bertalanffy muotoili yleisen systeemiteorian keskeiset käsitteet. 1949 hän esitti, että perusero elävien ja elottomien systeemien välillä on jälkimmäisen sopeutuva organisaatio, jonka jälkeen vitalistinen biologia joutui väistymään. (François, 1999, 5, 7.) 1978 ilmestyi Millerin systeemiteorialle tieteellistä statusta ja tutkimukselle suuntaa antava kirja *Living systems*, joka käsitteli mm. eläviä systeemejä, ja jossa hän myös teki systeemien luokituksen hierarkian tasojen ja monimutkaisuuden perusteella. 1980 Maturana keksi autopoieesin, itsetuotannon (itsensä kopioimisen tai replikaation) käsitteen, jolla on käyttöä mm. kognition ja elävien systeemien itsetuotantoon liittyvissä ongelmissa, ja johon liittyvät tärkeät itseviittavuuden ja autonomian käsitteet. (François, 1999, 10- 11.) Von Neumann aloitti automaattien tutkimuksen (theory of automation) (1956 ja 1966), joka perustuu Turingin teoreettiselle mallille tietokoneesta (1950), ja hän tutki myös joidenkin systeemien kykyä tuottaa itsensä uudelleen (François, 1999, 7-8), jota aihetta myöhemmin tutki unkarilainen Csanyi (1989) (François, 1999, 14). Näistä ja muista, mm. Langtonin, Brooksin ja kumppaneiden (1989) tutkimuksista sai alkunsa keinoelämän (engl. artificial life, AL) tutkimus, joka on johtanut mm. luonnollisten ja keinotekoisten systeemien sosiaalisen konstruktion samankaltaisuuden huomaamiseen, soluautomaateista sosiaalisiin hyönteisiin ja luultavasti myös ihmisyhteisöihin. Sosiaalisuus on yksi yleisimpiä aihealueita, joita tutkitaan monitieteellisesti systeemiteorian ja kybernetiikan keinoin. (François, 1999, 15.) Systeemiteoreettista lähestymistapaa on käytetty vaihtelevasti myös ekonomiassa (mm. Odum 1971, Georgescu-Roegen 1971, Daly 1973), globaalin hallinnoinnin suunnitteluun (Boulding

1953), sosiaalitieteissä (Berrien 1968, Buckley 1967 ja 1968) (François, 1999, 13- 14), matemaattiseen mallinnukseen, informaatioteoriaan, käytännön sovelluksiin laskennassa, ekologiassa, ja perheterapiassa (Heylighen ja Joslyn 1992).

teoriaa

Systeemi tarkoittaa toisiinsa liittyvien osien muodostamaa kokonaisuutta. Osien välisten suhteiden ansiosta systeemillä kokonaisuutena on ns. emergenttejä ominaisuuksia, joita sen osilla itsenäisesti tarkasteltuna ei ole. (Banathy, B.H. 1997.) Näitä kokonaisuuksia ei voida ainakaan kokonaan selittää pelkästään sen tiedon avulla, joka saadaan tutkimalla osia erillisinä yksikköinä. Tällaisista kokonaisuuksista hän mainitsee esimerkkeinä mm. elävät oliot ja jotkut sosiaaliset ilmiöt. (Bertalanffy 1968, 36- 37.)

Systeemi voidaan määritellä vuorovaikuttavien elementtien joukoksi. Vuorovaikutus tarkoittaa, että elementit, p, ovat suhteissa, R, siten että elementin p käyttäytyminen R:ssä eroaa sen käyttäytymisestä toisessa suhteessa, R'. Systeemin kaikki ominaisuudet eivät ole selitettävissä sen erillisten osien ominaisuuksilla, vaan ne voivat olla erilaisia ja uusia eli emergenttejä. Kuitenkin jos systeemin osat ja niiden väliset suhteet tunnetaan, systeemin käyttäytyminen voidaan päätellä. (Bertalanffy 1968, 19, 55- 56) Systeemi voidaan määritellä myös formaalisti: Systeemi koostuu objekteista ja objektien välisistä suhteista eli relaatioista: Systeemillä S tarkoitetaan joukkojen M ja S järjestettyä paria S = (M, R). Joukon M elementtejä sanotaan systeemin S objekteiksi. Joukko R taas on M:lle määriteltyjen relaatioiden joukko. Systeemin S joukkoa M sanotaan sen objektijoukoksi, ja joukkoa R sen relaatiojoukoksi. (Lin, Y. 1999, 352.) On osoitettu, että kaikkia formaaleja kieliä voidaan kuvata systeemeinä (M, K), jossa M on kielen kaikkien sanojen joukko ja K on kielen kaikkien kielioppisääntöjen joukko. Jokainen teoria voidaan kuvata systeeminä (M, T, ∪, K), jossa M ja K on määritelty kuten edellä ja T on joukko perusperiaatteita, joiden päälle teoria on kehitetty (Lin, 1989). Esimerkiksi matematiikka voidaan kuvata systeeminä (M, T, ∪, K) jossa T on ZFC -aksioomasysteemin kaikkien aksioomien joukko (Kuratowski ja Motowski, 1976). (Lin, Y. 1999, 352.)

Eri tieteenaloilta voidaan löytää formaalisti keskenään identtisiä tai isomorfisia lakeja. Joissakin tapauksissa lait koskevat tiettyjä systeemien luokkia tai aliluokkia, ja ne pätevät riippumatta niihin kuuluvien olioiden ominaisuuksista. Yleinen systeemiteoria tutkii systeemien yleisiä (joillekin systeemeille yhteisiä) ominaisuuksia, ja luokittelee systeemejä sen mukaisesti. (Bertalanffy 1968, 37- 38.) Se perustuu oletukseen, että on olemassa yleisiä järjestymisen periaatteita, jotka pätevät kaikkiin tai useisiin systeemeihin, esim. riippumatta siitä, ovatko ne fysikaalisia, kemiallisia, biologisia, kognitiivisia tai sosiaalisia. (Heylighen 1998.) Systeemiteoria on poikkitieteellinen tutkimusala ilmiöiden abstraktista organisaatiosta, riippumatta niiden substanssista, tyypistä tai tilallisesta tai ajallisesta skaalasta. Siinä tutkitaan sekä kaikille monimutkaisille entiteeteille yhteisiä periaatteita että (usein matemaattisia) malleja, joita käytetään niiden kuvaamiseen. Sen avulla eri oppialojen (mm. fysiikan, kemian, biologian, teknologian ja sosiologian) tutkimuskohteet voidaan ehkä ymmärtää samoilla käsitteillä ja periaatteilla, jolloin niiden tiedon yhdistäminen helpottuu. (Heylighen ja Joslyn 1992.)

Systeemit voidaan jakaa suljettuihin ja avoimiin. Ennen 1950-lukua fysiikassa tutkittiin lähinnä suljettuja systeemejä: sellaisia joiden ei oletettu olennaisesti vuorovaikuttavan niiden ulkopuolisen maailman kanssa. Tehtiin malleja esim. aurinkokunnasta, atomista tai heilurista olettaen, ettei muu universumi vaikuttaisi niihin, mikä teki mahdolliseksi laskea niiden tulevaisuuden tilat täsmällisesti, koska kaikki tarvittava informaatio on tiedossa. (Heylighen 1998.) Kuitenkin suurelle osalle ilmiöistä tällainen yksinkertaistus on käytännössä mahdotonta. Jos elävä olento erotettaisiin ympäristöstään, se kuolisi nopeasti hapen, veden ja ravinnon puutteeseen. Tunnetut elävät olennot ovat esimerkkejä avoimista systeemeistä: ne eivät tule toimeen ilman materian ja energian vaihtoa ympäristönsä kanssa. Avoimet systeemit ovat vuorovaikutuksessa toisten systeemien kanssa. Vuorovaikutus tapahtuu siten, että systeemi saa syötteen (engl. input) ympäristöstä, ja palauttaa vasteen (engl. output). Yleensä systeemin vaste on sen syötteen suora tai epäsuora tuote. Systeemin syötteen ja vasteen erottamiseksi täytyy tietää raja, joka erottaa sen ympäristöstään. Esimerkiksi eliöiden iho erottaa ne ympäristöstään. Kokonaisuutta, jossa systeemin syöte muuttuu sen vasteeksi, voidaan sanoa systeemin suoritukseksi (engl. throughput). (Heylighen 1998.)

Systeemejä lähemmin tarkasteltaessa voidaan huomata, että nekin koostuvat ympäristönsä kanssa vuorovaikuttavista systeemeistä. Systeemiin nähden sen osat ovat alisysteemejä (engl. subsystemes), joiden yli- tai supersysteemi (engl. supersystem) se on. Ylemmältä hierarkian tasolta voidaan saada kattava käsitys kokonaisuudesta, ja alemmalta sen osista, osien osista, jne. Esimerkiksi voidaan mitata kokonaisen kaupungin käyttämä polttoainemäärä (syöte), ja sen tuottaman saasteen määrä (vaste) tietämättä paljonko ja miltä osin kukin henkilö vastasi saasteen tuotannosta. Tällöin systeemiä (kaupunkia) tarkastellaan ikään kuin mustana laatikkona, josta havaitaan vain sen syöte ja vaste tietämättä, mitä systeemin sisällä tapahtuu. Vaikka näkökulma ei aina olekaan tyydyttävä, se on usein ainoa käytännössä mahdollinen. Systeemiä voidaan tarkastella ns. valkoisena laatikkona, jos sen sisäiset prosessit voidaan havaita. (Heylighen 1998.)

Systeemitutkimuksen neljä osa-aluetta ovat:
1) Systeemifilosofia
2) Systeemitiede
3) Systeemimetodologia
4) Systeemisovellukset
Lisäksi se voidaan jakaa tietoa ja sovelluksia tuottavaan tutkimukseen. (Banathy, B.H. 1997.)

1) Systeemifilosofia on systeemeihin liittyvien käsitteiden filosofista tutkimusta ja kehittämistä sekä maailman ilmiöiden tarkastelua ja selittämistä niiden avulla. Tätä ovat tehneet mm. Bunge, Bahm ja Laszlo. 2) Systeemitieteessä tutkitaan systeemejä tieteellisesti, määritellään, luokitellaan ja etsitään niiden toiminnan yleisiä periaatteita, ja toisaalta selitetään ilmiöitä systeemeihin liittyvien käsitteiden ja mallien avulla ja kehitetään niitä. Lisäksi varsinkin yleisessä systeemiteoriassa etsitään kaikkia systeemejä koskevia yleisiä toimintaperiaatteita, ja yhdistetään tieteellistä tietoa rajoittumatta erityisesti joihinkin tieteenaloihin. 3) Systeemimetodologiassa etsitään menetelmiä ja työkaluja, jotka sopivat parhaiten tietyn tyyppiseen systeemiin, tietyn tavoitteen, ongelman ja tilanteen kannalta. Se voidaan lisäksi jakaa systeemejä koskevan tiedon hankkimiseen ja esittämiseen liittyvien menetelmien etsimiseen ja toisaalta systeemien suunnittelussa ja kehittämisessä käytettävien menetelmien, strategioiden ja työkalujen etsimiseen. 4) Systeemisovellus tarkoittaa

systeemitutkimuksen tuottamien lähestymistapojen, käsitteiden, mallien, menetelmien ja työkalujen soveltamista tiettyyn tarkoitukseen. Tällöin valitaan systeemin tyyppiin ja tavoitteeseen sopivat tiedot ja menetelmät. (Banathy, B.H. 1997, Lin 1999, 9.)

Nykyinen tutkimus

Nykyiseen systeemiteorian kehitykseen liittyvät ainakin seuraavat kehitykset. Biologisen tutkimuksen tärkein kehitys on mm. ajatteluun, muistamiseen ja unohtamiseen, käsitteenmuodostukseen ja luovuuteen liittyvän hermojärjestelmää koskevan tiedon lisääntyminen. Toinen tärkeä alue on tekoälyyn (keinotekoisiin älykkäisiin systeemeihin) liittyvät kysymykset, kuten kykenemmekö rakentamaan autonomisia keinotekoisia älykkäitä systeemejä lähitulevaisuudessa? Mitä voimme oppia niistä? Kuinka suhtaudumme niihin yksilöinä ja yhteisöinä? Onko näillä systeemeillä etiikkaa ja oikeuksia? Kolmanneksi vastavuoroinen eri kulttuurien keskinäinen ymmärtäminen, ja eri tieteenalojen tiedon yhdistämisen projektit saattavat hyötyä yhteisestä kielestä. Systeemiteoreettis-kyberneettinen käsitteiden ja mallien metakieli, epistemologia ja ontologia näyttävät olevan askel tähän suuntaan, koska sitä voidaan käyttää hyvin monilla aihealueilla (François 1997). Näyttää siltä, että olemme tiellä uuteen yleiseen tieteenfilosofiaan, jolla on hyvin käytännöllisiä sovelluksia monilla eri alueilla ja myös eri alojen tiedon yhdistämiseen. (François, 2000, 19- 20.) (ks. myös Heylighen ja Joslyn 1992 ja Lin 2003.) Systeemiteorian ja sen luojien poliittisen merkityksen tarkastelua kts. (Debora 2003).

3.1.4 Älykkäät systeemit -maailmankatsomuksen kehitykseen vaikuttaneita tai muuten sille läheisiä ongelmia

- Mikä on laskettavissa?

Jos aivot ovat fysikaalinen laskentalaitteisto, tärkeäksi kysymykseksi tulee, mikä on laskettavissa? Älykkäiden systeemien ajattelu rajoittuu viime kädessä siihen, mikä on laskettavissa. Fysikaalinen laskettavuus rajaa sekä aivojen että kaikkien rakennettavissa olevien tietokoneiden suorituskykyä. Vuonna 1931 Kurt Gödel (1906- 1978) osoitti ns.

epätäydellisyysteoreemassaan, että missä tahansa kielessä, joka on riittävän ilmaisukykyinen kuvaamaan luonnollisten lukujen ominaisuuksia, on tosia lauseita, joiden totuus on ratkaisematon (engl. undecidable) siinä mielessä, että niiden totuutta ei voida selvittää millään algoritmilla. Tämän motivoimana Alan Turing (1912- 1954) lähti selvittämään, mikä ylipäätään on laskettavissa. Laskettavuuden käsite on ongelmallinen, koska sille ei ole voitu antaa formaalista määritelmää, mutta kuitenkin on yleisesti hyväksytty Churchin-Turingin teesi, jonka mukaan Turingin kone voi laskea minkä tahansa laskettavan funktion. Turing myös osoitti, että on olemassa funktioita, joita Turingin kone ei voi laskea. (Russell ja Norvig 2003, 8.) Näistä algoritmisesti ratkeamattomista ongelmista tärkeimpiä on pysähtymisongelma (engl. halting problem), jossa tehtävänä on selvittää, pysähtyykö annetun algoritmin suoritus, vai jatkuuko se loputtomiin, kenties kiertäen silmukkaa. (Ukkonen 1993, 29). Algoritmi määrittelee täsmällisen, äärellisessä ajassa suoritettavissa olevista perusaskelista muodostuvan operaatiojonon, joka käsittelee algoritmille annetun syötteen, ja tuottaa sen perusteella jonkin tulosteen (Ukkonen 1993, 28). Vaikka laskettavuuden periaatteelliset rajat on merkittävä tutkimuskohde, älykkäiden systeemien mahdollisten kykyjen ja tekoälyn kehityksen kannalta yhtä tärkeä on kysymys käytännön rajoituksista. Reaalimaailmassa toimiessaan älykkään systeemin täytyy usein tehdä hyvin nopeita päätöksiä, mikä rajoittaa käyttökelpoisten algoritmien määrää rajusti. Käyttökelpoisuus määräytyy tavoitteen ja käytettävissä olevien resurssien mukaisesti. Ollakseen käyttökelpoinen, algoritmin vaaditaan yleensä toimivan polynomisessa ajassa, eli että sen aikavaatimus on jokin syötteen pituuden polynomi. Algoritmia sanotaan eksponentiaaliseksi, jos sen aikavaatimus kasvaa nopeammin kuin mikään polynomi. Niin hitaita voidaan käyttää vain hyvin lyhyillä syötteillä (Ukkonen 1993, 30). Eksponentiaalisuuden takia tietokoneiden jatkuva nopeutuminen ei juuri auta merkittävien ongelmien ratkaisemiseksi, vaan joudutaan kehittämään parempia ratkaisualgoritmeja. (Russell ja Norvig 2003, 9.) Wolframin (2002, 713- 714) mukaan ennen nykyaikaisia tietokonesovelluksia saatettiin olettaa, että laskennan käsite soveltuu vain abstraktien elementtien systeemeihin. Nykyään laskennan elementteinä voi olla monentyyppistä dataa, numeroita, tekstiä, kuvaa jne. Tämä tuntuu osoittavan, että on mahdollista pitää mitä tahansa säännönmukaista prosessia laskentana, riippumatta siitä, minkälaisia elementtejä se sisältää.

- Voivatko koneet olla älykkäitä?

Heikko tekoäly (engl. weak AI) on kanta, jonka mukaan koneet voivat simuloida älykkyyttä, eli että ne voivat toimia ikään kuin olisivat älykkäitä. Vahvan tekoälyn (engl. strong AI) mukaan taas koneet voivat todella olla älykkäitä. Suurin osa tekoälytutkijoista olettaa, että ainakin heikko tekoäly on totta, eivätkä välitä vahvan tekoälyn totuudesta. Heille ei ole väliä, simuloiko ohjelma älykkyyttä vai onko se todella älykäs, vaan tärkeintä on että se toimii. (Russell ja Norvig 2003, 947.)

Jotta voitaisiin tunnistaa älykkyys, se täytyy määritellä. Turing oli pioneeri myös älykkyyden määritelmän kehittäjänä tekoälytutkimuksen piirissä. Kuuluisassa artikkelissaan "Computing machinery and intelligence" (1950) hän ehdotti, että koneiden älykkyyttä voitaisiin testata. Tässä ns. Turingin testissä älykkyys määritellään vertailukokeella, jossa testaaja keskustelee päätelaitteen välityksellä joko tietokoneen tai toisen ihmisen kanssa. Mikäli testaaja ei kykene keskustelemalla tunnistamaan kumppaniaan koneeksi, voidaan konetta pitää älykkäänä. (Hyvönen 1993, 15.) Tekoälytutkijat eivät ole olleet kovin kiinnostuneita Turingin testin läpäisystä. (Russell ja Norvig 2003, 948.) Tämä johtunee siitä, että tekoälylaitteet on yleensä suunniteltu jotain tiettyä tehtävää varten, eikä sanallinen ihmisen matkiminen ole ollut heidän mielestään hyödyllinen tavoite. Jotkut filosofit ovat yrittäneet todistaa, että ei ole mahdollista että koneet voisivat toimia älykkäästi ja argumentoineet tekoälytutkimuksen lopettamisen puolesta (mm. Sayre 1993).

Kyvyttömyyden argumentin (engl. argument from disability) mukaan kone ei koskaan voi tehdä X:ää. Joitakin X:iä ovat seuraavat: olla ystävällinen, olla aloitteellinen, omata huumorintajua, erottaa oikeata väärästä, oppia kokemuksesta, tehdä jotain täysin uutta. Ainakin tähän asti tekoälytutkimus on edistynyt, ja koneet voivat tehdä nykyään asioita, joihin ne eivät ole ennen kyenneet. Tietokoneet voivat nykyään suorittaa monia vaativia tehtäviä yhtä hyvin tai paremmin ihmiset: ne pelaavat shakkia, pokeria, oikolukevat tekstiä, ohjaavat autoja ja helikoptereita, diagnosoivat sairauksia ja tekevät tieteellisiä löytöjä. Joissakin tehtävissä koneet eivät edelleenkään ole hyviä, ja näihin kuuluu mm. Turingin testissä tarvittava vapaamuotoisen keskustelun käyminen. (Russell ja Norvig 2003, 948-949.) Matemaattisen argumentin (engl. the mathematical objection) mukaan koneet ovat mielen kyvyiltään ihmisiä rajoitetumpia, koska niitä formaaleina systeemeinä rajoittaa Gödelin

epätäydellisyysteoreema. Sen mukaan missä tahansa formaalissa systeemissä F, joka on riittävä aritmetiikkaan, on mahdollista muodostaa ns. Gödel-lause G(F), jolla on seuraavat ominaisuudet: G(F) on F:n lause, mutta sitä ei voida todistaa F:ssä, ja jos F on konsistentti, niin G(F) on tosi. John Lucas (1961) ja Sir Roger Penrose (1989, 1994) ovat väittäneet, että koneet eivät voi todistaa oman Gödel-lauseensa totuusarvoa, mutta ihmiset eivät ole tällä tavalla rajoitettuja. Vaikka myönnettäisiinkin, että koneiden todistuskyky on rajallinen, ei ole todisteita siitä, että ihmiset eivät olisi samalla tavalla rajoitettuja. (Russell ja Norvig 2003, 950.)

Russellin ja Norvigin mukaan tekoäly voidaan määritellä parhaan ohjelman etsinnäksi annetulle arkkitehtuurille. Tämän käsityksen kannalta tekoäly on mahdollista määritelmän mukaan: mille tahansa digitaaliselle arkkitehtuurille, jossa on *k* bittiä muistitilaa, on olemassa 2^*k* agenttiohjelmaa (engl. agent programs), joista täytyy etsiä paras. (Russell ja Norvig 2003, 947-948.) Tämä määritelmä toimii hyvin insinööriongelmana, ja tässä mielessä tekoäly on mahdollista. Filosofit ovat kuitenkin kiinnostuneita vertaamaan kahta eri arkkitehtuuria, ihmisen ja koneen. Lisäksi he ovat perinteisesti asettaneet kysymyksen seuraavasti: "Voivatko koneet ajatella?". Russellin ja Norvigin mielestä ongelma on huonosti rajattu. Tämän huomaamiseksi voidaan tarkastella kahta kysymystä:

- Voivatko koneet lentää?
- Voivatko koneet uida?

Suurin osa ihmisistä vastaa ensimmäiseen kysymykseen myöntävästi: esimerkiksi lentokoneet voivat lentää, mutta toiseen kieltävästi: laivat ja sukellusveneet kulkevat vedessä, mutta sitä ei sanota uimiseksi. Kuitenkaan kummallakaan kysymyksellä tai niiden vastauksilla ei ole merkitystä ilmailijoiden tai laivanrakennusinsinöörien työelämään, lentokoneiden tai sukellusveneiden suunnitteluun tai ominaisuuksiin, vaan ne liittyvät siihen, kuinka tiettyjä sanoja käytetään. Sana "uida" (engl. swim) tarkoittaa vedessä liikkumista ruumiinosia liikuttamalla, kun taas sanan "lentää" määritelmässä ei ole vastaavaa rajoitusta menetelmien suhteen. Ajattelevien koneiden käytännöllinen mahdollisuus on ollut olemassa vasta noin 50 vuotta, ja niinpä sanan "ajatella" käyttötapa ei tässä suhteessa ole vielä ehtinyt vakiintua. (Russell ja Norvig 2003, 948, 953- 954.)

Sille kannalle, että koneet voivat ajatella, löytyy ainakin kaksi perustetta: Ihmiset ovat koneita ja ihmiset voivat ajatella, siis: jotkut koneet voivat ajatella. On ehkä mahdollista rakentaa koneita, jotka suorittavat riittävän samankaltaisia prosesseja, että niiden voidaan sanoa ajattelevan, kunhan määritellään, mitä ajattelulla tällöin tarkoitetaan. Fritzin mukaan koska (esimerkiksi hänen rakentamansa) keinotekoiset älykkäät systeemit voivat suorittaa samat sisäiset ja ulkoiset toiminnot kuin ihminenkin, ne ajattelevat eivätkä vain simuloi ajattelua (Fritz 2007, Artificial intelligent systems).

Kysymykseen, voivatko koneet olla älykkäitä, voidaan vastata seuraavasti: Universaalin älykkyyden määritelmän mukaan älykkyyden määrä kuvaa agentin kykyä saavuttaa tavoitteitaan eri ympäristöissä. Kyky oppia ja ymmärtää, ratkaista ongelmia, suunnitella jne. sisältyvät implisiittisesti tähän määritelmään. (Legg ja Hutter 2007a, 12, Legg ja Hutter 2006, 73- 80, kts. myös tämän tekstin kappale 3.2.2 Älykkyys) Nykyään on käytössä paljon koneita, jotka tämän määritelmän mukaan ovat älykkäitä.

- Voimmeko rakentaa älykkäitä koneita?

Älykkäälle koneelle voidaan rakentaa myös sellaisia aisteja, joita ei ihmisellä ole ja ne voivat toimia hyvin erilaisissa ympäristöissä. Aistit yhdistetään hierarkkiseen muistijärjestelmään, joka toimii samoilla periaatteilla kuin aivokuori. Muistijärjestelmää opetetaan samaan tapaan kuin lapsia. Opettamisen myötä se rakentaa mallin ympäristöstään sellaisena kuin se havaitsee sen omien aistiensa kautta, jolloin se pystyy tekemään analogioita aiempiin kokemuksiin, tekemään ennusteita tulevasta, ehdottamaan ratkaisuja ongelmiin ja saattamaan tietonsa ihmisten käytettäväksi. (Hawkins 2005, 212- 213.) Suurin tekninen haaste älykkäiden koneiden rakentamisessa on muistin luominen. Muistin luomisen suurimpia teknisiä haasteita ovat suorituskyky ja yhdistettävyys. Aivokuorella on noin 32 biljoonaa synapsia. Jos kutakin synapsia edustaa vain kaksi bittiä (jotka antavat meille neljä mahdollista arvoa synapsia kohden) ja kussakin tavussa on kahdeksan bittiä (jolloin yksi tavu voi edustaa neljää synapsia), niin silloin tarvitaan noin kahdeksan biljoonaa tavua muistia. Tavallisen henkilökohtaisen tietokoneen (PC:n) levyllä oli vuonna 2005 noin sata miljardia tavua, joten tarvitsisimme noin 80 kovalevyä muistia saadaksemme samankokoisen muistin kuin ihmisen aivokuoressa on. Tämän kokoinen muisti on nykyään (2005) rakennettavissa laboratoriossa. Vähempikin riittää

moniin sovelluksiin. Ihmisen aivoissa on aivokuoren alla suuri määrä valkeaa ainetta. Valkea aine koostuu miljoonista tätä kautta kulkevista aksoneista, ja ohuen aivokuoren alla se yhdistää aivokuoren hierarkian eri alueet toisiinsa. Yksittäisellä aivokuoren solulla voi olla yhteys 5000- 10 000 muuhun soluun. Sähköjohdot lähettävät signaaleja paljon nopeammin kuin neuronien aksonit. Sirun yksittäinen johto voidaan jakaa ja sen vuoksi käyttää useisiin erilaisiin yhteyksiin, kun taas aivoissa kukin aksoni kuuluu vain yhdelle neuronille. Esimerkiksi puhelinjärjestelmissä kaikilla puhelimilla on yhteiset suhteellisen vähälukuiset, mutta hyvin suorituskykyiset linjat. Tämä menetelmä toimii niin kauan kuin kunkin linjan kapasiteetti on paljon suurempi kuin kapasiteetti, joka tarvitaan välittämään yksittäinen keskustelu. (Hawkins 2005, 213- 216.)

Älykkyys ei välttämättä tarkoita ihmismäisyyttä. Älykkyyttä mitataan hierarkkisen muistin ennustuskyvyllä, ei ihmisen kaltaisella käyttäytymisellä. (Hawkins 2005, 212- 213.) Usein elokuvissa tai kirjallisuudessa älykkäillä koneilla on hyvin paljon samankaltaisuutta ihmisten kanssa. Ihmisen mieli ei kuitenkaan ole ainoastaan aivokuoren luomus, vaan siihen vaikuttavat myös vanhojen aivojen emotionaaliset järjestelmät sekä ihmisen kehon monimutkaisuus. Ihminen tarvitsee ihmisenä olemiseen koko biologisen koneistonsa. Älykkäät koneet vastaavat aivokuorta, ja niillä on aistit, mutta loput ovat valinnaisia ominaisuuksia. Koneella ei ole ihmistä muistuttavaa mieltä, ellei sitä varusteta ihmistä muistuttavilla emotionaalisilla järjestelmillä ja ihmisen kaltaisilla kokemuksilla. (Hawkins 2005, 209- 212.) Älykkäät koneet voisivat mieltää maailmaa minkä tahansa luonnossa esiintyvän aistin kautta sekä ihmisen suunnittelemien uusien aistien avulla. Ääniluotain, tutka ja infrapunanäkö ovat todennäköisiä esimerkkejä, mutta ne voisivat kokea myös aidosti eksoottisia, vieraita aistimaailmoja. Aivokuoren algoritmi pyrkii pohjimmiltaan löytämään malleja maailmasta. Se ei aseta mallien fyysisiä alkuperiä paremmuusjärjestykseen. Niin kauan kuin aivokuorelle tulevat syötteet eivät ole sattumanvaraisia ja niissä on sisältöä tai tilastollista struktuuria, älykäs järjestelmä muodostaa niistä muuttumattomia muistikuvia ja ennusteita. Ei ole olemassa syytä, miksi näiden syötemallien olisi oltava analogisia eläinten aisteihin tai edes olla peräisin todellisesta maailmasta. (Hawkins 2005, 231.) On olemassa useita tapoja, joilla aivoja muistuttavat koneet voisivat ylittää omat kykymme. Älykkäät koneet saattavat ajatella ja oppia miljoona kertaa ihmistä nopeammin, muistaa valtavia määriä yksityiskohtaista informaatiota tai muodostaa

erittäin abstrakteja malleja. Ne saattavat ajatella kolmella, neljällä tai useammalla ulottuvuudella. (Hawkins 2005, 234- 236.)

- Pitäisikö meidän rakentaa älykkäitä koneita?

Miksi rakentaa keinotekoisia älykkäitä systeemejä? Syy on sama kuin muidenkin välineiden ja työkalujen rakentamiselle. Ne auttavat ihmisiä elämään miellyttävämmin. Ne saattavat joskus vapauttaa ihmiset (älyllisten ja aineellisten) resurssien niukkuudesta johtuvista huolista. (Fritz 2007, Artificial Intelligent Systems.) Ovatko robotit ihmiselle vaarallisia? Robotti, jonka päätavoite on miellyttää ihmisiä, voi olla suureksi avuksi, mutta robotti, jonka päätavoite on sen oma selviytyminen, olisi hyvin vaarallinen. Koska robotit tulevat ajattelemaan paljon nopeammin ja täsmällisemmin kuin ihmiset jälkimmäisenkaltaiset robotit voisivat käyttää kaikki saatavissa olevat resurssit ja ihmiset olisivat avuttomia. (Fritz 2007, Consequences of Artificial Intelligent Systems for our Human Society). Tällaisten robottien tulisi olla laittomia ja sellaiset tulisi tuhota niin pian kuin mahdollista

Uudet tiedon alueet ja uusi teknologia pelottavat ihmisiä aina, ja niin myös itsenäiseen ajatteluun ja toimintaan kykenevien koneiden mahdollisuus. (Hawkins 2005, 217.) Kaikkea teknologiaa voidaan soveltaa hyviin ja pahoihin tarkoituksiin, mutta se on vain väline, jota käytetään eri tarkoituksiin. On hyödyllisempää kehittää etiikkaa ja politiikkaa, joka säätelisi teknologian käyttöä kuin yrittää estää tieteen edistymistä.

Älykkäisiin koneisiin liittyvät pelot perustuvat usein oletukseen, että älykkyys synnyttäisi niihin joitain mahdollisesti ihmisille haitallisia motiiveja. Ihmisellä äly on sulautettu vanhojen aivojen emotionaalisiin vietteihin, pelkoon, vainoharhaisuuteen ja intohimoon. Älykkäillä koneilla ei ole näitä kykyjä. Niillä ei ole henkilökohtaisia pyrkimyksiä, ne eivät himoitse valtaa, sosiaalista asemaa tai aistillista tyydytystä. Niillä ei ole himoja, riippuvuuksia eikä mielialahäiriöitä. Niissä ei ole mitään, mikä muistuttaa ihmisen emootioita, elleivät ihmiset näe paljon vaivaa suunnitellakseen niille sellaisia. Tärkeimmät älykkäiden koneiden sovellukset ovat siellä, missä ihmisen älyllä on vaikeuksia eli alueilla, jossa aistimme ovat riittämättömät, tai tylsinä pitämissämme tehtävissä. Yleensä näillä tehtävillä on hyvin vähän

emotionaalista sisältöä. (Hawkins 2005, 219- 220.) On asetettava rajoituksia sille, minkälaisia koneita saadaan valmistaa ja mitä ihmiset saavat niillä tehdä.

Monilla uusilla teknologioilla on ollut tahattomia negatiivisia sivuvaikutuksia: polttomoottorin mukana tulivat ilmansaasteet ja moottoritiet, ydinfission myötä tulivat Tshernobyl ja globaalin tuhon pelko. (Russell ja Norvig 2003, 960.) Tekoälyn kehitykseen liittyviä ongelmia ovat luultavasti ainakin seuraavat:

- Ihmiset voivat menettää työnsä automaatiolle.
- Ihmisillä voi olla liian paljon (tai liian vähän) vapaa-aikaa.
- Ihmiset saattavat menettää tunteen omasta ainutlaatuisuudestaan.
- Ihmiset saattavat menettää joitakin yksityisyysoikeuksistaan (yksityisyydensuoja).
- Tekoälysysteemien käyttö voi johtaa vastuun menettämiseen.
- Tekoälyn menestys voi tarkoittaa ihmislajin loppua. (Russell ja Norvig 2003, 960.)

Ei tiedetä, kuinka tekoälyn edistys vaikuttaa ihmisten työpaikkojen määrään. Jotkin ammatit saattavat olla tulevaisuudessa täysin tai osittain automatisoitu, ja alan ihmiset voivat joutua työttömiksi. Toisaalta tekoäly saattaa myös synnyttää uusia aloja ja ammatteja ihmisille tai tekoäly voi tehdä sellaista työtä, mitä ennen sitä ei ollut olemassa. (Russell ja Norvig 2003, 960.) Jos työpaikkojen tai vaadittavien työtuntien määrä vähenee, ihmiset voivat työn vähetessä jäädä ilman riittävää toimeentuloa, ellei muuta tulonlähdettä kehitetä. Tekoälyn käyttäminen saattaa vähentää paitsi työpaikkojen määrää, myös jäljelle jäävien työpaikkojen ihmisiltä vaatimaa työtuntien määrää, ja siten lisätä ihmisten vapaa-ajan määrää. Toisaalta esimerkiksi joillakin tietotyön aloilla ihmiset työskentelevät 24 tuntia vuorokaudessa toimivien tietokonesysteemien parissa, ja pysyäkseen työn tahdissa, heidän täytyy työskennellä pidempään kuin ennen. Kilpailuun perustuvassa informaatiotaloudessa, tuotot eivät aina ole selkeässä suhteessa työhön sijoitetun ajan kanssa, vaan voi olla suuri hyöty olla vähän kilpailijoitaan parempi; työskenteleminen 10 % enemmän saattaa lisätä tuloja 100 %. Niinpä ihmisillä on jatkuva paine työskennellä enemmän. Tekoäly myös lisää nopeutta, jolla tieteelliset ja teknologiset keksinnöt syntyvät. (Russell ja Norvig 2003, 961.)

Keinotekoisten älykkäiden systeemien käyttäminen aiheuttaa taloudellisen mullistuksen, joka on jo alkanut tietokoneiden käytön myötä. Kestää vielä jonkin aikaa kunnes todella älykkäät, halvat, monikäyttöiset robotit tulevat saataville. Kun se tapahtuu, monet työtehtävät katoavat, ihmisten täytyy hankkia muita töitä, ja tarvittavien työtuntien määrä vähenee. Näin on tapahtunut ennenkin uusien teknologioiden ilmaantumisen myötä. Pitkällä aikavälillä näyttää siltä, että työ, eli maksettu toiminta vähenee merkittävästi tai loppuu kokonaan. Robottien tekemä työ tyydyttää kaikki materiaaliset tarpeet ja ihmiset voivat tehdä mitä tahansa haluavat. (Fritz 2007, Economic effects.) On monia tapoja miten muutos ihmisten työhön perustuvasta yhteiskunnasta robottien työhön perustuvaan voidaan tehdä. Fritz tarjoaa järkevää ratkaisua (katso: Fritz 2007, Consequences of artificial intelligent systems for our human society). Keinotekoiset älykkäät systeemit robotteina, älykkäänä automaationa ja neuvonantajaohjelmina tekevät kaiken työn jota ihmiset eivät halua tehdä. Ihmiset voivat vapaasti nauttia elämästä. Uusi "teollinen vallankumous" ja muutos työhön perustuvasta yhteiskunnasta vapaa-aikaan perustuvaan yhteiskuntaan täytyy kuitenkin tehdä hallitusti. Laaja työttömyys voidaan välttää jakamalla työ kaikkien halukkaiden kesken vähentämällä viikoittaisten työtuntien määrää. Lopulta työtä on niin vähän, että tulojen ja ostovoiman ylläpitämiseksi tarvitaan muita keinoja. Tällainen voisi olla "sosiaalinen osinko". Jokainen yhteiskunnan jäsen olisi valtion osakas, ja saisi kuukausittain osinkonsa. Varat tähän tulisivat automatisoitujen tehtaiden tuotoista (voitoista). (Fritz 2007, Consequences of Artificial Intelligent Systems for our Human Society).

Tekoälyn myötä ihmiset saattavat menettää tunteen omasta ainutlaatuisuudestaan. Weizenbaum teoksessaan Computer power and human reason (1976) esitti joitakin tekoälyn synnyttämiä uhkia yhteiskunnalle. Yksi pääargumenteista oli, että tekoäly tekee mahdolliseksi ajatuksen ihmisistä automaatteina. Hänen mielestään tämä voi aiheuttaa ihmisten autonomian eli itsemääräämisoikeuden ja jopa ihmisyyden menettämisen. Russell ja Norvig huomauttavat, että ihmiset ovat selviytyneet jo monista ainutlaatuisuuden tunnettaan vahingoittavista iskuista: Kopernikuksen teorian myötä maan ei enää ajateltu olevan aurinkokunnan keskipisteessä (Kopernikus 1543), ja Darwinin evoluutioteorian myötä ihmistä alettiin pitää yhtenä

eläinlajina muiden joukossa (Darwin 1871). Jos tekoäly menestyy, se voi olla yhtä suuri uhka nykyajan ihmisten moraalisille oletuksille kuin Darwinin ja Kopernikuksen teoriat aikoinaan.

Ihmiset saattavat menettää yksityisyydensuojaansa. Tekoälyteknologia, mm. kielenkääntäminen, puheentunnistus, salasanojen etsiminen mahdollistaa tehokkaan valvonnan, ja jos sitä ei rajoiteta, ihmiset saattavat menettää yksityisyytensä. Toisaalta terrorismin uhan takia ihmiset voivat olla entistä halukkaampia lisäämään valvontaa. (Russell ja Norvig 2003, 961.)

Tekoälysysteemien käyttö voi johtaa vastuun menettämiseen. Jos lääkäri käyttää eksperttisysteemiä diagnoosin tekoon, kuka on vastuussa, jos diagnoosi on väärä? Tähän asti USA:ssa oikeusistuimissa lääketieteellisiä eksperttisysteemejä on pidetty lääkärien työkaluina samaan tapaan kuin oppi- ja lähdekirjoja. Lääkärin vastuu on ymmärtää päätöksiin johtavat syyt ja päättely, ja käyttää omaa arvostelukykyään systeemin suositusten hyväksymisessä. Tulevaisuudessa on mahdollista, että jos eksperttisysteemit tulevat selvästi luotettavimmiksi ja täsmällisemmiksi kuin ihmiset, lääkärit voivat joutua korvausvelvollisiksi, jos he eivät käytä eksperttisysteemin suosituksia. (Russell ja Norvig 2003, 962.)

Tekoälyn menestys voi tarkoittaa ihmislajin loppua. Monet teknologiat voivat mahdollistaa suuren tuhon vääriin käsiin joutuessaan. Jos robotit on suunniteltu ihmisiä palveleviksi niin kuin mitkä tahansa työkalut, niistä ei luultavasti ole vaaraa. Jotkut ihmiset saattavat kuitenkin suunnitella robotteja, jotka toteuttavat heidän tavoitteitaan, eivätkä välitä muiden ihmisten tavoitteista, tai vahingoittavat muita ihmisiä. Heitä täytyy pitää rikollisina. Ihmiset käyttävät älyään aggressiivisesti, koska heillä on luonnonvalinnan kautta perittyjä aggressiivisia taipumuksia. Ihmisten rakentamilla koneilla ei kuitenkaan tarvitse olla aggressiivisia taipumuksia, ellei niille nimenomaan rakenneta niitä, jolloin aggressiivisuus on lähtöisin ihmisestä. (Russell ja Norvig 2003, 963.) Vernon Vingen (1933) nimeämä teknologinen singulariteetti tarkoittaa ajankohtaa, jolloin ihmiset saavat aikaan teknologian, joka tekee mahdolliseksi ihmisiä älykkäämpien koneiden rakentamisen. Teknologisen singulariteetin on ennustettu tulevan n. vuonna 2020 (Vinge), 2050 mennessä (Moravec 2000) ja Ray Kurzweill kirjassaan The age of intelligent machines (2000) 2099 mennessä. Jonkinlaisia, esimerkiksi

tietoisia robotteja voi olla moraalitonta kohdella pelkkinä koneina. Toisaalta myös robottien pitää käyttäytyä moraalisesti, eli niihin täytyy ohjelmoida teoria oikeasta ja väärästä. Science fiction -kirjailijat ja elokuvantekijät ovat käsitelleet aihetta robottien oikeuksista ja velvollisuuksista, alkaen Isaac Asimovista (1942) Steven Spielbergiin (A.I. 2001) ja eteenpäin. (Russell ja Norvig 2003, 963- 964.)

3.2 Älykkäät systeemit

Tässä osassa määritellään älykkään systeemin käsite. Jos olio on systeemi ja älykäs, sitä voidaan sanoa älykkääksi systeemiksi. Älykkään systeemin käsitteen yksityiskohtaiseksi määrittelemiseksi määritellään ensin systeemin ja älykkyyden käsitteet.

3.2.1 Systeemit

Yleisesti systeemi koostuu objekteista ja objektien välisistä suhteista eli relaatioista: systeemillä S tarkoitetaan joukkojen M ja S järjestettyä paria S = (M, R). Joukon M elementtejä sanotaan systeemin S objekteiksi. Joukko R taas on M:lle määriteltyjen relaatioiden joukko. Niinpä Systeemin S joukkoa M sanotaan sen objektijoukoksi ja joukkoa R sen relaatiojoukoksi. (Lin, Y. 1999, 352.) On osoitettu, että kaikkia formaaleja kieliä voidaan kuvata systeemeinä (M, K), jossa M on kielen kaikkien sanojen joukko ja K on kielen kaikkien kielioppisääntöjen joukko. Jokainen teoria voidaan kuvata systeeminä (M, T, ∪, K), jossa M ja K on määritelty kuten edellä ja T on joukko perusperiaatteita, joiden päälle teoria on kehitetty (Lin, 1989). Esimerkiksi matematiikka voidaan kuvata systeeminä (M, T, ∪, K) jossa T on ZFC -aksioomasysteemin kaikkien aksioomien joukko (Kuratowski ja Motowski, 1976). (Lin, Y. 1999, 352.)

Systeemiteorian käsitteet ovat osa älykkäät systeemit -maailmankatsomuksen kieltä ja muodostavat sen ontologian. Jonkinlaista systeemiteoreettista kieltä suositellaan maailmankatsomuksen kieleksi sen täsmällisyyden ja yleiskäyttöisyyden ansiosta. (Vidal 2007,

24.) Systeemiteoriaa käsittelevässä kappaleessa (3.1.3) esiteltiin joitakin systeemiteorian käsitteitä.

Systeemi on:
1. Osa universumia.
2. Vahvempia tai enemmän korrelaatioita on (tai voidaan havaita) systeemin osan ja toisen osan välillä kuin systeemin osan ja muun universumin välillä. (Fritz 2007, Definition of the concept "system".)

Universumi on "kaikki mikä on olemassa, kokonaisuutena tarkasteltuna" (Fritz 2007, Glossary, Universe). Sana on peräisin latinasta: "universum "kaikkien olemassa olevien olioiden kokonaisuus, " universus "kokonainen, yhdistetty, " uni- "yksi" ja versus "käännetty kohti, " vertere "kääntyä" ". (Merriam-Webster online dictionary, ja Merriam-Webster student dictionary: universe.)

Universumissa on tai voidaan havaita säännönmukaisuuksia, korrelaatioita, joiden perusteella sitä voidaan jakaa osiin. Korrelaatiolla tarkoitetaan kahden muuttujan välistä riippuvuutta. Tilastotieteessä korrelaatiota kuvataan korrelaatiokertoimen arvona, joka kuvaa riippuvuuden suuruutta ja suuntaa välillä [-1, 1]. Positiivinen korrelaatio saa maksimissaan arvon +1 ja muuttujien riippuvuus on tällöin täydellinen, ja vastaavasti korrelaatiokertoimen ollessa -1 on täydellinen negatiivinen korrelaatio. Korrelaatiokertoimen ollessa nolla riippuvuutta muuttujien välillä ei ole. (StatSoft, Inc. 2007.) Struktuuri on universumin osa, jolla on rajallinen tilallinen laajuus. Transformaatio taas on tilallisten tai ajallisten suhteiden muutos universumin osassa. (Fritz 2007, Correlation, Transformation, Structure.) Universumia voidaan jakaa struktuureihin (rakenteisiin) ja transformaatioihin (muutoksiin). Korrelaatioita voi olla struktuurien välillä, transformaatioiden välillä tai niiden kesken. Toisaalta ne itse edellyttävät jonkinlaisia alemman tason korrelaatioita, jotta ne voitaisiin ylipäänsä erottaa taustasta.

Systeemit ovat universumin osia. Systeemejä erotetaan muusta universumista korrelaatioiden perusteella: Määrittelemme systeemin kun mielessämme erotamme muusta universumista osia,

joilla on vahvempia tai enemmän korrelaatioita keskenään (Fritz 2007, Definition of the concept "system"). Systeemeillä on monia korrelaatioita, vahvempia ja heikompia, ja olisi vaikeaa ja turhaa ottaa huomioon niitä kaikkia ja yleensä systeemiä määriteltäessä otetaankin huomioon vain vahvimmat korrelaatiot, (Fritz 2007, Correlation) tai muuten tärkeimmät. Eri älykkäät systeemit ovat myös herkempiä havaitsemaan itselleen tärkeämpiä korrelaatioita.

Korrelaatio ja syy voidaan erottaa siten, että vain älykkäät systeemit voivat olla jonkin syitä (Fritz 2007, Correlations, Cause and Effect). Syyn käsite on yleensä määritelty seuraavalla tavalla: Kausaliteetti edellyttää, että on olemassa lakeja joiden mukaisesti tiettyyn luokkaan kuuluvan olion B ilmaantuminen riippuu toiseen luokkaan kuuluvan olion A ilmaantumisesta. A:ta sanotaan syyksi, ja B:tä seuraukseksi. Syyn täytyy edeltää tai olla samanaikainen seurauksen kanssa. Syyn ja seurauksen täytyy olla joko suorassa tai välillisessä tilallisessa yhteydessä toisiinsa. (Born 1949, Sowan 2000 viittaamana.) Fritzin käsityksen mukaan jos älykkäällä systeemillä A on tavoite B ja se toimii tuottaakseen B:n, niin silloin A onnistuessaan on B:n syy. Luonnossa, johon älykäs systeemi ei vaikuta, on vain korrelaatioita, ei syitä. (Fritz 2007, Correlations, Cause and Effect.)

Frontieeri (engl. frontier) on systeemin raja, joka erottaa sen muusta universumista. Se erottaa universumin osat, joilla on vahvempia tai useampia korrelaatioita keskenään, muusta universumista. Kaikki systeemin frontieerin ulkopuolella oleva johon systeemi vaikuttaa tai joka vaikuttaa systeemiin, on systeemin ympäristöä (engl. environment). Toisin sanoin systeemin ympäristöä on kaikki se universumista, joka on kommunikaatiossa (engl. communication) systeemin kanssa. Kaikki kommunikaatio systeemin ja sen ympäristön välillä kulkee frontieerin läpi. (Fritz 2007, Frontier and environment, communication.)

Kommunikaatio on materian tai energian liikettä kahden universumin osan välillä. Materia tai energia voi kantaa informaatiota. Fritzin käsityksen mukaan informaatio on käsitteiden ja vastesääntöjen summa, joka voidaan erottaa kommunikaatiosta. (Fritz 2007, Communication, information.) Informaatiota voidaan mitata objektiivisesti ja subjektiivisesti. Objektiivinen mitta ilmaisee, kuinka paljon kohteessa on informaatiota ja subjektiivinen sitä, kuinka paljon informaatiota tietty subjekti kohteesta erottaa. Objektiivisen informaation mitta on osa

algoritmista informaation teoriaa (AIT), jossa kohteen informaatiomäärä mitataan seuraavasti: Jos systeemille A voidaan antaa sopivassa kielessä rakennekuvaus D, niin A:n "informaatiomäärä" riippuu siitä, kuinka pitkän ohjelman Turingin kone (idealisoitu tietokone) tarvitsee tuottaakseen kuvauksen D (Niiniluoto 1989, 21). Toisin sanoin objektin informaatiosisältö on sen lyhimmän kuvauksen pituus jollakin standardikielellä (Hutter 2000). Subjektiivinen (ja edellä Fritzin käyttämä) mitta informaation määrästä on, kuinka paljon hyödyllistä säännönmukaisuutta (käsitteitä ja vastesääntöjä) tietty älykäs systeemi voi erottaa tietystä kommunikaatiosta (datasta). Informaatioon liittyvän käsitteistön tutkimus on nykyään kiihkeää filosofiassa. Tästä voi ottaa selvää mm. lähteistä: (Floridi 2004 ja 2007). Tilanne varsinkin ns. semanttisen informaation tutkimuksessa on vielä ilmeisesti hyvin keskeneräinen.

3.2.2 Älykkyys

Älykkäät systeemit ovat systeemejä, jotka ovat älykkäitä. Mitä älykkyys on? Miten älykkyyttä on syntynyt joihinkin systeemeihin? Miten älykkäitä systeemejä on syntynyt universumiin?

Älyn historia

Älykkyys on evolutiivinen etu, koska sen avulla voidaan mallintaa, ennustaa ja manipuloida ympäristöä (Yudkowsky 2002, 1). Tämä koskee myös sosiaalista ympäristöä. Älykkyys on hyödyllistä vain, jos maailmalla on struktuuria eli säännönmukaisuutta, ja on sen vuoksi ainakin osittain ennustettavissa. Muisti, ennustaminen ja käyttäytyminen olisivat kaikki hyödyttömiä, jos maailmalla ei olisi struktuuria. "Kaikki (älykäs) käyttäytyminen, olipa se sitten ihmisen, etanan, yksisoluisen eliön tai puun käyttäytymistä, on keino hyödyntää maailman struktuuria lisääntymiseen." (Hawkins 2005, 181- 182.) Kaikki eliöt ovat jossakin määrin älykkäitä, koska ne hyödyntävät muistia ja ennustamista lisääntyäkseen tehokkaammin. Lammikossa elävällä yksisoluisella eliöllä on siima, jonka avulla se pystyy uimaan. Solun pinnalla on molekyylejä, jotka havainnoivat ravintoaineita. Lammikon kaikilla alueilla ei ole yhtä paljon ravintoaineita. Kun eliö ui lammikon toiselle alueelle, se pystyy havaitsemaan muutoksen. Solu hyödyntää kemiallista tietoisuuttaan uimalla kohti paikkoja, joissa on

enemmän ravintoaineita. Se ennustaa, että tietyllä tavalla uimalla se löytää enemmän ravintoaineita. Ennusteen tekemiseen tarvitaan muistia. Tässä tapauksessa muisti on eliön DNA. Yksisoluinen eliö ei opi elinaikanaan, vaan oppiminen on tapahtunut evoluution aikana, ja se on tallentunut eläimen DNA:han. Jos maailman struktuuri muuttuisi äkillisesti, yksisoluinen eläin ei oppisi mukautumaan, koska se ei pysty muuttamaan DNA:taan tai sen seurauksena syntyvää käyttäytymistään, vaan laji oppii ainoastaan monien sukupolvien mittaisissa evoluutioprosesseissa. (Hawkins 2005, 182.) Jokainen laji käyttäytyy hieman eri tavalla ja hyödyntää maailman struktuurista hieman eri osia omiin tarkoituksiinsa. (Hawkins 2005, 183.) Elämänaikainen oppiminen tuli mahdolliseksi, kun evoluution aikana neuronien välisistä yhteyksistä tuli muokattavia. Neuroni pystyi lähettämään signaalin tai olemaan lähettämättä signaalia sen mukaan, mitä äskettäin tapahtui. Eliön käyttäytymistä pystyttiin muokkaamaan nyt elämän aikana ja se pystyi oppimaan ympäristönsä struktuuria elinaikanaan. Muovattava hermojärjestelmä oli todella suuri evolutiivinen etu, ja se johti uusien lajien nopeaan kehitykseen - kaloista etanoihin ja etanoista ihmisiin. (Hawkins 2005, 183.)

Aivokuori on kaikkein viimeisimmäksi kehittynyt hermokudos. Kaikilla nisäkkäillä on vanhat aivot, joiden päällä sijaitsee aivokuori. Hierarkkisen rakenteensa, esitystapainvarianttiensa ja analogiaan perustuvan ennustamisensa avulla aivokuori antaa nisäkkäille mahdollisuuden hyödyntää maailman struktuuria huomattavasti enemmän kuin mihin aivokuorettomat eläimet pystyvät. (Hawkins 2005, 183- 184.) Kasvattamalla aivokuoren kokoa ihmisaivot voivat oppia monimutkaisempia malleja maailmasta ja tehdä monimutkaisempia ennusteita. Ihminen näkee syvempiä analogioita, enemmän struktuureja struktuurissa kuin muut nisäkkäät. (Hawkins 2005, 184.)

Ihmisen aivot ovat merkittävän suuret sekä absoluuttisesti että suhteessa muuhun ruumiiseen. Ihmisen kognitiivisia adaptaatioita suosineita olosuhteita ei tunneta tarkkaan. Viimeisimmät mallit perustuvat ekologisiin olosuhteisiin ja siihen liittyvään sosiaaliseen ongelmanratkaisuun. Yhden käsityksen (Richerson, Bettinger ja Boyd 2005, Calvin 2001) mukaan ihmisen aivojen räjähdysmäinen kehitys tapahtui kun ilmastonvaihtelut olivat erittäin suuria. 100 000- 10 000 vuotta sitten välillä oli erittäin lämmintä ja välillä äärimmäisen kylmää, minkä jälkeen on vallinnut lähes tasainen lämmin ilmasto jossa oli vain yksi lyhyt kylmä jakso. Ihmisen

sopeutumisponnistelut ovat siis olleet äärimmäisiä. Richard Alexanderin mukaan kun esi-isämme kykenivät yhä enemmän hallitsemaan luonnon tapahtumia, kilpailu saman lajin edustajien kanssa tuli tärkeämmäksi, erityisesti sosiaalisen kyvykkyyden suhteen. Simpanssien kanssa samankaltaisten sukulaisuuteen tai vastavuoroisuuteen perustuvien liittojen lähtökohdista alkoi autokatalyyttinen sosiaalisten kykyjen kilpavarustelu, joka johti kokoelmaan ihmislajille tunnusomaisia ominaisuuksia, kuten kätkettyyn ovulaatioon, molempien vanhempien osallistumiseen lastenhoitoon, monimutkaiseen sosiaalisuuteen, ja poikkeavaan kognitiivisten kykyjen kokoelmaan. (Flinn et al. 2005, 10- 11.) Evoluution ja ihmisen psykologian yhteydestä katso myös: (Buss 2005 ja 2007). Monia tyypillisiä hominidien adaptaatioina pidettyjä kykyjä ajatellaan nykyään olleen myös apinoilla, jotka olivat hominidien esi-isiä. Ihmisten kognitiivisten kykyjen kehitys alkoi hominidien esi-isinä olleiden apinoiden sopeutumisesta Euraasian metsien vaikeisiin elinoloihin, ja hominideille ainutlaatuiset kognitiiviset adaptaatiot kehittyivät sopeutumisena vielä vaikeampiin metsättömän ruohomaaston ekologisiin olosuhteisiin, ja ovat suoraa jatketta edellisistä. (Russon ja Begun 2004, 353- 366.)

Nykyihmiselle kieli on hyvin tärkeä esimerkiksi tieteessä, teknologiassa, taiteessa, politiikassa, kaupankäynnissä, ja henkilökohtaisissa suhteissa. Kuitenkaan ei tiedetä, mikä oli kielen ensimmäinen tehtävä, miksi se ilmaantui ja mikä oli sen tehtävä varhaisissa yhteisöissä. Ainakin kielellisellä kyvyllä tuntuu olevan tärkeä suhde ihmisen sosiaaliseen organisaatioon ja sen ympäristössä selviytymiseen. Yhden selityksen mukaan kieli oli alun perin (ja on nykyään) tärkeä kommunikaatioväline metsästäjäryhmän jäsenten välillä, esimerkiksi metsästyksen suunnittelua ja koordinointia varten, sekä sotureiden ja perheenjäsenten välillä. (Sabbatini 2001, Language and evolution.) Kielen avulla ihminen pystyy herättämään esiin muistikuvia ja luomaan uusia rinnastuksia mentaalisista kohteista toiselle ihmiselle. Sen avulla voimme saada toiset ihmiset kokemaan ja oppimaan asioita, joita he eivät ehkä koskaan oikeasti näe. Kielen avulla voidaan siirtää elämän aikana opittuja malleja nykyisille yhteisön jäsenille ja seuraaville sukupolville. Ilman kieltä elävät eläimet eivät välitä lähellekään yhtä paljon tietoa jälkeläisilleen. Rotta voi oppia monia malleja elinaikanaan, mutta se ei välitä yksityiskohtaista uutta informaatiota jälkeläisilleen. Kielen kehitys edellytti suurta aivokuorta, joka kykeni käsittelemään syntaksin ja semantiikan sisäkkäisiä rakenteita. Se edellytti myös

täydellisemmin kehittynyttä motorista aivokuorta ja lihaksistoa, joiden avulla ihminen pystyy esittämään pitkälle kehittyneitä ja tarkasti artikuloituja äänteitä ja eleitä. (Hawkins 2005, 185.)

Ihmisen älykkyydellä ja kyvyllä tehdä työkaluja on myös vahva yhteys. Työkalujen valmistamisen teknologian ilmaantuminen hominidien evoluution aikana merkitsee suurta eroa apinoihin, ja niiden jäänteet ovat varhaisin löydetty todiste kulttuurisesta traditiosta, jolla on arvoa selviytymisen kannalta, ja joka perustuu oppimiseen. Ihmiset tekivät työkaluja ensimmäisen kerran 3 tai 4 miljoonaa vuotta sitten. Ne tehtiin luultavasti puusta tai luusta, mutta vain niiden jälkeen noin 2,5 miljoonaa vuotta sitten käytettyjä kivityökaluja on löydetty, koska ne säilyvät paremmin. Ihmiset käyttivät niitä toisten ihmisten ja eläinten tappamiseen, luiden rikkomiseen, lihan viipaloimiseen, oksien katkomiseen ja puukeppien teroittamiseen. Hominidit ja varhaiset ihmiset luultavasti elivät kasvien ja hedelmien keräilyllä, haaskansyönnillä ja metsästyksellä, ja työkalut tekivät näistä kaikista toiminnoista entistä tehokkaampaa. Ne myös mahdollistivat tehokkaamman metsästyksen ja lihapitoisemman ruokavalion, joka mahdollisti suurempien aivojen ylläpidon ja pitkät metsästysmatkat. (Sabbatini 2001, Tool making, hunting and war.) Kyky tehdä työkaluja kehittyi luultavasti vähäpuisen maaston vuoksi kehittyneen kaksijalkaisuuden ansiosta, joka vapautti kädet työkalujen valmistamista, kantamista ja käyttämistä varten, ja teki myös mahdolliseksi ihmiselle ainutlaatuisen peukalon kehittymisen. Koska työkalujen valmistus ei ole synnynnäinen kyky eikä siis leviä geenien kautta, ainoa keino levittää sitä sukupolvelta toiselle on kulttuurisen perinteensiirron kautta, kielen, imitaation ja harjoittelun avulla. Tätä voidaan pitää käytännöllisen koulutuksen alkuperänä. Jotkin muutkin eläimet (ainakin apinat) voivat oppia imitaation avulla, mutta paljon rajoitetummin. (Sabbatini 2001, Tool making, hunting and war.)

Älykkyyden kehitys voidaan jakaa kolmeen aikakauteen:
1) Eliöt hyödynsivät DNA:ta muistivälineenä. Yksilöt eivät pystyneet oppimaan ja sopeutumaan elinaikanaan, vaan pystyivät ainoastaan välittämään DNA:han perustuvia muistikuvia jälkeläisilleen geeniensä avulla.
2) Luonto keksi muokattavat hermojärjestelmät, jotka pystyvät nopeasti muodostamaan muistikuvia. Yksilö pystyi tuolloin oppimaan maailmansa struktuuria ja sopeuttamaan

käyttäytymistään elinaikanaan. Yksilöt eivät tosin vieläkään pystyneet välittämään tätä tietoa jälkeläisilleen muuten kuin suoralla havainnoinnilla.
3) Kolmas ja toistaiseksi viimeinen, ihmiselle ainutlaatuinen aikakausi alkaa kielen keksimisestä ja suuren aivokuoremme laajenemisesta. Ihminen oppii paljon maailman struktuurista elinaikanaan, ja voi tehokkaasti välittää tätä tietoa toisille ihmisille kielen avulla. Ihmisestä on tullut planeetan sopeutuvin olento ja ainoa, joka pystyy siirtämään tietoa maailmasta suurille massoille. Ihmispopulaatio on kokenut räjähdysmäistä kehitystä, koska se voi oppia ja hyödyntää hyvin suurta osaa maailman struktuurista ja välittää sitä toisille ihmisille. Ihminen menestyy kaikkialla, olipa sitten kyseessä sademetsä, autiomaa, jäinen tundra tai betoniviidakko. Suuren aivokuoren ja kielen yhdistelmä on johtanut ihmislajin jatkuvaan menestykseen. (Hawkins 2005, 186.)

Älyn määritelmä

Älykkyys on käsite, jota käytämme jokapäiväisessä elämässämme, ja sillä näyttää olevan melko konkreettinen, vaikkakin ehkä naiivi merkitys. Voidaan esimerkiksi sanoa, että hyvät arvosanat koulukokeesta saanut on älykäs, tai ehkä kissaa joka on oppinut menemään piiloon kuultuaan sanan eläinlääkäri, voidaan sanoa älykkääksi. Vaikka tämä intuitiivinen käsite ei aiheuta ongelmia arkipäiväisessä käytössä, niin sen täsmällinen määrittely on vaikeaa. (Legg ja Hutter 2007b, 3.) Toimiva määritelmä voi auttaa ilmiön ymmärtämistä ja manipulointia. Ongelma on akuutti etenkin, kun tarkastellaan systeemejä, jotka ovat merkittävästi erilaisia kuin ihmiset. Miten arvioidaan eläinten, tietokoneiden tai mahdollisten maapallon ulkopuolisten olioiden älykkyyttä?

Älykkyys ja sen mittaaminen on tunteita nostattava ongelma mm. siksi, että se liittyy läheisesti kysymykseen, kuinka ihmisiä arvioidaan. Onko työntekijä toista parempi? Ovatko miesten keskiarvoiselta älykkyys korkeampi kuin naisten? Ovatko valkoiset älykkäämpiä kuin mustat? Tämän tuloksena älykkyystestit ja niiden luojat ovat usein joutuneet yleisön kritiikin kohteeksi. On poliittisesti latautunut kysymys, voiko yksinkertainen testi olla osatekijänä tietyn rodun, sukupuolen, sosiaalisen luokan, kulttuurin hylkäämisessä. (Legg ja Hutter 2007b, 3.)

On kehitetty määritelmiä ja testejä ihmisten älykkyyden mittaamiseen, vertailuun ja esimerkiksi erityisen älykkäiden tai "lahjakkaiden" tunnistamiseksi, esimerkiksi jotta heidän koulutustaan voitaisiin kehittää tai heikommin menestyviä auttaa (mm. Binet ja Simon 1916). Älykkyyttä mitataan usein mm. koulutukseen, liiketoimintaan ja sodankäyntiin liittyvillä aloilla, koska se ennustaa luotettavasti käyttäytymistä: Älykkyys korreloi merkittävästi koulutuksen onnistumisen ja suoritustason kanssa. Reen ja Earlesin (1992, 86- 89) mukaan yleinen älykkyys g on paras yksittäinen työssä suoriutumisen ennustaja. Korrelaatio vaihtelee välillä r = 0.33- 0.76. G korreloi vahvasti myös mm. tulojen, sosiaalisen aseman ja koulutustason kanssa (Geary 2005). Älykkyysosamäärä eli ÄO (engl. intelligence quotient, IQ) on mitta, joka ilmaisee yksilön älykkyyden suhteellisen aseman toisten samaan populaatioon kuuluvien joukossa. Sen mitta-asteikolla ei ole luonnollista alkuperää tai yksikköä, vaan perustuu konventioon. Se yrittää mitata multidimensionaalista kvaliteettia (eri kykyjen joukkoa), mutta ongelmana on, että tätä ei aina tehdä samalla tavalla, mikä vaikeuttaa pätevien vertailujen tekemistä. Yleinen älykkyys (engl. general intelligence, g) muistuttaa älykkyysosamäärää, mutta yhdistää erilaisia älykkyyksiä yksidimensionaaliseksi mittaluvuksi. Molemmat ovat epäsuoria kuvauksia älykkyydestä, koska kertovat vain vähän siitä, mitä prosesseja aivoissa tapahtuu älykkään toiminnan synnyttämiseksi. Kehittyneempi kuvaus yhdistäisi älykkyyden suoremmin relevantteihin aivojen prosesseihin ja rakenteisiin. (Bartholomew 2004, 143- 144.) Kriitikoiden mukaan ihmisillä on yleensä jonkin verran erilainen ja laajempi käsitys älykkyydestä kuin se, mitä älykkyystesteissä mitataan. Toiseksi voidaan väittää, että psykometriassa ylipainotetaan g:tä. Kolmanneksi mitä tahansa normaalijakaumaa käytetäänkin, määritelmän mukaan yksilön älykkyysosamäärä ilmaisee aina kognitiivista suoritusta suhteessa johonkin laajempaan ryhmään. Selvästi tämä on ongelmallista koneiden yhteydessä, koska jotkut koneet voivat olla hyvin eri suuruusluokkaa kuin toiset. Niinpä laajempaan käyttöä varten absoluuttinen mitta on merkityksellisempi kuin perinteiset älykkyysosamäärää mittaavat testit. (Legg ja Hutter 2007b, 7.) Älykkyystestien rakenteesta ja historiasta ks. Kaufman (2000). Eläinten älykkyyden mittaamisesta ks. Zentall (2000) ja Herman ja Pack (1994).

Yksi keskeinen kysymys on, pitäisikö älykkyyttä pitää yhtenä vai monena kykynä. Älykkyyttä monena kykynä pitävien lähestymistapojen historia alkoi Thurstonesta (1938) ja häntä seurasi mm. Sternberg (1985). Nykyään suosittu lähestymistapa on Gardnerin (1993) moniälykkyysteoria, jonka mukaan ihmisen älykkyyden komponentit ovat riittävän erillisiä, jotta niitä voidaan pitää eri älykkyyksinä. Ihmisen aivojen rakenteeseen perustuen hän pitää eri älykkyyksinä kielellistä, musiikillista, loogis-matemaattista, avaruudellista ja visuaalista, liikunnallista, intrapsyykillistä (itseymmärrys) ja interpsyykillistä (kyky olla toisten kanssa vuorovaikutuksessa) kykyä, ja lisäsi tähän joukkoon myöhemmin kahdeksannen, ns. luonnon ymmärtämisen kyvyn. Vaikka Gardnerin teoria on kiinnittänyt suuren yleisön huomion, jää nähtäväksi, missä määrässä sillä on pysyvää vaikutusta ammattipiireissä. (Legg ja Hutter 2007b, 10.) Toisessa ääripäässä ovat Spearman (1927) ja hänen seuraajansa, jotka pitävät älykkyyttä yhtenä kykynä. Tässä älykkyys ajatellaan hyvin yleiseksi mentaaliseksi kyvyksi, joka on kaikkien muiden mentaalisten kykyjen taustalla. Todisteena tästä he pitävät sitä, että yksilön suoritustaso järkeilyssä, assosioinnissa, kielellisessä ja tilallisessa ajattelussa, jne. korreloivat positiivisesti. Tätä korrelaatiota Spearman nimittää g-faktoriksi. G viittaa yleiseen älykkyyteen (engl. general intelligence). Nykyään Spearmanin työn jatkajia on mm. Gottfredson (2002). (Legg ja Hutter 2007b, 10.)

On myös kehitetty määritelmiä ja testejä koneiden mahdollisen älykkyyden tunnistamiseksi, mittaamiseksi ja vertailemiseksi toisten koneiden ja ihmisten välillä (mm. Turing 1950, Wang 1995, Kurzweil 2000). Tämä on ollut lähinnä tekoälytutkimuksen lähestymistapana. Vaikka keskustelu voi olla vähemmän poliittisesti latautunutta, keskeiset aiheet ovat vieläkin vaikeammat. Koneilla voi olla fyysisiä toteutuksia, sensoreita, aktuaattoreita, kommunikaation keinoja ja ympäristöjä, jotka ovat aivan erilaisia kuin meidän kokemamme. Tämä tekee koneälyn käsitteen yleisestä määrittelystä erityisen vaikean. Joissakin tapauksissa koneilla voi olla ominaisuuksia, joita voidaan samastaa ihmisälykkyyden kanssa ja silloin on järkevää kuvailla konetta älykkääksi. Toisissa tilanteissa tämä näkemys on liian rajoittunut ja antroposentrinen (ihmiskeskeinen). Ideaalisesti voitaisiin mitata hyvin monien eri systeemien älykkyyttä: ihmisten, koirien, robottien, eksperttisysteemien, luokittelusysteemien ja ennusteita tekevien algoritmien. (Legg ja Hutter 2007b, 3-4.)

Kolmanneksi on yritetty kehittää universaalia älykkyyden määritelmää, joka sopisi kaikkien kohteiden älykkyyden tunnistamiseen, mittaamiseen ja vertailuun, olivat ne sitten eläimiä tai ihmisiä, koneita tai kenties ulkoavaruuden olioita (mm. Legg ja Hutter 2007a, Fritz 2007). Filosofisesta näkökulmasta tämä pyrkimys on tärkein, ja siitä voi tinkiä vain, jos se osoittautuu liian vaikeaksi. Kattavampi käsite (ceteris paribus) on tehokkaampi väline maailman selittämiseen. Onneksi uskottavia tämän tyypin vaihtoehtoja on olemassa. Leggin ja Hutterin mielestä määritelmä ei saisi olla rajoittunut tiettyihin aisteihin, ympäristöihin tai tavoitteisiin, eikä myöskään tiettyyn laitteistoon (teknisen toteutuksen tapaan), kuten silikoniin tai biologisiin neuroneihin. Sen tulee perustua periaatteisin, jotka eivät todennäköisesti muutu ajan mukana. Lisäksi älykkyyden määritelmä olisi ideaalitapauksessa formaalisti ilmaistavissa, objektiivinen ja sen perusteella voitaisiin kehittää käytännön älykkyystestejä. (Legg ja Hutter 2007b, 4.)

Huolimatta pitkästä tutkimuksen ja väittelyn historiasta, älykkyyden standardimääritelmää ei ole saavutettu. Tämä on johtanut jotkut ajattelemaan, että älykkyyttä voidaan kuvailla jollakin tarkkuudella, mutta ei määritellä täsmällisesti (mm. Bartholomew 2004, 144- 150). Kuitenkin eri määritelmät ovat keskenään hyvin samankaltaisia. (Legg ja Hutter 2007a, 2.) Vaikka älykkyyden määritelmän ja mittaamisen yksityiskohdista kiistellään, on olemassa laaja konsensus älykkyyden tieteellisen määritelmän pääpiirteistä. On laajasti todettu, että kun standardiälykkyystestejä on oikein sovellettu ja tulkittu, ne kaikki mittaavat samaa asiaa. Lisäksi se mitä ne mittaavat, on yksilöissä melko pysyvä ominaisuus, jonka perusteella voidaan tehdä merkittäviä ennustuksia yksilön elämästä. (Legg ja Hutter 2007b, 3.) Standardi älykkyystestit ovat tilastollisesti vakaimpia ja luotettavimpia psykologisia testejä. Kysymys ei olekaan siitä, ovatko nämä testit hyödyllisiä tai käyttökelpoisia vaan siitä, mittaavatko ne todella älykkyyttä. Jotkut asiantuntijat uskovat että ne mittaavat älykkyyttä, ja toiset että ne mittaavat vain joitain älykkyyden aspekteja. (Legg ja Hutter 2007b, 5.) Esimerkkejä älykkyyden määritelmistä, joita ovat kehitelleet mm.

1) yleensä psykologien muodostamat tutkimusryhmät:

- "Älykkyys on kykyä hankkia ja käyttää tietoa" (The american heritage dictionary, fourth edition, 2000) (Legg ja Hutter 2007a, 3).

- "Älykkyys on hyvin yleinen mentaalinen kyky johon kuuluu muun muassa kyky järkeillä, suunnitella, ratkaista ongelmia, ajatella abstraktisti, ymmärtää monimutkaisia ajatuksia, oppia nopeasti ja oppia kokemuksesta" (Gottfredson 1997, 13- 23. 52: n ekspertin allekirjoittama lausunto).
- "Yksilöt eroavat toisistaan kyvyssään ymmärtää monimutkaisia ajatuksia, sopeutua ympäristöönsä, oppia kokemuksista, järkeillä ja selviytyä vaikeuksista ajattelun avulla" (American Psychological Association 1996.)

2) yksittäiset psykologit:

- "henkilöllä on älykkyyttä siinä määrin kuin hän on oppinut tai voi oppia sopeutumaan ympäristöönsä" (Colvin S. S., lainattu teoksessa: Stenberg (ed.) 2000).
- Älykkyys on tietty joukko kognitiivisia kykyjä, joiden ansiosta yksilö voi sopeutua mihin tahansa ympäristöön (Simonton D. K. 2003).
- Ihmisillä on eri tavoitteita: jotkut haluavat hyvät kouluarvosanat, toiset tulla hyväksi koripallon pelaajiksi, näyttelijäksi tai muusikoksi. Älykkyys on kyky saavuttaa mikä tahansa haluttu tavoite omassa sosiokulttuurisessa ympäristössä. (Stenberg R. J. 2003)

3) tekoälytutkijat:

- "Minkä tahansa systeemin. . . , joka käyttäytyy sopeutuakseen ympäristöönsä tavoitteidensa saavuttamiseksi, voidaan sanoa olevan älykäs." (Fogel 1995).
- "Älykkyys on kykyä käyttää optimaalisesti rajallisia resursseja - aika mukaan lukien - tavoitteiden saavuttamiseksi" (Kurzweil 2000).
- "Älykkyys on informaatiota prosessoivan systeemin kykyä sopeutua ympäristöönsä, jossa sillä on käytettävissään ei-täydellinen määrä tietoa ja resursseja" (Wang 1995).
- John McCarthyn mukaan älykkyys on laskennallinen osa kyvystä saavuttaa tavoitteita ympäristössä (McCarthy 2004, 1).

Ongelmaa voidaan lähestyä seuraavasti: Otetaan eksperttien antamia älykkyyden määritelmiä ja erotetaan niiden olennaiset ominaisuudet. Nämä piirteet formalisoidaan matemaattisesti ja tuotetaan yleinen mitta kaikille agenteille. (Legg ja Hutter 2007b, 1.) Älykkyys:

- on yksittäisen agentin ominaisuus, kun se vuorovaikuttaa ympäristönsä kanssa.
- liittyy agentin kykyyn saavuttaa tavoitteensa. Tämä edellyttää jonkin tavoitteen olemassaoloa. Tavoitteet voivat olla erilaisia. Älykkyyden kannalta olennaista on,

kykeneekö yksilö valitsemaan toimintonsa niin että se saavuttaa tavoitteensa. Mitä suurempi on yksilön kyky saavuttaa eri tavoitteita eri ympäristöissä, sitä älykkäämpi se on.

- riippuu siitä, kuinka kyvykäs agentti on sopeutumaan eri tavoitteisiin ja ympäristöihin. Yleensä yksilö ei tiedä kaikkea tavoitteen saavuttamisen kannalta relevanttia tietoa ympäristöstään, joten sen täytyy oppia ja sopeutua. (Legg ja Hutter 2007b, 11- 12.)

Nämä ominaisuudet yhdistämällä saadaan universaali älykkyyden määritelmä, jonka mukaan älykkyyden määrä kuvaa agentin kykyä saavuttaa tavoitteitaan eri ympäristöissä. Kyky oppia ja ymmärtää, ratkaista ongelmia, suunnitella jne. sisältyvät implisiittisesti tähän määritelmään. (Legg ja Hutter 2007a, 12, Legg ja Hutter 2006, 73- 80.) Fritzin käsitys älykkyydestä vastaa edellistä: Älykkyys on systeemin kykyä saavuttaa tavoitteensa (Fritz 2007, Our Definition Of Intelligence). Tarkemmin ilmaistuna *se on systeemin kykyä antaa oikea vaste tiettyyn syötteeseen jonkin tavoitteen kannalta*, ja *systeemillä on enemmän älykkyyttä, jos se saavuttaa tavoitteensa useammin tai nopeammin.* (Fritz 2007, What is intelligence? Are Learning Rates A Measure Of Intelligence? Our Definition Of Intelligence.) Määritelmä edellyttää, että systeemillä on jokin tavoite, ja että se voi toimimalla vaikuttaa ympäristöönsä saavuttaakseen tämän tavoitteen. Määritelmä sisältää kolme olennaista komponenttia: agentin, ympäristön ja tavoitteen, se siis perustuu agentti-ympäristö -viitekehykseen: Agentin ja ympäristön tulee olla vuorovaikutuksessa toistensa kanssa. Agentin tulee voida lähettää signaaleja ympäristöön ja ottaa vastaan ympäristöstä tulevia signaaleja ja ympäristön tulee voida saada signaaleja ja lähettää niitä agentille. Agentin ympäristöön lähettämiä signaaleja voidaan sanoa toiminnoiksi (engl. actions) ja ympäristöstä saamia signaaleja sanotaan havainnoiksi (engl. perceptions). (Legg ja Hutter 2007b, 15- 16.) Jotkut objektit kuten kivet, ilmamolekyylit tai vesi eivät toimi, vaan niiden liike johtuu niihin kohdistuneesta ympäristöstä tulevasta voimasta. (Fritz 2007, What is intelligence?, Acting on the environment.) Sen sijaan älykkäät systeemit toimivat, mikä tarkoittaa, että ne käyttävät energiaa muuttaakseen ympäristöään jonkin tavoitteen saavuttamiseksi. Toiminto on suhteessa systeemin saamaan aistisyötteeseen tai tietoon ympäristön nykyisestä tilanteesta. Toimiminen edellyttää tavoitetta, koska ilman tavoitetta systeemillä ei ole syytä valita syötteeseen liittyvää toimintoa. (Fritz 2007, What is intelligence?, Acting on the environment, Action.) Agentilla täytyy olla ainakin yksi tavoite.

(Legg ja Hutter 2007b, 15- 16.) Tavoite on tietty tilanne, johon systeemi pyrkii. (Fritz 2007, What is intelligence?, Acting on the environment, Objectives.) Tavoite on usein määritelty ns. menestyksen kriteeristöllä (engl. performance criteria) jonka täyttämällä agentti saavuttaa tavoitteensa. Suoritusarvo (engl. performance measure) kuvaa, kuinka hyvin toiminta täyttää menestyksen kriteerit. Tavoitteen saavuttaakseen agentti suorittaa toimintoja usein ympäristöstään saamansa palautteen perusteella. Menestyksellinen toiminta vastaa ympäristön tilan muuttumista tavoitetilaksi toiminnan avulla. Ei ole olemassa yhtä ja pysyvää menestyksen kriteeristöä, joka sopisi kaikille agenteille, vaan agenteilla on eri tavoitteita, ja menestys määräytyy niiden saavuttamisen mukaan. (Russell ja Norvig 2003, 35.) Usein tavoite voidaan jakaa alitavoitteisiin, joiden saavuttaminen siirtää systeemiä tavoitetilaa kohti. (Fritz 2007, What is intelligence?, Acting on the environment, Objectives.) Agentti voi olla älykäs ilman tavoitetta, jonka saavuttamiseen se voi käyttää älykkyyttään, ja siinäkin tapauksessa, että agentti ei halua käyttää älyään tavalla joka vaikuttaa sen ympäristöön. Molemmissa tapauksissa agentin älykkyyttä ei voida havaita. Älykkyys voidaan havaita, kun agentilla on jokin tavoite, jota se aktiivisesti ympäristöönsä vaikuttamalla yrittää saavuttaa. Mistä agentti tietää, mikä sen tavoite on? Yksi mahdollisuus on, että tieto tavoitteesta on alun perin ohjelmoitu agenttiin. Toiseksi se voidaan kommunikoida agentille. Agentti voi myös saada signaalin itsestään tai ympäristöstään, joka kertoo, kuinka hyvä sen nykyinen tilanne on. Tätä signaalia voidaan sanoa esim. palkinnoksi (engl. reward). Tällöin agentin tavoitteeksi tulee palkinnon maksimointi. (Legg ja Hutter 2007b, 15- 16.) Ilmeisesti mikä tahansa peli, haaste, ongelma tai testi voidaan ilmaista tämän yksinkertaisen viitekehyksen avulla ilman suurta vaivaa. Täytyy huomata, että tämä ns. agentti-ympäristö -viitekehys ei sano mitään siitä, miten agentti tai ympäristö todella toimii, se vain kuvaa niiden roolit. (Legg ja Hutter 2007b, 17.) Määritelmä esitetään formalisoidummassa muodossa lähteessä: (Legg ja Hutter 2007b, 17- 24). Yksi keskeinen haaste on kehittää universaalin älykkyyden käytännöllinen testi. Testin perusrakenne voisi olla seuraava: Testi toimisi arvioimalla agentin suoritusta laajalla valikoimalla simuloituja ympäristöjä, ja yhdistämällä eri ympäristöissä saadun suoritusarvon yhdeksi älykkyysarvoksi. Tämä tehtäisiin painottamalla agentin suoritus ympäristön monimutkaisuuden perusteella. (Legg ja Hutter 2007b, 39.)

Älykkyys ja rationaalisuus

Tällä tavalla määriteltynä älykkyys on lähellä sitä, mitä usein sanotaan rationaalisuudeksi. Russellin ja Norvigin mukaan agentti on rationaalinen kun se toimii oikein. Oikea toiminto on se, jonka suorittamalla agentti menestyy parhaiten, eli täyttää parhaiten menestyksen kriteeristöä (engl. performance criteria), ja saa siten korkeimman suoritusarvon (engl. performance measure). (Russell ja Norvig 2003, 34). Jokaiselle mahdolliselle perseptisekvenssille rationaalinen agentti valitsee toiminnon, jonka odotetaan maksimoivan sen menestyksen (menestyksen kriteeristön perusteella), sen perseptisekvenssin (havaintojen) ja mahdollisen sisään rakennetun tiedon perusteella. (Russell ja Norvig 2003, 35- 36.) Rationaalista päätöksentekoa on tutkittu paljon mm. tekoälyn ja taloustieteen aloilla ja tätä tietoa voidaan nyt käyttää älykkyyden ymmärtämiseen:

Se, mikä on rationaalista milloin ja missä tahansa, riippuu neljästä asiasta.
1) Menestyksen määritelmästä tai kriteeristöstä.
2) Agentin tiedosta ympäristöstään.
3) Toiminnoista, jotka agentti voi suorittaa.
4) Agentin perseptisekvenssistä nykyhetkeen saakka. (Russell ja Norvig 2003, 37,38.)

Oikein toimimisen kyvylle voidaan määritellä asteita:
1) Täydellinen agentti maksimoi *todellista toiminnan tulosta*. Se on kaikkitietävä, joten se tietää mahdollisten toimintojensa tulokset etukäteen ja voi toimia tämän tiedon mukaisesti. Sen toiminnan tulos on aina paras mahdollinen. Kaikkitietävyys on mahdotonta tai epätodennäköistä käytännössä, ellei sitten kehitetä kristallipallojen tai aikakoneiden suoritusta.
2) Rationaalinen agentti (engl. perfectly/ideal rational agent) maksimoi *odotettua toiminnan tulosta* (engl. expected utility) ympäristöstä saamansa tiedon perusteella. Se ei ole kaikkitietävä, ja voi siis tehdä vain epävarmoja ennusteita. Rationaalisilta agenteilta ei vaadita täydellistä toimintaa, koska rationaalinen valinta riippuu vain valinnan hetkeen asti saadusta perseptisekvenssistä (havainnoista). Rationaalisen agentin toiminnon valintaa koskeva tieto perustuu tähän perseptisekvenssiin. Siksi tavoitteen kannalta relevantin tiedon hankkiminen ja oppiminen on yleensä tärkeää rationaaliselle agentille, mutta ei kaikissa tapauksissa: Jos

agentilla on tarpeeksi tietoa ympäristöstään, sen ei tarvitse oppia. Jos ympäristö ei muutu, agentille riittää tarpeellisen tiedon hankkiminen, jonka jälkeen sen ei enää tarvitse oppia. Jos se tuntee ympäristön kokonaan, sen ei tarvitse havaita eikä oppia. Täydellinen rationaalisuus on erittäin vaikeaa tai mahdotonta monimutkaisissa ympäristöissä, koska laskennalliset vaatimukset ovat liian kovat.

3) Laskennallisesti rationaalinen agentti (engl. calculatively rational agent) tekee lopulta rationaalisen valinnan, kunhan se saa riittävän laskenta-ajan. Esimerkiksi laskennallisesti rationaalinen shakkiohjelma valitsee oikean siirron, mutta voi käyttää 10^50 kertaa liian pitkän ajan sen laskemiseen.

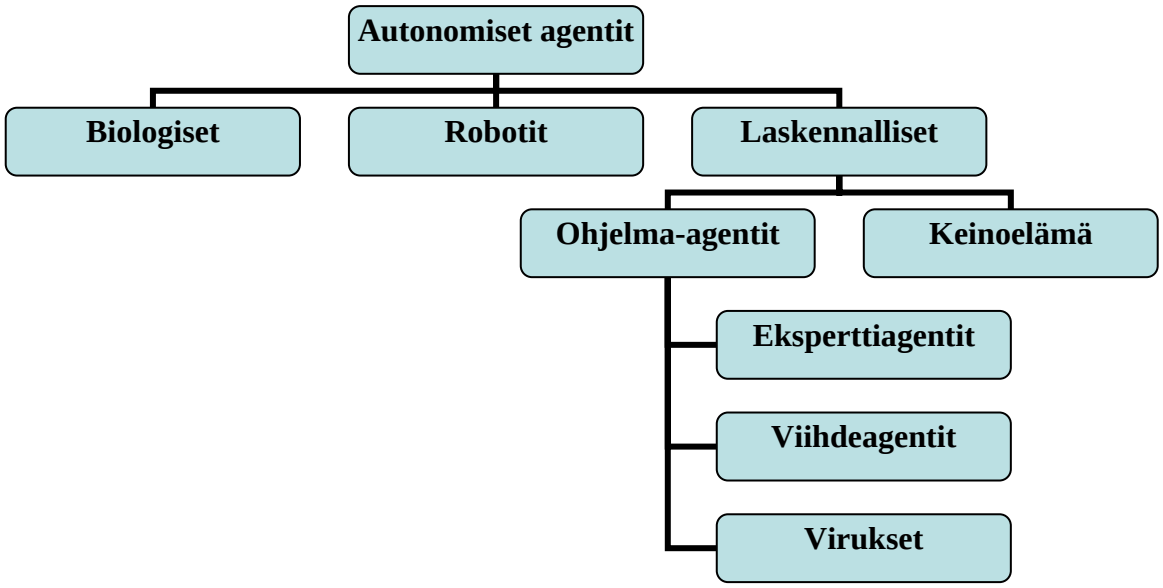
4) Rajoitetusti optimaalinen agentti (engl. bounded optimal agent, (BO)) toimii mahdollisimman hyvin sen laskennallisten resurssien puitteissa. Se ei välttämättä löydä täydellistä ratkaisua, mutta kuitenkin niin hyvän kuin sen tiedolla, muistitilalla ja nopeudella on siinä tilanteessa mahdollista. Russellin ja Norvigin mukaan rajoitetusti optimaalinen agentti on tekoälytutkimukselle sopivin tavoite, koska sen toiminta on parasta, mitä käytännössä voidaan saavuttaa. Jokaiselle tehtävälle, koneelle ja ympäristölle on olemassa BO-ohjelma. Ongelma on tietysti sen löytäminen. Tekoälytutkimus voidaan määritellä rajoitetun optimaalisuuden tutkimukseksi. (Russell ja Norvig 1995, 845- 846 ja 2003, 37) Käytännöllisestä rationaalisuudesta ks. myös Pollock (2006).

Agentilta sanotaan puuttuvan autonomiaa siinä määrin, kuin sen käyttäytyminen tukeutuu suunnittelijan siihen ohjelmoimaan tietoon eikä agentin omiin persepteihin. Rationaalisen agentin täytyy olla autonominen, koska sen täytyy voida oppia kompensoidakseen epätäydellistä tai virheellistä aikaisempaa tietoa (engl. prior knowledge). Käytännössä agentit ovat harvoin täysin autonomisia alusta asti. Tämä johtuu siitä, että jos agentilla on vähän tai ei ollenkaan aikaisempaa tietoa, se ei voisi muuta kuin toimia satunnaisesti. Eläimillä on usein refleksejä säilyäkseen hengissä siihen asti kunnes ne kykenevät oppimaan itse. Samoin on hyödyllistä rakentaa keinotekoisille älykkäille systeemeille tietoa ennen kokemusta, ja kyky oppia. Kun agentti on saanut riittävästi kokemuksia ja oppinut niistä, se voi tulla riippumattomaksi aikaisemmasta tiedosta. Oppimisen kyvyllä varustettu agentti voi menestyä monenlaisissa eri ympäristöissä. (Russell ja Norvig 2003, 38.)

Vaikka älykkyys ja rationaalisuus niin kuin ne edellä on määritelty, ovat hyvin lähellä toisiaan, ne voidaan ehkä erottaa seuraavasti. Älykkyys kuvaa agentin yleistä kykyä saavuttaa tavoitteita. Koska intuitiivisesti agentin älykkyyteen eivät kuulu sen ulkoiset resurssit (ympäristön hyödylliset osat) eivätkä välittävät resurssit (sensorien ja aktuaattorien hyödylliset kyvyt), vaan vain sisäiset resurssit (hyödylliset sisäiset rakenteet ja prosessit), niin rajoitetummin älykkyys voidaan määritellä agentin laskennalliseksi kyvyksi muodostaa oikea toiminto tavoitteen saavuttamiseksi. Rationaalisuus taas voidaan määritellä seuraavasti: Rationaalinen agentti valitsee aina sille mahdollisista toiminnoista aina sen, joka parhaiten auttaa sitä saavuttamaan tavoitteensa. Muuten agentti on irrationaalinen. Siten rationaalisuus olisi osa älykkyyttä: Älykkyys = (kyky muodostaa toimintoja) + (rationaalisuus, eli kyky valita niistä oikea).

3.2.3 Älykkäät systeemit

Älykkäitä systeemejä ovat systeemit, jotka ovat älykkäitä: mm. ihmiset, korkeammat eläimet, robotit, alkuasukasheimot, yritykset, valtiot ja mahdollisesti jotkin ulkoavaruuden oliot. (Fritz 2007, Create a scientific ethics.) Käsitteitä agentti ja älykäs systeemi käytetään tässä tekstissä samassa merkityksessä, vaikka muuten näiden välillä voikin olla jonkin verran eroa. Sana agentti tulee latinan sanasta agere: tehdä, toimia. Agentti on yksinkertaisesti jokin, mikä toimii. Näin laajan määritelmän mukaan esimerkiksi entsyymi on agentti. Russel ja Norvig täsmentävät sitä seuraavasti: Agentti on mikä tahansa, mitä voidaan tarkastella sensoreillaan ympäristöään havainnoivana ja siinä aktuaattoreillaan toimivana (Russell ja Norvig 2003, 32). Laajempaa toiminnan käsitteen tutkimusta: (Wilson 2008). Älykäs systeemi on systeemi, joka voi muokata käyttäytymistään ympäristössään tavoitteensa mukaisesti. Agentit, jotka eivät voi muokata käyttäytymistään tietyn tavoitteen mukaisesti, eivät ole älykkäitä systeemejä.



Kaavio 1: Agenttien taksonomia Franklinin ja Graesserin (1996) mukaan. Biologiset, robotit ja laskennalliset agentit näyttävät kuuluvan eri lajeihin (natural kinds) (Keil 1989). Laskennalliset agentit voidaan jakaa ohjelma-agentteihin ja keinoelämään. Ohjelma-agentit voidaan jakaa tiettyä tehtävää suorittamaan tarkoitettuihin eksperttiagentteihin, viihdetarkoitukseen luotuihin agentteihin, ja tietokoneviruksiin. (Franklinin ja Graesserin 1996) Taksonomia ei ole kovin kehittynyt. Jaot eivät tunnu kovinkaan välttämättömiltä.

Fritz kumppaneineen on muodostanut älykkään systeemin käsitteensä keinotekoisten älykkäiden systeemien rakentamisessa saatujen käytännön kokemusten perusteella. Se on tulosta monista muutoksista, joita alkuperäiselle hypoteesille tehtiin. Teorian ja ohjelman muuttelu ja niiden kokeilu käytännössä opettivat heille mitä älykkääseen toimintaan tarvitaan, ja miten älykkäiden systeemien käsite tulisi määritellä. (Fritz 2007, Preliminary remarks on the theory of intelligent systems.) Määritelmän on tarkoitus olla hyödyllinen ja hänen mukaansa se on luultavasti myös paranneltavissa. (Fritz 2007, Definition of the intelligent system.)

Fritzin mukaan seuraava älykkään systeemin käsite näyttää selittävän melko hyvin joitakin keinotekoisiin (esimerkiksi älykkäät tietokoneet, robotit) ja luonnollisiin älykkäisiin systeemeihin (esimerkiksi eläimet, mukaan lukien ihmiset) sekä niiden yhteisöihin liittyviä ilmiöitä ja sitä voidaan käyttää niiden tutkimuksessa, ja on myös hyödyllinen joidenkin filosofisten kysymysten ymmärtämiseksi. (Fritz 2007, Definition of the intelligent system,

Artificial intelligent systems, Persons as intelligent systems, Ethics as a science, A scientific philosophy.)

Älykäs systeemi:
1) On systeemi.
2) Se oppii olemassaolonsa aikana. Toisin sanoin se aistii ympäristöään ja oppii joka tilanteessa toiminnon, joka auttaa sitä saavuttamaan tavoitteensa.
3) Se toimii, joko mielessään tai ulkoisesti, ja toimimalla saavuttaa tavoitteensa useammin kuin satunnaisesti.
4) Se kuluttaa energiaa sisäisiin prosesseihinsa ja ulkoiseen toimintaansa. (Fritz 2007, Definition of the intelligent system.)

Määritelmä edellyttää että:
1) Systeemin täytyy olla olemassa.
2) Systeemillä täytyy olla ympäristö, jonka kanssa se voi vuorovaikuttaa.
3) Sillä täytyy olla kyky saada kommunikaatioita ympäristöstään, että se voisi tehdä mallin nykyisestä tilanteesta.
4) Sillä täytyy olla tavoite.
5) Saavuttaakseen tavoitteensa sen täytyy valita vaste (reaktio).
6) Sen täytyy kyetä oppimaan, eli muuttamaan ja luomaan vasteita, koska sama vaste voi joskus olla hyödyksi ja joskus haitaksi.
7) Sen täytyy kyetä toimimaan, esimerkiksi jaloilla tai käsillä, mutta yleisesti sanottuna efektoreilla (tai aktuaattoreilla (Russell ja Norvig 1995, 777)). Jos joku näistä edellytyksistä puuttuu, niin älykäs systeemi ei ilmeisesti voi toimia. Tämä voidaan varmistaa tarkistamalla jokaisen edellytyksen kohdalla, toimisiko älykäs systeemi ilman sitä. (Fritz 2007, Definition of the intelligent system.)

Agentin määritelmässä tulevat ilmi jokaisessa agentissa välttämättömät fysikaaliset osat: Agentti = arkkitehtuuri + ohjelma. Arkkitehtuuriin kuuluvat ainakin sensorit, aktuaattorit ja jonkinlainen laskentalaitteisto, jolla agentin ohjelma käy (engl. run). Agentin ohjelma liittää toiminnot havaintoihin (tämä prosessi voi tietysti olla hyvin monimutkainen). Ohjelman pitää

sopia arkkitehtuuriin: Jos ohjelma suosittaa toiminnoksi kävelyä, arkkitehtuurilla on hyvä olla esim. jalat tätä varten. (Russell ja Norvig 2003, 44.) Ilman sensoreja (aisteja) älykäs systeemi ei voi saada tietoa ympäristöstään, jolloin se ei voi valita vastettaan suhteessa ympäristön ominaisuuksiin (edellytys 3). Ilman aivoja (laajassa mielessä laskentalaitteistoa ja siinä käyvää agentin ohjelmaa) se ei voi tallentaa ja prosessoida syötteitään ja tuottaa niiden perusteella sen tavoitteisiin ja ympäristöön sopivaa vastetta (edellytykset 4, 5, ja 6). Ilman aktuaattoreja se ei voi suorittaa valitsemiaan vasteita, jos niihin kuuluu ympäristön manipuloiminen (edellytys 7).

Älykäs systeemi on ns. syöte-vaste -systeemi (engl. stimulus-response system) (Fritz 2007, Overview of the intelligent system); sen yleisenä toimintaperiaatteena on toistaa syöte-vaste -sykliä: Se saa ympäristöstään syötteen, valitsee siihen (tavoitteensa kannalta) sopivan vasteen, ja suorittaa sen. Vaste yleensä muuttaa ympäristön tilaa, ja systeemi saa uuden syötteen, ja sykli alkaa alusta. Kaikkien älykkäiden systeemien toimintaperiaate on sama, vaikkakin se voidaan käytännössä toteuttaa monin eri tavoin. Yksityiskohtaisemmin esitettynä älykäs systeemi toimii toistaen seuraavien askelten muodostamaa ohjelmaa:

1) **Se saa syötteen:** Syöte on älykkään systeemin ympäristöstään vastaanottamien kommunikaatioiden summa. Sitä osaa älykkäästä systeemistä, joka ottaa vastaan kommunikaatioita ympäristöstä, voidaan sanoa sensoriksi, aistiksi tai aistinelimeksi. Älykkäällä systeemillä on yleensä hyvin rajallinen määrä sensoreita, ja ne voivat ottaa vastaan yleensä vain tietyn tyyppisiä kommunikaatioita, esimerkiksi ääntä tai valoa (ilmanpaineen vaihteluita tai sähkömagneettista säteilyä tietyllä taajuus- ja voimakkuusalueella). (Fritz 2007, Overview of the intelligent system, Stimulus.) Näiden ja muiden, esimerkiksi aivojen laskennallisten resurssien rajoitusten takia älykkäällä systeemillä voi olla vain hyvin rajoitetusti tietoa ympäristöstään. Agentin sensorien syötettä tiettynä aikana sanotaan myös perseptiksi (engl. percept). Perseptisekvenssi (engl. perceptual sequence) on kaikki agentin havainnot (syötteet) nykyhetkeen asti. (Russell ja Norvig 2003, 32- 34.)
2) **ajattelee:** Syötteen perusteella aivot rakentavat muistissa olevien, aikaisempien kokemusten perusteella muodostettujen käsitteiden avulla esityksen sen nykyisestä tilanteesta. (Fritz 2007, Present situation, What is a "Present situation?".) Ensiksi se arvioi, onko nykyinen tila sen tavoitetila. Jos on, se joko ei tee mitään, tai sitten muodostaa uuden tavoitteen. Muussa

tapauksessa se valitsee nykyiseen tilanteeseen sopivan vastesäännön muististaan. Sopiva vastesääntö on se, jonka suorittaminen auttaa älykästä systeemiä saavuttamaan tavoitteensa. Käytännössä nykyinen tavoite on usein alitavoite, joka on päätavoitteen saavuttamiseen johtavan suunnitelman osa. (Fritz 2007, Overview of the intelligent system, Response, Response rule, Brain, Mind, Experiences.) Älykkään systeemin käyttäytyminen tunnetaan, kun tiedetään sen valitsemat vastesäännöt kaikille mahdollisille perseptisekvensseille (syötesekvenssille). Agentin funktio (engl. agent function) liittää minkä tahansa annetun perseptisekvenssin johonkin toimintoon. Teoriassa jokaiselle agentille voidaan ilmoittaa agentin funktio taulukoimalla kaikki mahdolliset agentin sensorien syötteet ja niihin liittyvät toiminnot. Suurelle osalle agenteista tällainen taulukko olisi joko erittäin suuri tai ääretön, ellei syötteiden määrää jollakin tavalla rajoiteta. Agentin funktion toteutus on agentin ohjelma (engl. agent program), joka on konkreettinen agentin arkkitehtuurissa käyvä ohjelma. (Russell ja Norvig 2003, 32-34.)

3) **ja palauttaa vasteen:** Se suorittaa valitsemansa vastesäännön vasteosan aktuaattoreillaan, esimerkiksi jaloillaan ja käsillään. Tämä aiheuttaa usein muutoksia ympäristöön (joka voi muuttua muutenkin), ja älykäs systeemi siirtyy takaisin askeleeseen (1). (Fritz 2007, Overview of the intelligent system, Stimulus, Response, Response rule, Brain, Mind, Experiences.)

Ensisilmäykseltä voi vaikuttaa uskomattomalta että kaikki älykkäiden systeemien (ihmisenkin) toiminnot ovat syöte-vaste -mekanismeja. Monimutkaisuuden takia näitä prosesseja voi olla vaikea hahmottaa. Tietokoneen toiminta voidaan jakaa monen eri tason tapahtumiseksi. Koodi rivit ryhmitellään blokeiksi, joita kutsutaan aliohjelmiksi, jotka ryhmitellään moduuleiksi, ja yhdistetään sovelluksiksi. Esimerkiksi tavallisessa taulukkolaskentaohjelmassa on niin monta aliohjelmaa ja moduulia, että yksi ihminen ei pysty ymmärtämään sitä kokonaan. Yhden pikselin saaminen näytölle edellyttää useita koodirivejä. Kokonaisen näytön piirtäminen laskentataulukkoon edellyttää, että tietokone suorittaa miljoonia konekielisiä käskyjä, jotka levitetään satoihin aliohjelmiin. Aliohjelmat herättävät muita aliohjelmia toistuvasti. On vaikea uskoa, että tietokone löytäisi oikean reitin moduulien ja aliohjelmien labyrintissä ja suorittaisi kaikki konekieliset käskyt niin nopeasti. (Hawkins 2005, 179- 180.) Samalla tavalla voi näyttää siltä kuin ihmisaivojen toiminta ei olisi vain syöte-vaste -toimintaa. Syöte-vaste -toiminta tapahtuu neuronien tasolla. Korkeammalla, neuronien joukkojen tasolla ovat yksinkertaiset käsitteet ja vastesäännöt. Näihin yksinkertaisiin vastesääntöihin perustuen aivot

oppivat seuraavan tason vastesääntöjä jne., kunnes ne lopulta saavuttavat mielen toimintojen monimutkaisen vastesääntöjen rakenteen. (Fritz 2007, Learned Mental Activity.)

Yksityiskohtaisempaa tietoa älykkäistä systeemeistä

Älykkäiden systeemien teorian ja sen kanssa yhteensopivan nykytiedon avulla voidaan ymmärtää paljon ihmisten, eläinten ja mm. robottien toiminnasta niin yksin kuin yhteisöissäkin. Olen koonnut selityksiä joistakin tärkeistä ominaisuuksia, joita useilla älykkäillä systeemeillä on. Ominaisuuksien toteutukset voivat olla erilaisia (esimerkiksi ihmisten ja robottien kognitio). Kaikkia toteutuksia ei voida esittää, vaan keskitymme niiden yleisiin kuvauksiin.

Ympäristöt

Systeemin ympäristö on se osa universumia, joka on kommunikaatiossa systeemin kanssa, mutta ei ole osa sitä (Fritz 2007, Glossary). On olemassa erilaisia ympäristöjä. Ympäristön laatu vaikuttaa suoraan siihen, millainen siinä menestyvän agentin täytyy olla. Agentin menestymisen kannalta ei ole tärkeää eroa sillä, onko ympäristö luonnollinen vai keinotekoinen. Ympäristön haasteellisuuden ratkaisee agentin käyttäytymisen, ympäristön, ja menestyksen määritelmän väliset suhteet (Russell ja Norvig 2003, 39). Joitakin usein agentin menestyksen kannalta tärkeitä ympäristöjen ominaisuuksia:

- Kokonaan havaittava tai osittain havaittava (engl. fully observable vs. partially observable). Ympäristö on kokonaan havaittava, jos agentti voi sensoreillaan saada selville ympäristön tilan täydellisesti, jokaisena ajanhetkenä. Se on efektiivisesti kokonaan havaittava, jos agentti voi sensoreillaan saada selville kaikki toiminnon valitsemisen kannalta relevantit ympäristön aspektit. Relevanttius riippuu menestyksen määritelmästä. Kokonaan havaittavissa olevat ympäristöt ovat agentin kannalta sikäli helppoja, että niissä sen ei tarvitse tallentaa ympäristön tilaa koskevaa osittaista tietoa,

joista sen täytyisi koota käsitys ympäristön nykyisestä tilasta. Ympäristö voi olla osittain havaittavissa esimerkiksi häiriöllisten (noisy), epätarkkojen (inaccurate) tai puutteellisten sensorien takia tai jos se on liian kompleksinen. (Russell ja Norvig 2003, 41.)

- Deterministinen tai stokastinen (engl. deterministic vs. stochastic). Ympäristö on deterministinen, jos ympäristön seuraava tila johtuu sen edellisestä tilasta. Muussa tapauksessa se on stokastinen (tilastollinen). Deterministisessä, kokonaan havaittavassa ympäristössä agentin ei tarvitse huolehtia epävarmuudesta. Ympäristö voi vaikuttaa stokastiselta, jos se on osittain havaittavissa. Näin voi käydä esimerkiksi silloin, kun ympäristö on kompleksinen, jolloin agentin on vaikea seurata (engl. keep track of) kaikkia ympäristön aspekteja. Suurin osa ympäristöistä ja niiden tiloista on niin kompleksisia, että ei ole agentin kannalta merkittävää, ovatko ne deterministisiä vai stokastisia, koska käytännössä niitä käsitellään stokastisina. (Russell ja Norvig 2003, 41, 43.)

- Episodinen tai sekventiaalinen (engl. episodic vs. sequential). Episodisissa ympäristöissä agentin kokemukset jakaantuvat itsenäisiin jaksoihin ja seuraava episodi ei johdu edellisistä episodeista. Jokaisessa episodissa agentti havaitsee ja suorittaa toiminnon. Agentin tietyssä episodissa valitsema toiminto johtuu vain siitä eikä muista episodeista. Esimerkiksi monet luokittelutehtävät ovat episodisia: Jos agentin tehtävänä on erottaa tuotteen vialliset osat kun ne kulkevat liukuhihnalla, sen edellinen päätös pitää jotakin osaa kunnossa olevana tai viallisena, ei vaikuta sen seuraavaan päätökseen. Sekventiaalisessa ympäristössä nykyinen päätös voi vaikuttaa kaikkiin tuleviin päätöksiin. Esimerkiksi shakin pelaaminen ja taksin ajaminen ovat sekventiaalisia: kummassakin lyhyen aikavälin toiminnoilla voi olla pitkän aikavälin seurauksia. Episodiset ympäristöt ovat paljon yksinkertaisempia kuin sekventiaaliset ympäristöt, koska niissä agentin ei yleensä tarvitse tehdä kovin pitkiä ja monimutkaisia suunnitelmia. (Russell ja Norvig 2003, 42.)

- Staattinen tai dynaaminen (engl. Static vs. dynamic). Jos ympäristö voi muuttua agentin suunnitellessa toimintoaan, ympäristö on agentin kannalta dynaaminen. Muuten se on staattinen. Staattiset ympäristöt ovat helpompia agentille, koska sen ei tarvitse seurata ympäristöään yrittäessään valita toimintoaan, eikä sen tarvitse välittää ajan kulumisesta. Dynaamiset ympäristöt voivat muuttua agentin valitessa toimintoaan, ja jos agentti ei ole valinnut toimintoaan ympäristön muuttuessa, tulos on sama kuin jos agentti olisi päättänyt olla toimimatta. (Russell ja Norvig 2003, 42.)

- Diskreetti tai jatkuva (engl. discrete vs. continous). Diskreetti/jatkuva erottelua voidaan soveltaa ympäristön tilaan, tapaan jolla aika käsitetään ja agentin persepteihin ja toimintoihin. Esimerkiksi shakkipeli on ympäristö, jossa on äärellinen määrä diskreettejä tiloja. Shakissa perseptit ja toiminnot ovat diskreettejä. Taksin ajaminen on ongelma, jossa tilat ovat jatkuvia ja aika on jatkuva. Taksin ja muiden ajoneuvojen nopeus- ja paikkaparametrien arvot muuttuvat ilman katkoja. Taksia ajavan agentin toiminnot, esimerkiksi ohjaaminen jne. , ovat myös jatkuvia. (Russell ja Norvig 2003, 42.)

- Yksiagenttinen tai moniagenttinen (engl. single agent vs. multiagent). Esimerkiksi agentti, joka ratkaisee ristisanatehtävää, on yksiagenttisessa ympäristössä, ja agentti, joka pelaa shakkia, on kaksiagenttisessa ympäristössä. Shakki on ympäristö, jossa kilpaillaan (engl. competitive environment). Kun agentti ajaa taksia, se on moniagenttisessa ympäristössä, jossa törmäysten välttäminen maksimoi kaikkien agenttien menestystä. Taksia ajava agentti on ympäristössä, jossa tehdään osittain yhteistyötä (engl. cooperative environment). (Russell ja Norvig 2003, 42.)

Vaikein ympäristö on osittain havaittavissa, stokastinen, sekventiaalinen, dynaaminen, jatkuva ja moniagenttinen (Russell ja Norvig 2003, 43), ja monien agenttien harmiksi useimmat reaaliset ympäristöt ovat juuri tällaisia.

Aistit

Kommunikaatio on materian tai energian liikettä kahden universumin osan välillä. Materia tai energia voi kantaa informaatiota. Informaatio on käsitteiden ja vastesääntöjen summa, joka voidaan erottaa kommunikaatiosta. (Fritz 2007, Communication, information.) Älykäs systeemi saa ympäristöstä kommunikaatioita eri energiamuotoja tunnistavien aistinelintensä eli sensoriensa kautta. Älykkäillä systeemeillä on usein hyvin rajallinen määrä sensoreita, ja nekin ovat yleensä rajoittuneita vain tietyn tyyppisten kommunikaatioiden vastaanottamiseen, koska rajalliset resurssit kannattaa käyttää vain varmasti relevantin tiedon hankkimiseen ja käsittelyyn. Nämä tosiasiat rajoittavat älykkään systeemin vastaanottamien kommunikaatioiden tyyppejä ja määrää ja siten myös informaatiota ja tietoa, joka sillä voi ympäristöstään olla. Esimerkiksi silmä tai televisiokamera voi ottaa vastaan vain tietyn taajuusalueen elektromagneettista säteilyä, mutta ei mitään muuta. Tietyn tyyppiset kommunikaatiot ovat tärkeitä joillekin älykkäille systeemeille ja toisen tyyppiset toisille. Esimerkiksi joillakin kaloilla on aistinelimet, jotka ottavat vastaan kommunikaatioita niiden ympäristön sähkökentistä, jotkin linnut voivat aistia magneettikenttiä ja jotkut keinotekoiset älykkäät systeemit voivat aistia röntgensäteitä. Ihmisen sensoreita ovat mm. silmät, korvat ja muut aistinelimet, ja aktuaattoreita kädet, jalat, suu, jne. (Fritz 2007, Frontiers of systems and their crossing, Sense organs, limitations.) (Ks. myös Viitala 2004, 23- 44.) Robotin sensoreina voivat olla esimerkiksi kamera tai tutka, ja aktuaattoreina erilaisia moottoreilla toimivia osia. Ohjelma-agentti (engl. software agent) saa sensorien syötteenä näppäimenlyöntejä, kansioiden sisältöjä (engl. file contents) ja nettipaketteja (network packets), ja toimii ympäristössään kirjoittamalla tiedostoja ja lähettämällä nettipaketteja. (Russell ja Norvig 2003, 32.)

Aivot

Käsittelemme yleisimpiä aivojen toimintoja, joita älykkäillä systeemeillä voi olla. Välillä keskitytään ihmisen erityispiirteisin, mutta useimmiten siihen, minkälaisia yleisiä periaatteita näillä toiminnoilla on, riippumatta varsinaisesta toteutustavasta. Aivoilla tarkoitetaan tässä sitä fysikaalista laitteistoa, joka toteuttaa agentin (älykkään systeemin) ohjelman, eli liittää sopivat toiminnot havaintoihin. Aivot on älykkään systeemin fysikaalinen osa jossa mielen prosessit

tapahtuvat (Fritz 2007, Glossary: Brain). Agentti voidaan jakaa arkkitehtuuriin ja ohjelmaan. Aivot voidaan määritellä siksi arkkitehtuurin osaksi, jossa agentin keskitetty informaationprosessointi tapahtuu. Biologisilla eliöillä, joilla on keskushermosto, aivot kehittyivät koska keskitetty informaationprosessointi tuotti paremman päätöksentekokyvyn, jonka hyödyt kelpoisuuteen painoivat niiden aineenvaihdunnallisia ja muita kustannuksia enemmän. (Gintis 2007, 1-3.) Sama hyöty on syynä aivojen rakentamiseen myös keinotekoisille älykkäille systeemeille.

Ihmisten aivojen perusmekanismit ovat kaikkien aivotoimintojen taustalla olevia mekanismeja (Fritz 2007, Fundamental mechanisms of the brain), jotka ovat tarpeellisia aivojen varsinaisen tehtävän kannalta. Evoluution aikana aivot ovat kehittyneet sellaisiksi, että nämä toiminnot ilmaantuvat. Ne ovat valmiina ihmisen syntyessä, jotta aivot voivat alkaa ohjaamaan vartaloa ja oppimaan. Syntyessä ihmisellä ei ole käsitteitä eikä vastesääntöjä, muita kuin vastesääntöjä muistuttavat vaistot. Täsmällisemmin sanottuna aivot alkavat toimia ennen syntymää, kun ne ovat kasvaneet tarpeeksi. (Fritz 2007, Fundamental physiological mechanisms.) Aivojen perusmekanismit ovat:

1) Aistidatan vastaanottaminen.
2) Aistidatan muuttaminen käsitteiksi, ja käsitteiden tallentaminen.
3) Monen toisiinsa liittyvän käsitteen kokoaminen korkeamman tason käsitteeksi.
4) Abstraktien käsitteiden luominen konkreeteista esimerkeistä.
5) Konkreettien vastesääntöjen oppiminen kokemuksista, ja niiden tallentaminen.
6) Erityyppisten yleistettyjen vastesääntöjen luominen ja tallentaminen.
7) Nykyisen tilanteen esittäminen aistimuksiin viittaavilla käsitteillä.
8) Päätavoitteen kannalta sopivan vastesäännön valitseminen.
9) Valitun vastesäännön mukaisen vasteen suorittaminen. (Fritz 2007, Fundamental physiological mechanisms.)

Kykymme kuvailla ihmisen kognitiota tavalla tai toisella perustuu ajatukseen, että ihmiset ovat fysikaalisia ja biologisia olioita. (Newell 1990, 42.) Organismi on yleensä vuorovaikutuksessa ympäristönsä kanssa. Ympäristö on yleensä monimutkainen ja dynaaminen. Mieli on nimi kontrollisysteemille, joka on kehittynyt organismeihin ohjaamaan sen käyttäytymistä. Tämän

käyttäytymisen tavoitteena on säädellä organismin ja ympäristön välisiä suhteita, organismin hyödyksi. Aivot taas ovat elin tai arkkitehtuurin osa, jossa mielen prosessit tapahtuvat. Mieli tuottaa vastefunktioita (engl. response function). Organismi tuottaa vasteita ympäristön tilojen funktiona (Newell 1990, 43). Vastefunktio on olennaisesti sääntö, joka kertoo, mikä toiminto suoritetaan, kun ympäristö on havaintojen perusteella tietyssä tilassa. Vastefunktio voidaan ilmaista seuraavassa muodossa: ”Jos ympäristö on tilassa X, niin suoritetaan toiminto Y ”. Mieltä on mahdollista tarkastella yhtenä suurena vastefunktiona. Tämä voidaan tehdä tarkastelemalla mieltä organismin menneisyyden kaikkien ympäristöjen kokonaissumman funktiona tulevaisuuden vasteisiin. Kuitenkin yleensä organismien käyttäytymistä kuvataan niin että käyttäytyminen voi koostua monista eri vastefunktioista, joka edellyttää jonkinlaista organismin analysoimista osiin. (Newell 1990, 44.) Aivojen perusmekanismit ovat fysikaalisia prosesseja, joita voidaan kuvata laskennallisesti. Wolframin (2002, 713- 714) mukaan ennen nykyaikaisia tietokonesovelluksia saatettiin olettaa, että laskennan käsite soveltuu vain abstraktien elementtien systeemeihin. Nykyään laskennan elementteinä voi olla monentyyppistä dataa, numeroita, tekstiä, kuvaa jne. Tämä tuntuu osoittavan, että on mahdollista pitää mitä tahansa säännönmukaista prosessia laskentana, riippumatta siitä, minkälaisia elementtejä se sisältää. Niinpä on mahdollista pitää luonnon ja keinotekoisia prosesseja laskentana. Säännöt joita luonnon prosessit seuraavat ovat luonnonlakeja.

Kognitiivisen psykologian lähtökohtana oleva näkemys aivoista informaatiota prosessoivana laitteena voidaan jäljittää William Jamesiin (1842- 1910). Tästä lähtökohdasta Kenneth Craik (1943) spesifioi kolme tietoperustaisen agentin askelta: 1) Se kääntää ärsykkeen sisäiseksi esitykseksi, 2) kognitiiviset prosessit manipuloivat esitystä tuottaakseen uusia sisäisiä esityksiä, ja 3) nämä käännetään ulkoiseksi toiminnaksi. Myöhemmin tietokonemallinnuksen kehittyminen teki mahdolliseksi kognitiotieteen luomisen, kun George Millerin, Noam Chomskyn, Allen Newellin ja Herbert Simonin merkittävät artikkelit MIT:ssä 1956 osoittivat, että tietokonemallinnusta voidaan käyttää muistin, kielen ja loogisen ajattelun tutkimuksessa. Nykyään psykologien keskuudessa on yhteinen näkemys, että kognition teorian tulee olla tietokoneohjelman kaltainen, mikä tarkoittaa, että sen tulisi olla yksityiskohtainen informaationkäsittelymekanismi, jolla jokin kognitiivinen funktio voitaisiin toteuttaa. (Russel ja Norvig 2003, 13- 14.) Uskottavia vaihtoehtoisia selittämisen lähtökohtia ei mielestäni tällä

hetkellä ole. Toisaalta tutkimuksen menestys viittaa siihen, että vaihtoehtoja ei ainakaan vielä tarvitse etsiä. Lisäksi näihin oletuksiin perustuvat selitykset yhdistävät eri tieteenalojen tietoa ja edistävät tieteen yhtenäisyyttä.

Ihmisen aivojen toiminta neuraalisella tasolla

Selkärankaisten aivot voidaan jakaa kolmeen pääalueeseen: taka-aivoihin, keskiaivoihin ja etuaivoihin. Taka-aivot säätävät perustavimpia toimintoja, niitä jotka ovat elämälle välttämättömiä. Keskiaivot pitävät yllä valvetilaa ja karkeita, erillisiä käyttäytymisreaktioita. Etuaivot säätävät monimutkaisia käyttäytymisen ja mielen prosesseja. Kaikilla selkärankaisilla on kaikki kolme osaa ja jopa evolutiivisesti kehittyneimmät etuaivot perustuvat yleisrakenteeseen, joka ilmenee jokaisessa selkärankaislajissa. (LeDoux 2003, 45- 49.)

Ihmisen ajattelun perusta ovat aivosolut eli neuronit ja niiden väliset muutoskykyiset yhteydet, synapsit (LeDoux 2003, 352). Aivot koostuvat aivosoluista, eli neuroneista, jotka muista soluista poiketen viestivät suoraan keskenään hermosyiden ja synapsien avulla. Hermosyitä on kahta lajia, aksoneja ja dendriittejä. Dendriitti eli tuojahaarake on hermosolun syötekanava ja aksoni eli viejähaarake sen tulostekanava. Synapsi taas on neuronin viestejä lähettävä ja vastaanottava piste. Synapsit ovat pieniä neuronien välisiä rakoja. Synapsi on se kohta, missä edellisen neuronin aksoni kohtaa seuraavan neuronin dendriitin. Kun neuroni käynnistyy, sähköimpulssi kulkee sen hermosyytä pitkin ja aiheuttaa kemiallisten välittäjäaineiden erittymisen sen päätteestä. Välittäjäaine ajautuu synapsin poikki ja tarttuu vastaanottavan neuronin dendriittiin, jolloin rako ylittyy. Olennaisesti kaikki aivojen toiminnot perustuvat synaptiseen tiedonvälitykseen. (LeDoux 2003, 12, 52- 60.)

Neuroniryhmä, jonka synapsit kytkevät yhteen on piiri, ja sarja hierarkkisesti järjestyneitä piirejä muodostaa tietyn aivojärjestelmän (LeDoux 2003, 352). Tieto välittyy alueelta toiselle hierarkkisten piirien sarjaa pitkin. Kullakin hierarkian tasolla tiedonkäsittelyä kuitenkin säädellään muunlaisilla piireillä. Paikallisen piirin kytkennät muuttavat käsittelyä kullakin hierarkian tasolla ja myös määrittävät, miten helposti mikin tietyn alueen toiminta vaikuttaa

seuraavaan alueeseen. Järjestelmä on monimutkainen piiri, jolla on erityinen tehtävä, kuten näkö, kuulo, tai vaaran havaitseminen ja siihen reagoiminen. Kaikkia aivojärjestelmiä voidaan pitää sarjana hierarkkisesti järjestyneitä piirejä, jotka synaptiset kytkennät yhdistävät tietyn tehtävän hoitamiseksi. (LeDoux 2003, 61- 78.) Konvergenssivyöhykkeet taas yhdistävät eri aivojärjestelmistä tulevaa tietoa (LeDoux 2003, 352).

Aivojärjestelmät oppivat ja tallentavat samoja asioita koskevaa tietoa tietoa, koska ne kokevat samat tapahtumat, mutta prosessoivat näiden tapahtumien eri piirteitä. Jotta näkisi esimerkiksi omenan eikä pyöreähköä, punaista möykkyä, näköjärjestelmän eri alajärjestelmien prosessoimat ärsykkeen erilaiset piirteet on yhdistettävä. Synkronia eli samanaikaisuus koordinoi järjestelmien rinnakkaista muovautuvuutta: yksittäisten alueiden solut laukaisevat aktiopotentiaaleja synkronisesti. Esimerkiksi muoto ja väri yhtyvät havaituksi esineeksi, kun tiettyä muotoa ja tiettyä väriä prosessoivat solut toimivat samanaikaisesti. Hebbiläinen muovautuvuus tarkoittaa, että kun solun vahvat ja heikot syötöt toimivat samanaikaisesti, heikko reitti vahvistuu, koska se liittyy eli assosioituu vahvaan reittiin. Niinpä hebbiläinen muovautuvuus sitoo samanaikaisesti toimivat solut yhteen siten, että kun sama tai samankaltainen ärsyke havaitaan seuraavan kerran, samat solut ja kytkennät käynnistyvät uudelleen. (LeDoux 2003, 332- 343.)

Synkronian ja hebbiläisen muovautuvuuden avulla yhdessä vaikuttavat syöttötiedot sekä eri alueiden vuorovaikutukset luovat assosiaatioita. Myös säätöjärjestelmät koordinoivat rinnakkaista muovautuvuutta: ne käynnistyvät koko aivoissa merkittävien ärsykkeiden vaikutuksesta, joita ovat uudet, odottamattomat ärsykkeet tai ärsykkeet, jotka käynnistävät emotionaalisen kiihtymisen. Emotionaaliset tilat monopolisoivat aivojen voimavarat. Tämän seurauksena oppiminen koordinoituu järjestelmien yli erittäin täsmällisellä tavalla ja varmistaa, että oppiminen on olennaista senaikaiselle emotionaaliselle tilanteelle. Konvergenssivyöhykkeet yhdistävät rinnakkaista muovautuvuutta. Ne ottavat vastaan tietoa muilta aivojen alueilta ja yhdistävät toisten alueiden itsenäisesti prosessoimaa tietoa. Otsalohkon etuosan aivokuoressa on tärkeitä konvergenssivyöhykkeitä. Kun tieto on yhdistetty, se voi vaikuttaa syöttävien alueiden toimintaan. Kognitiiviset psykologit sanovat tiedon virtaamista alemmista ylempiin prosessointiasemiin alhaalta ylös -prosessoinniksi ja

virtausta ylemmistä alempiin asemiin ylhäältä alas -prosessoinniksi. Työmuistin kyky yhdistää eri järjestelmistä tulevaa tietoa ja säilyttää se väliaikaisesti mielen toimintoja varten on esimerkki alhaalta ylös -prosessoinnista. Työmuistin kyky käyttää prosessointinsa tulosta säätääkseen sitä, mitä tarkkailemme, on taas esimerkki ylhäältä alas -prosessoinnista. (LeDoux 2003, 332- 356.)

Työmuisti on mielen työn perusta ja se liittyy kaikkiin ajattelun ja ongelmanratkaisun piirteisiin. Työmuistissa oleva tieto on se, jota tällä hetkellä ajattelee tai josta on tietoinen. Koska työmuistin tieto on väliaikaista, sen sisältöä on päivitettävä jatkuvasti. Työmuistin sisältö ei riipu vain nykyisestä tilanteesta, vaan myös siitä, mitä tiedämme ja millaisia kokemuksia meillä on ollut aikaisemmin, ts. pitkäkestoisesta muistista. Työmuistin yleisjärjestelmä koostuu työtilasta ja joukosta mielen toimintoja, joita sanotaan eksekutiivisiksi toiminnoiksi (toiminnan ohjaukseksi) ja ne kohdistetaan työmuistissa olevaan tietoon. Vaikka työtilassa on kunakin hetkenä vain rajallinen määrä tietoa, se voi säilyttää ja liittää toisiinsa eri erikoistuneista järjestelmistä tulevaa erilaista tietoa. Kyky yhdistää eri järjestelmien tietoa johtaa esineiden ja tapahtumien abstrakteihin esityksiin. Se on kehittynyt erityisen hyvin ihmisellä ja todennäköisesti se vaikuttaa ihmisen ajattelun ainutlaatuisuuteen. Toiminnan ohjaus ohjaa erikoistuneet järjestelmät tarttumaan tiettyihin ärsykkeisiin ja jättämään muut huomiotta sen perusteella, mitä työmuisti kulloinkin työstää. Monimutkaisissa, monenlaisia mielen toimintoja edellyttävissä tehtävissä toiminnan ohjaus suunnittelee mielen askelten jonon, ja aikatauluttaa tiettyjen toimintojen käynnistymisen siirtäen tarpeen mukaan tarkkaavaisuuden toiminnosta toiseen. (LeDoux 2003, 197- 223.)

Aivokuoren toiminta

Evoluution kulussa luonto havaitsi, että lisäämällä eläimelle muistijärjestelmän ja syöttämällä siihen aistivirran eläin pystyisi muistamaan menneitä tapahtumia. Kun eläin havaitsi olevansa samassa tai samanlaisessa tilanteessa kuin aiemmin, muistot johtivat ennusteeseen siitä, mitä todennäköisesti tapahtuu seuraavaksi. Älykkyys ja ymmärrys alkoivat muistijärjestelmästä, joka loi ennusteita aistivirrasta. Jonkun asian tietäminen tarkoittaa, että siitä pystyy tekemään

ennusteita. Aivokuori kasvoi suuremmaksi ja alkoi pystyä tallentamaan pidemmälle kehittyneitä muistoja, jolloin se pystyi tekemään ennusteita aiempaa monimutkaisempien suhteiden perusteella. Ihmisen aivokuori on erityisen suuri, ja sen vuoksi sillä on valtava muistikapasiteetti. Se ennustaa jatkuvasti mitä näemme, kuulemme ja tunnemme, useimmiten tiedostamattamme. (Hawkins 2005, 111.)

Aivokuoren tehtävä on selvittää, kuinka syötteet liittyvät toisiinsa, muistaa niiden välisten suhteiden sekvenssi, ja käyttää tätä muistoa ennustaakseen syötteiden käyttäytymistä tulevaisuudessa. Jokainen sen alue muodostaa esitystapainvariantteja hierarkiassa alapuolellaan sijaitsevien syötealueiden perusteella. Laajat kokonaisuudet ovat tallentuneina hierarkian huipulle ja pienemmät lähemmäs hierarkian alaosaa. (Hawkins 2005, 129, 130.) Kukin aivokuoren alue oppii sekvenssejä, kehittelee sen jälkeen tuntemilleen sekvensseille nimen ja välittää sen jälkeen nämä nimet aivokuoren hierarkiassa seuraavaksi korkeammille alueille. (Hawkins 2005, 133- 134.) Ihminen pystyy pohtimaan maailmaa, liikkumaan siinä ja tekemään ennusteita tulevaisuudesta, koska aivokuori on luonut mallin maailmasta. Kaikki kohteet muodostuvat yhdessä esiintyvistä osatekijöistä; tämä on kohteen määritelmä. Annamme jollekin asialle nimen, koska tietyt piirteet tai ominaisuudet esiintyvät jatkuvasti yhdessä. Kasvoja sanotaan kasvoiksi, koska niissä on yleensä kaksi silmää, nenä ja suu ja ne esiintyvät yhdessä. Kohteet koostuvat usein pienemmistä kohteista, jotka ovat suurempien esineiden ja kohteiden osia. Aivokuoren oppimisalgoritmi löytää ympäristön hierarkkiset rakenteet ja vangitsee ne. Kun rakennetta ei ole, ihminen hämmentyy. (Hawkins 2005, 131-132.) Todennäköisyys lukuisien syötemallien tapahtumiselle samassa suhteessa yhä uudelleen on erittäin pieni. Siksi ennustaminen on luotettava keino saada selville, että maailman eri tapahtumat ovat sidoksissa toisiinsa. (Hawkins 2005, 133- 134.)

Jokaisella aivokuoren alueella on nimi jokaiselle sen tuntemalle sekvenssille. Tämä nimi on joukko soluja, joiden kollektiivinen impulssien lähetys edustaa kohteiden joukkoa sekvenssissä. Solut pysyvät aktiivisina niin kauan kuin tuota sekvenssiä esitetään, ja nimi välitetään hierarkiassa ylöspäin seuraavana olevalle alueelle. Kokoamalla ennustettavat sekvenssit nimetyiksi kohteiksi jokaisella hierarkian tasolla saavutetaan yhä vakaampi tila mitä ylemmäs edetään. Tällä tavalla syntyy ns. esitystapainvariantteja. Vastakkainen vaikutus

ilmenee, kun malli siirtyy hierarkiassa alaspäin. Vakiona pysyvät mallit jakautuvat sekvensseiksi. Ensiksi alue luokittelee syötteensä yhdeksi mahdollisuudeksi rajallisesta määrästä mahdollisuuksia, ja sitten se etsii sekvenssejä. (Hawkins 2005, 135- 136, 139, 141, 142.)

Käsitteet

Käsitteitä tarvitaan ja käytetään ns. kognitiivisen ekonomian saavuttamiseksi, kokemusten tallentamiseen, käyttämiseen ja järjestämiseen: jakamalla ympäristön kohteet luokkiin vähennetään opittavan, havaittavan, muistettavan ja tunnistettavan informaation määrää (Collins & Quillian 1969). Käsitteiden avulla voidaan tehokkaasti esittää tietoa, kommunikoida ja tehdä ennusteita. Esimerkiksi jos luokittelemme eläimen kissaksi (eikä leijonaksi), voimme ennustaa että se ei luultavasti vahingoita meitä. (Eysenck 2005, 294- 295.) Kategoriat ovat kohteiden luokkia ja käsitteet ovat mielen esityksiä näistä kohteiden luokista (Eysenck 2005, 294).

Filosofit käyttävät käsitteen käsitettä monin eri tavoin, osaksi koska sitä käytetään monissa eri projekteissa ja eri tavoitteisiin. Sekaannusta lisää se, että väittely käsitteistä heijastaa syvästi vastakkaisia lähestymistapoja mielen, kielen ja filosofian itsensä tutkimukseen. (Margolis ja Stephen 2007.) Näkemyksiä käsitteiden ontologisesta statuksesta on ainakin kaksi:

1) Käsitteet ovat kognitiivisille agenteille ominaisia kykyjä (Brandom 1994, Millikan 2000). Esimerkiksi kissan käsite olisi kyky erottaa kissa ei-kissoista ja tehdä joitain päätelmiä niistä. Yksi syy siihen, miksi jotkut filosofit ovat omaksuneet tämän käsityksen, on epäily mentaalisten representaatioiden olemassaoloa ja hyödyllisyyttä kohtaan, koska on mm. väitetty, että tämä selitys sisältää samoja ongelmia kuin mihin yrittää vastata: tietoa ensimmäisestä kielestä ei pitäisi selittää mallilla tiedosta toisesta kielestä (Dummett 1993, 98.) Kriitikoiden mukaan se ei pysty selittämään kykyä tuottaa uusia ajatuksia eikä muitakaan mielen prosesseja, vaan kyvyt on selitettävä esimerkiksi mentaalisten representaatioiden avulla

(Chomsky 1980). (Margolis ja Stephen 2007.) Tällainen lähestymistapa ei tunnu lisäävän käsitteitä koskevaa ymmärrystä.

2) Käsitteet ovat mielen esityksiä. Lähtökohtana on representationaalinen mielen teoria (engl. representational theory of mind, RTM), jonka mukaan ajattelu tapahtuu sisäisessä esitys-systeemissä. Varhaiset RTM:n kannattajat (esim. Locke (1690/1975) ja Hume (1739/1978) nimittivät yksinkertaisimpia representaatioita ideoiksi ja pitivät niitä mentaalisina kuvina. Sen sijaan uudemmat RTM:n versiot olettavat, että ajattelu perustuu ns. ajattelun kieleen, jolla on oma puhutusta ja kirjoitetusta kielestä ainakin jonkin verran eroava syntaksi ja semantiikka (Fodor 1975). Käsitteet mielen esityksinä on nykyään kognitiotieteen lähtökohta (Pinker 1994), ja se on suosittu mielenfilosofiassa, mm. koska sen perusteella on helppoa ymmärtää miten ihmisillä voi olla rajaton määrä ajatuksia (Fodor 1987). Joidenkin mielestä RTM on liian läheisessä suhteessa vanhentuneeseen arkijärkipsykologiaan (engl. commonsense psychology) (Churchland 1981), tai että laskennallinen mallinnus tarjoaa vaihtoehtoja varsinkin RTM:n ajattelun kieli -versioon (ks. Port ja Gelder 1995 dynaamisista systeemeistä, Elman et al. 1996 konnektionismista). (Margolis ja Stephen 2007.)

Mentaalinen representaatio on ns. laskennallisen mielen teorian (engl. computational theory of mind, CTM) peruskäsite. (Pitt 2007.) Nykyään mielenfilosofit tyypillisesti olettavat tai toivovat, että mieltä koskevat tosiasiat voidaan selittää luonnontieteiden avulla. Tämä oletus on myös kognitiotieteellä, joka yrittää tuottaa selityksiä aivoja ja keskushermostoa koskevan tiedon avulla. Kognitiotieteen eri osa-alueet (mm. kognitiivinen ja laskennallinen psykologia ja kognitiivinen ja laskennallinen neurotiede) olettavat monia erilaisia struktuureja ja prosesseja, joista monet eivät ole suoraan mentaalisten tilojen ja prosessien implikoimaa. Kuitenkin yleensä oletetaan, että mielen tilat ja prosessit voidaan selittää mentaalisten representaatioiden termistöllä. (Pitt 2007.) Kognitiotieteessä filosofisesti relevantit debatit ovat keskittyneet aivojen ja keskushermoston kognitiiviseen arkkitehtuuriin ja tieteellisen ja arkijärjen mentaalisuuden selitysten yhteensopivuuteen. (Pitt 2007.)

Johtavan representationaalisen mielen teorian (RTM) version, laskennallisen mielen teorian (CTM) mukaan aivot on jonkinlainen tietokone ja mielen prosessit komputaatioita

(laskennallisia prosesseja). Kognitiiviset tilat koostuvat laskennallisista suhteista erilaisiin mentaalisiin representaatioihin ja kognitiiviset prosessit ovat tällaisten tilojen sekvenssejä. CTM:n kannattajat jakautuvat nykyään klassisten ja konnektionististen arkkitehtuurien kannattajiin. Klassisistien (mm. Turing 1950, Fodor ja Pylyshyn 1988, Newell ja Simon 1976) mukaan mentaaliset representaatiot ovat symbolisia struktuureja, joilla tyypillisesti on semanttinen sisältö, ja mielen prosessit ovat niiden manipulointia. Konnektionistien (mm. Smolensky 1988, Rumelhart 1989) mukaan mentaaliset representaatiot toteutuvat aktivaatiosekvensseinä yksinkertaisten prosessorien (solmujen, engl. node) verkossa ja mielen prosessit koostuvat näiden aktivaatiosekvenssien leviämisestä verkossa. (Pitt 2007.) Konnektionistien kanta on ihmisten kognition kohdalla järkevämpi, koska ihmisen aivojen arkkitehtuuri on selvästi lähempänä tätä käsitystä. Aivojen toiminnasta tarjoavat uskottavan alustavan selityksen P. Churchland (1995), ja Hawkins (2005). Kognitiiviset prosessit voidaan toteuttaa monella eri tavalla.

Tärkeimpiä teorioita käsitteiden rakenteesta ovat klassinen teoria, prototyyppiteoria ja teoriateoria:
1) Klassisen teorian mukaan käsite koostuu yksinkertaisemmista käsitteistä, jotka ilmaisevat välttämättömiä ja riittäviä ehtoja kohteen käsitteen alaan kuulumiselle. Perinteinen esimerkki on käsite POIKAMIES, joka koostuu leksikaalisista käsitteistä NAIMATON ja MIES. Luokittelu voidaan ymmärtää prosessiksi, jossa tarkastetaan, täyttääkö kohde käsitteen määrittelevien osien ehdot. Yksi tämän teorian ongelmia on, että sen olettamaa käsitteiden määritelmällistä rakennetta on ollut vaikea analysoida esiin: Siitä asti kun Gettier (1963) kyseenalaisti tiedon perinteisen määritelmän (TIETO = PERUSTELTU TOSI USKOMUS), ja tyydyttävä vaihtoehtoista määritelmää ei ole löydetty, jotkut ovat ajatelleet, että käsitteillä ei ehkä sittenkään ole määritelmällistä erillisiin yksinkertaisempiin käsitteisiin analysoitavaa rakennetta. (Margolis ja Stephen 2007.) Lisää klassisen teorian kritiikkiä ks. (Murphy 2002).

2) Prototyyppiteorian mukaan käsitteellä on probabilistinen rakenne: Kohde kuuluu käsitteen alaan, jos se on riittävän samankaltainen prototyypin kanssa. Prototyyppiteoriassa luokittelu ymmärretään prosessiksi, jossa vertaillaan kohteen samankaltaisuutta sen prototyyppiin ja toisiin kohteisiin. (Margolis ja Stephen 2007.) Vaikuttaa epätodennäköiseltä, että yksi

prototyyppi voisi kattaa kaikki eri mahdollisuudet muuta kuin harvojen hyvin kapeasti määriteltävien käsitteiden kohdalla (Murphy 2002, 42). Tämän kanssa samantyyppinen on esimerkkiteoria (engl. example theory), jonka mukaan esimerkiksi koiran käsite on yksilön kaikkien muistamiensa koirien joukko. Luokittelu perustuu siihen, muistuttaako uusi kohde näitä käsitteeseen jo kuuluvia esimerkkejä. Sen ongelmia ovat mm. kuinka mitata samankaltaisuutta esimerkin kanssa, ja hierarkkisten rakenteiden muodostumisen käsitteiden välillä selittäminen. Lisäksi se on laskennallisesti epäilyttävä, koska kohteen vertailu esimerkkien kanssa vie enemmän laskentaresursseja kuin mitä esimerkiksi ihmiset näyttävät luokitteluun käyttävän (Murphy 2002, 49- 58, 484- 486, 490- 491.)

3) Teoriateorian (engl. theory theory) mukaan käsitteet ovat samalla tavalla suhteessa toisiinsa kuin tieteellisen teorian termit, ja luokittelu on prosessi, joka muistuttaa tieteellistä teoretisointia (Carey 1985, Gopnik ja Meltzoff 1997). Teoriateoria selittää erityisen hyvin vaikeat (laajat tai epäselvät, paljon kohteita tai niiden variaatioita sallivat käsitteet) luokittelutapaukset toisin kuin prototyyppiteoria. Sen mukaan käsitteiden kehittyminen lapsuudessa muistuttaa teorian muutosta tieteessä. Sitä on kritisoitu mm. siitä, että koska sen mukaan käsitteen sisältö määrittyy sen roolista teoriassa ja koska ihmisillä ei yleensä ole täysin samat teoriat, niin eri ihmisten käsitteitä ei välttämättä voida vertailla (Fodor ja Lepore 1992). (Margolis ja Stephen 2007.) (Muphy käyttää tästä nimeä the knowledge approach (2002, 60- 64, lisää teorian kritiikkiä ks. s. 487.))

Murphyn mukaan käsitteiden teorian tulee olla yhdistelmä vaihtoehtoja (2) ja (3): Sen tulee perustua ensisijaisesti prototyyppiteoriaan eli sen pitää olla kuvaus koko käsitteestä, sen tyypillisistä ominaisuuksista niiden tärkeyden mukaan painotettuina (2). Kuitenkin tämän kuvauksen tulee olla osa suurempaa tiedonesittämis-skeemaa, jossa sillä voi olla hierarkkinen paikka ja sopiva teoreettinen viitekehys (3). Ilmeisesti tämä yhdistää molempien vahvoja puolia ja kompensoi heikkouksia. Kuitenkin lisää tutkimusta tarvitaan (Murphy 2002, 488-498), ja monissa eri kannoissa voi olla oikeaan osunutta (Margolis ja Stephen 2007).

Mielestäni Fritzin, Hawkinsin ja Churchlandin näkemysten perusteella voidaan esittää uskottava teoria käsitteistä. Käsite on ajattelun peruselementti, informaation tallenne, joka voi

olla tallennettuna esimerkiksi hermosoluihin (mm. ihmisellä) tai elektroniikkaan (joillakin roboteilla). Älykäs systeemi luo käsitteitä prosessoimalla sen sensorien ympäristöstä vastaanottamia kommunikaatioita. (Fritz 2007, Elementary concepts.) Aistinelimet ovat ensimmäinen askel tämän informaation hankkimisen prosessissa. Kun aivot saavat tämän aisti-informaation (hermoimpulsseina) se prosessoi sen niin, että tilalliset ja ajalliset suhteet joidenkin impulssien välillä huomioidaan. Jos nämä suhteet ovat samankaltaisia ennen saadun informaation kanssa, se tulkitaan käsitteellä, joka älykkäällä systeemillä ennestään on. Muussa tapauksessa se luo uuden käsitteen ja tulkitsee sen tällä uudella käsitteellä. Suoraan aisti-informaatioon perustuvia käsitteitä sanotaan elementaarikäsitteiksi (engl. elementary concepts). (Fritz 2007, Elementary concepts.) Myöhemmin älykkään systeemin aivot käyttävät näitä elementaarikäsitteitä rakentaakseen niistä monimutkaisempia yhdistelmä- eli komposiittikäsitteitä (engl. built-up concepts). Mielen prosessit käyttävät elementaarikäsitteitä ja komposiittikäsitteitä. Muistissa olevat käsitteet ovat usein yhteydessä toisiinsa, ja ne voivat muodostaa verkon tai verkoston. (Fritz 2007, Built-up concepts.)

Käsite on informaatiorakenne, joka liittää tiettyihin ominaisuuksiin tietyt muut ominaisuudet. Käsitteet sisältävät kohteitaan koskevaa, agentille hyödyllistä tietoa. Jos agentti on oppinut, miten tietty kohde jossakin suhteessa käyttäytyy, ja jos jotkin toiset kohteet muistuttavat jollakin tavalla tätä kohdetta, agentti voi koettaa olettaa, että näiden toisten kohteiden käyttäytyminen on jollakin tavalla samanlaista kuin ensimmäisen. Jos tämä oletus eri kokeilujen perusteella pitää paikkansa, agentti voi tästä eteenpäin olettaa, että mikä tahansa, mikä täyttää ainakin jossain määrin kohteen tuntomerkit, ainakin jossakin määrin käyttäytyy tällä tavalla. Tällainen tietorakenne, joka liittää tietyt ominaisuudet tiettyihin toisiin ominaisuuksiin, on käsite. Jos käsite ainakin yleensä pitää paikkansa (toimii käytännössä), se voi säästää agentin laskennallisia resursseja, koska tällöin joitakin kohteita ei tarvitse käsitellä uusina, vaan voidaan käyttää hyväksi jo ennestään hankittua tietoa. Käsite voidaan yksinkertaisimmillaan ilmaista seuraavana informaatiorakenteena: "Jos kohde x täyttää käsitteen Y tuntomerkit (käsitteen määrittelevät ominaisuudet) paremmin kuin muiden käsitteiden määrittelevät ominaisuudet, esitä kohde käsitteellä Y." Tällöin kohteen oletetaan käyttäytyvän niin kuin muidenkin käsitteen alaan kuuluvien kohteiden (eli instanssien). Tällainen rakenne voidaan toteuttaa monella eri tavalla.

Käsite ei ole sama kuin sen kohde, vaan materiaalinen rakenne, kokoelma informaatiota aivoissa. Omenan käsite on eri asia kuin pöydällä oleva omena. Pöydällä on jotain, josta tulee näkyviä elektromagneettisia aaltoja. (Fritz 2007, The concept is NOT the "thing"!) Silmän retinan ensimmäinen kerros konvertoi osan tästä säteilystä ("näkyvä valo") hermoimpulsseiksi. Toisen ja siitä seuraavien kerrosten hermosolut kokoavat ja prosessoivat niitä, esimerkiksi erottavat rajoja ja kulmia jne... Näin aistien saama kommunikaatio muuttuu informaatioksi. Aivot yhdistelevät eri hermoimpulsseja, jotka voivat sisältää informaatiota esimerkiksi väreistä ja muodoista ja joistakin muista aspekteista, ja muodostavat niistä koherentin kuvan. Aivot antavat tälle informaatiokokonaisuudella nimen. Tämä nimi on käsite ja informaatio, johon se viittaa, on sen sisältö. Aivot myös muokkaavat, yhdistelevät ja tiivistävät käsitteen sisällön informaatiota. Esimerkiksi ne voivat lisätä "siemenet" omenan käsitteeseen, kun älykäs systeemi halkaisee omenan ja huomaa ne ensi kerran. Lisäksi siihen voidaan lisätä informaatio, että älykäs systeemi voi kommunikoida siitä toisten kanssa suomen kielellä käyttämällä symboleita kuten kirjaimet o-m-e-n-a. (Fritz 2007, The process.) Käsitteillä on yhteydet (linkki) toisiin, joiden osia ne itse ovat. Esimerkiksi käsitteellä "rengas" on linkki käsitteeseen "auto". Käsite "auto" puolestaan on osa käsitettä "ajoneuvo", ja sillä vastaava linkki. Käsitteillä on myös linkit toiseen suuntaan, omiin osiinsa. Osilla voi olla omia osia. (Fritz 2007, Total and parts.) Elementaarikäsitteet ovat kaikkein konkreettisinta informaatiota, mitä aivoilla on. Osiin menevien linkkien sijaan niillä on yksityiskohtaista informaatiota aistimuksista tai elementaaritoiminnoista. (Fritz 2007, Abstracts and concretes.)

Käsitteen merkki on informaatiorakenne, jonka tietty tulkinta liittää tiettyyn käsitteeseen. Merkkejä käytetään, koska niitä on usein helpompi prosessoida ja kommunikoida kuin käsitteitä. Kun merkki kommunikoidaan esim. äänen (puhe, musiikki) tai valon (kuva, kirjoitus) välityksellä (koodaamalla ja dekoodaamalla se eri muotoihin), systeemille, jolla on oikea merkin tulkinta, merkki viittaa suunnilleen samanlaisiin käsitteisiin tai vastesääntöihin tämän toisen systeemin tietokannassa. Tulkinta määrittelee, mitkä merkit viittaavat mihinkin käsitteisiin. Käsitteet taas viittaavat sisäisesti tai ulkoisesti havaittaviin kohteisiin. Jos johonkin esim. konventioon perustuen niiden aivoissa samankaltaisilla käsitteillä on sama esimerkiksi äänien tai kirjoituksen avulla kommunikoitava symboli, toinen älykäs systeemi

voi aktivoida toisessa tämän käsitteen kommunikoimalla toiselle sen symbolin. Käsitteellä on siis kolme aspektia, jotka täytyy erottaa toisistaan:
1) Käsite, jota käytetään mielen prosesseissa.
2) Symbolit joita käytetään käsitteen kommunikoimiseen.
3) Käsitteen suurelta osin tuntematon kohde ympäristössä (Fritz 2007, Extensive summary).

Miten käsitteet toteutuvat hermoverkoissa? Kun tietty neuronien joukko aktivoituu tietyllä tavalla, sen osiin yhteydessä oleva hierarkiassa ylempänä oleva neuroni tai neuronien joukko aktivoituu. Tämä ylempi joukko on käsitteen symboli, ja alempi sen sisältö. Symbolia voidaan käyttää helpommin mielen prosesseissa, koska se edustaa suurta määrää alemman tason informaatiota. Hawkinsin termein ihmisellä käsite on esitystapainvariantti, pysyvä mallisekvenssi (neuronien tasolla: hermosolujen ja niiden välisten yhteyksien joukko), joka aktivoituu aina riittävän stimulaation seurauksena.

Jos mallit liittyvät toisiinsa siten, että alue voi oppia ennustamaan, mikä malli esiintyy seuraavaksi, niin aivokuoren alue muodostaa pysyvän esitystavan tai muistikuvan tuosta sekvenssistä. Aivot käsittelevät abstraktisia ja konkreettisia kohteita samalla tavalla. Ne molemmat ovat mallisekvenssejä, jotka esiintyvät yhdessä ajan kuluessa ja ennustettavalla tavalla. Todennäköisyys lukuisien syötemallien tapahtumiselle samassa suhteessa yhä uudelleen on erittäin pieni. Siksi ennustaminen on luotettava keino saada selville, että maailman eri tapahtumat ovat sidoksissa toisiinsa. Kukin aivokuoren alue oppii sekvenssejä, kehittelee sen jälkeen tuntemilleen sekvensseille nimen ja välittää sen jälkeen nämä nimet aivokuoren hierarkiassa seuraavaksi korkeammille alueille. (Hawkins 2005, 133- 134.)

Jokaisella aivokuoren alueella on repertuaari sen tuntemia sekvenssejä. Jokaisella aivokuoren alueella on nimi jokaiselle sen tuntemalle sekvenssille. Tämä nimi on joukko soluja, joiden kollektiivinen impulssien lähetys edustaa kohteiden joukkoa sekvenssissä. Nämä solut pysyvät aktiivisina niin kauan kuin tuota sekvenssiä esitetään, ja tämä nimi on juuri se, mikä välitetään hierarkiassa ylöspäin seuraavana olevalle alueelle. Niin kauan kuin syötemallit ovat osa ennustettavaa sekvenssiä, alue esittää vakiona pysyvän nimen seuraavaksi korkeammalle alueelle. Kokoamalla ennustettavat sekvenssit nimetyiksi kohteiksi jokaisella hierarkian

alueella (tasolla) saavutetaan yhä vakaampi tila mitä ylemmäs edetään. Tällä tavalla syntyy esitystapainvariantteja. Vastakkainen vaikutus ilmenee, kun malli siirtyy hierarkiassa alaspäin. Vakiona pysyvät mallit jakautuvat sekvensseiksi. Ensiksi alue luokittelee syötteensä yhdeksi mahdollisuudeksi rajallisesta määrästä mahdollisuuksia, ja sitten se etsii sekvenssejä. Aivokuoren alueilla alhaalta ylös -luokitukset ja ylhäältä alas -sekvenssit ovat jatkuvassa vuorovaikutuksessa ja muuttuvat ihmisen koko elämän ajan. Kaikki aivokuoren alueet ovat plastisia, joten niitä voidaan muokata kokemusten kautta. Ihminen muistaa maailman muodostamalla uusia luokituksia ja sekvenssejä. Esitystapainvariantti voi millä tahansa aivokuoren alueella muuttua yksityiskohtaiseksi ennusteeksi siitä, miten se ilmenee ihmisen aisteissa, jos mallia siirretään hierarkiassa alaspäin. (Hawkins 2005, 135- 136, 139, 141, 142.)

Vastesäännöt

Vastesäännöt ovat älykkään systeemin ehkä tärkein osa. Ne ovat tietoa, johon älykäs systeemi on liittänyt tietyn toiminnon. Olemassaolonsa aikana älykäs systeemi saa kokemuksia. Kokemus on jotain, mitä älykkäälle systeemille on tapahtunut sen olemassaolon aikana. Kokemukseen sisältyvät 1) ilmaantunut tilanne, 2) sen tässä tilanteessa suorittama toiminto, ja 3) toiminnon tulokset, johon sisältyy ympäristön uusi tila ja arvio tästä tilasta. Jotta sillä olisi mahdollisuus toimia paremmin tulevaisuudessa, se tallentaa kokemuksen eli tilanteen, vasteen ja tuloksen kokonaisuudeksi, jota sanotaan vastesäännöksi. (Fritz 2007, Response rules.) Se vertailee syötteeseen mahdollisia eri toimintoja ja niiden tuloksia ja yrittää parantaa tulostaan valitsemalla paremman toiminnon nyt ja tulevaisuudessa kuin menneisyydessä. Vastesääntö voidaan ilmaista seuraavasti: "Tilanteessa A suoritetaan toiminto B, jolloin seuraa tila C, jonka haluttavuus on D."

Luonnollisilla älykkäillä systeemeillä toimintojen tulokset ilmaistaan usein emootioina, jotka tunnetaan toiminnan tuloksena. Emootiot ovat tulosta aivojen ja monien erilaisten eri puolilla kehoa olevien reseptorien kommunikaatioista. Reseptorit voivat viestittää aivoille esimerkiksi kivusta, nälästä, väsymyksestä tai näiden puuttumisesta. Keinotekoisilla älykkäillä systeemeillä aivot voivat saada emootioita (ainakin) kahdella tavalla. Ne voivat olla signaaleja,

joita ihminen lähettää sille tarkkaillessaan sen vasteita, esimerkiksi opetustilanteessa, tai ne voivat olla signaaleja, jotka on laskettu erilaisten kehon variaabelien (muuttujien) arvojen yhteenvedosta. Kehon variaabeleja voivat olla esimerkiksi käytettävissä olevan energian määrä, objektiin törmääminen, tai huomio toimintaviasta. (Fritz 2007, Response rules.) Keinotekoisilla älykkäillä systeemeillä vastesäännöt koostuvat pääasiassa tilanteen esityksestä ja vasteesta. Usein vastesääntöihin lisätään muutakin tietoa. Esimerkiksi voidaan lisätä aika milloin vastesääntöä viimeksi käytettiin, vastesäännön suorittamisen tuloksena saatu mielihyvä (nautinto) tai tuska, tai käsitteiden positiivinen tai negatiivinen paino tilanteen esityksessä, jne. (Fritz 2007, Response rules, Pleasure.)

Vaistot ovat vastesääntöjä, joita älykkäällä systeemillä on olemassaolonsa alusta. Ihmislapsen vaistoihin kuuluvat esimerkiksi kiinni pitäminen ja imeminen, jotka ovat olleet välttämättömiä selviytymiseen. Keinotekoisissa älykkäissä systeemeissä esimerkiksi uteliaisuutta ja joitakin muita yleisiä vastesääntöjä pidetään joskus hyödyllisinä ja ne liitetään sen alkuperäiseen ohjelmaan. (Fritz 2007, Response rules, Overview of the intelligent system.) Melkein mikä tahansa älykäs systeemi luultavasti hyötyy ainakin seuraavista vaistoista:

- Itsesäilytysvietti auttaa välttämään vaurioita.
- Uteliaisuus edistää oppimista.
- Toisten älykkäiden systeemien toimintojen kopiointi voi säästää vaivaa kun kaikkea ei tarvitse itse kokeilemalla oppia. (Fritz 2007, Overview of the intelligent system.)

Koska on mahdotonta ennustaa mihin kaikkiin eri tilanteisiin älykäs systeemi saattaa joutua, kaikkia sen vasteita ei kannata yrittää ohjelmoida etukäteen. On parempi, että se suureksi osaksi oppii käyttäytymään menneisyyden kokemustensa perusteella.

Mielen menetelmät (engl. mental methods) ovat vastesääntöjä, joiden toiminta tapahtuu aivoissa. Niillä ohjataan, hallitaan tai järjestellään (manage) informaatiota, esimerkiksi kun älykäs systeemi kuvittelee tai suunnittelee. (Fritz 2007, Mental methods.) Käsitteitä ja vastesääntöjä voi olla hyödyllistä yleistää. Yleistyksessä yksi käsite tai vastesääntö korvaa useampia. Yleistys eli generalisaatio on yksi mielen menetelmä. Esimerkiksi unen aikana älykäs systeemi voi korvata vastesäännön toiminto-osan elementaarikäsitteet komposiittikäsitteellä. Kun älykäs systeemi suorittaa vastesäännön toimintoa, se suorittaa

komposiittikäsitteen jokaisen osan. Mielen menetelmissä komposiittikäsitteitä on helpompaa käsitellä, koska vastesäännön toiminto-osassa on vain yksi käsite useamman sijaan. Myös vastesäännön tilanne-osan konkreetti käsite voidaan joskus korvata abstraktilla käsitteellä, jonka esimerkki (instanssi) konkreetti käsite on. Tällöin vastesääntöä voidaan käyttää moniin tilanteisiin. Esimerkiksi koko tilanteen korvaavaa abstraktia käsitettä voidaan käyttää kaikissa tilanteissa, jotka ovat tämän abstraktin käsitteen esimerkkejä. (Fritz 2007, Mental methods.)

Voidaan havaita, että älykäs systeemi toistaa usein samoja vastesääntöjen ketjuja tai sekvenssejä samankaltaisissa tilanteissa. Jos joidenkin toimintojen sekvenssi havaitaan aina hyödyllisiksi tietynlaisessa tilanteessa, älykäs systeemi tallentaa sen ja vastaa samankaltaiseen tilanteeseen tällä tietyllä toimintojen sekvenssillä, ja näin siitä tulee tapa. (Fritz 2007, Mental methods and chains of response rules: habits.)

Ongelmanratkaisu ja haku

Ongelmanratkaisua tarvitaan, kun agentin täytyy toimia tilanteessa, johon sillä ei ole valmiina vastesääntöä. Tavoitteen määrittely on ongelmanratkaisun ensimmäinen askel. Se perustuu tietoon nykytilanteesta ja agentin menestyksen määritelmästä (engl. performance measure). Tavoite on haluttu maailman tila, joka voidaan ilmaista mm. ehtojen konjunktiona. Joskus moni eri maailman tila voi täyttää halutun maailman tilan ehdot. Agentin tehtävä on selvittää, minkä toimintojen sekvenssin suorittamalla se saavuttaa tavoitteensa. (Russell ja Norvig 2003, 60.) Ongelman ratkaisu on se toimintojen sekvenssi, jonka suorittamalla agentti saavuttaa tavoitteensa. Agentin täytyy päättää, millaisia toimintoja ja tiloja se voi tässä tehtävässä ottaa huomioon. Ongelman muotoilu (engl. problem formulation) on päätöksentekoprosessi, jossa päätetään, mitä toimintoja ja tiloja ongelmanratkaisussa otetaan huomioon. (Russell ja Norvig 2003, 60.)

- Tila-avaruus on lähtötilasta seuraajafunktiolla saavutettavien tilojen joukko. Ongelma voidaan määritellä formaalisti neljällä komponentilla tila-avaruudessa: Lähtötila,

toimintojen joukko, tavoitetesti-funktio ja reitin kustannus -funktio. Tila-avaruudessa reitti lähtötilasta tavoitetilaan on ongelman ratkaisu.

- Lähtötila (engl. initial state) on tila, josta agentti aloittaa.
- Toimintojen joukko on agentille mahdollisten toimintojen kuvaus. Yleisin formulaatio käyttää seuraajafunktiota (engl. successor function). Tietylle tilalle *x*, SUCCESSOR-FN(*x*) palauttaa joukon [toiminto, seuraaja] järjestettyjä pareja, missä jokainen toiminto on yksi mahdollisista toiminnoista tilassa *x*, ja jokainen seuraaja on tila joka voidaan saavuttaa tilasta *x* suorittamalla toiminto. Lähtötila ja seuraajafunktio yhdessä määrittelevät ongelman tila-avaruuden (engl. state space) implisiittisesti. Ongelman tila-avaruus on kaikkien lähtötilasta tavoitettavissa olevien tilojen joukko. Reitti (engl. path) tila-avaruudessa on tilojen sekvenssi jota yhdistää toimintojen sekvenssi.
- Tavoitetesti-funktio (engl. goal test) ratkaisee, onko tietty tila tavoitetila. Joskus mahdollisten tavoitetilojen joukko voidaan esittää eksplisiittisesti, jolloin tavoitetesti yksinkertaisesti selvittää onko annettu tila yksi niistä. Joskus taas tavoite on määritelty abstraktien ominaisuuksien avulla, eikä eksplisiittisesti esitettävien tilojen joukon avulla. Esimerkiksi shakissa tavoitteena on päästä tilaan nimeltä shakkimatti, jossa vastustajan kuningas on uhattuna eikä voi paeta.
- Reitin kustannus -funktio (engl. path cost function) liittää jokaiseen reittiin sen kustannuksen lukuarvona. Ongelmanratkaisuagentti valitsee kustannusfunktion, joka heijastaa sen menestyksen mittaa (määritelmää) (performance measure; suoritusarvo). (Russell ja Norvig 2003, 62.)

Edelliset elementit määrittelevät ongelman, ja ne voidaan kerätä yhteen yhdeksi tietorakenteeksi, joka voidaan syöttää ongelmanratkaisualgoritmiin. Ratkaisun hyvyys voidaan mitata reitin kustannus -funktiolla; optimaalisella ratkaisulla on pienin reitin kustannus kaikkien ratkaisujen joukossa. (Russell ja Norvig 2003, 62.)

Monissa ongelmissa ratkaisuun johtavia askelia ei voida tietää etukäteen. Tällöin tarvitaan hakua (engl. search), jossa ratkaisua etsitään käymällä läpi eri vaihtoehtoja. Joillekin ongelmille on ominaista ns. kombinatorinen räjähdys, jossa ongelmaan tai sen ratkontaan liittyvien erilaisten mahdollisten valintavaihtoehtojen määrä kasvaa räjähdysmäisesti, kun

tehtävän koko kasvaa. Esimerkiksi kun yhteen päätökseen liittyy kymmenen vaihtoehtoa, kuuden vastaavan päätöksen sarja tarjoaa miljoona vaihtoehtoa. Tällaisissa tapauksissa ei voida käydä läpi kaikkia valintavaihtoehtoja, vaan ratkaiseminen perustuu vaihtoehtojen määrää rajoittavaan tietämykseen. Tällaista tietämystä ovat mm. erilaiset oletukset, rajoitukset ja valintastrategiat. Tietämystä käyttävää hakua kutsutaan heuristiseksi hauksi (engl. heuristic search). Käytettävän hakumenetelmän valintaan vaikuttavat useat käsiteltävän ongelma-alueen piirteet. On otettava huomioon, mikä on kaikkien mahdollisten vaihtoehtojen lukumäärä, onko tieto osin puutteellista tai epätarkkaa, ja onko se pysyvää. (Lehtola ja Honkela 1993, 168,169.)

Jos tavoitetila ei ole sama kuin lähtötila, hakua laajennetaan (expand)) generoimalla joukko uusia tiloja soveltamalla seuraajafunktiota nykyiseen tilaan. Jos yksikään näistä uusista tiloista ei ole tavoitetila, valitaan näistä joku tai jotkin joihin sovelletaan seuraajafunktiota jne. Tätä jatketaan niin kauan kunnes löydetään tavoitetila, tai lopetetaan kun tutkittavia tiloja ei ole enempää generoitavissa, tai kun resurssi, kuten aika loppuu. Hakustrategia (engl. search strategy) määrää sen, mistä solmusta hakua missäkin tapauksessa laajennetaan. (Russell ja Norvig 2003, 69.) Erilaisia hakustrategioita on olemassa suuri määrä. Useimmissa ongelmissa hakustrategiaa valittaessa on käytettävissä vain puutteellisesti tietoa, eikä valinta näin välttämättä kohdistu sopivimpaan hakustrategiaan. (Lehtola ja Honkela 1993, 170.) Hakustrategioita voidaan luokitella eri tavoin. Ei-informoitu hakustrategia (engl. uninformed search, blind search) on sellainen, jossa ei käytetä tiloista enempää tietoa kuin mitä ongelman määritelmässä on annettu. Niinpä ne voivat ainoastaan generoida uusia tiloja ja erottaa tavoitetila ei-tavoitetiloista. Informoitu hakustrategia (engl. informed search, heuristic search) on sellainen, jolla on kyky erotella ratkaisun löytymisen kannalta "lupaavimmat" tilat muista. Toisaalta kaikki hakustrategiat voidaan luokitella sen mukaan, missä järjestyksessä ne laajentuvat eri solmuista. (Russell ja Norvig 2003, 73.)

Hakustrategia voidaan ilmaista ns. ongelmanratkaisualgoritmilla, jonka vaste on joko epäonnistunut yritys, eli huti (engl. failure), tai ratkaisu. Lisäksi jotkut algoritmit saattavat jäädä päättymättömään silmukkaan, jolloin ne eivät koskaan anna vastetta. Algoritmin suoritustehoa voidaan arvioida neljän ominaisuuden perusteella:

- Täydellisyys (engl. completeness): Löytääkö algoritmi ratkaisun, jos se on olemassa?

- Optimaalisuus (engl. optimality): Löytääkö se optimaalisen ratkaisun, eli reitin kustannukseltaan alimman ratkaisun?
- Aikavaatimus (engl. time complexity): Kuinka kauan ratkaisun löytäminen kestää?
- Tilavaatimus (engl. space complexity): Kuinka paljon muistia tarvitaan haun suorittamiseen? (Russell ja Norvig 2003, 62, 71- 72.)

Päättely (engl. reasoning)

Tieto ja päättely ovat tärkeitä agenteille, koska ne mahdollistavat menestyksekkään käyttäytymisen, jota muuten olisi vaikea saavuttaa. Tieto siitä, mitä tuloksia omat toiminnot aiheuttavat, voi auttaa agenttia menestymään monimutkaisissa ympäristöissä. Toiseksi tietoa ja päättely käyttävä agentti voi yhdistellä yleistä tietoa nykyisiin persepteihin päätelläkseen muuten piilossa olevia ympäristön aspekteja nykyisessä tilanteessa. Tietoa ja päättelyä käyttävät agentit ovat usein joustavia. Ne voivat omaksua uusia tehtäviä, ja voivat saavuttaa kompetenssin nopeasti, jos niille ilmoitetaan tai jos ne oppivat uutta tietoa ympäristöstä, ja ne voivat sopeutua ympäristön muutoksiin päivittämällä relevanttia tietämystään. (Russell ja Norvig 2003, 194,195.)

Tietoa ja päättelyä päätöksenteossaan käyttävällä agentilla täytyy olla tietämyskanta (engl. knowledge base, KB) ja päättelymekanismi (engl. inference mechanism). Agentti tallentaa maailmaa koskevia lauseita tietämyskantaansa, käyttää päättelymekanismia johtaakseen uusia lauseita, ja käyttää niitä päättääkseen mitä toimintoja se suorittaa. Jokainen lause on ilmaistu tietämyksen esittämiseen tarkoitetulla kielellä (engl. knowledge representation language), ja esittää jotain väitettä maailmasta. Esityskielen (engl. representation language) määrittelevät sen syntaksi, joka määrittelee kielen oikein muodostettujen lauseiden rakenteen, ja semantiikka, joka määrittelee jokaisen lauseen totuuden jokaisessa mahdollisessa maailmassa tai mallissa. Agentilla täytyy myös olla keino lisätä uusia lauseita tietämyskantaan ja kysyä (engl. query), mitä lauseita siellä on. (Russell ja Norvig 2003, 195, 232.)

Tietämyskanta saattaa jo alkutilassa sisältää jotain taustatietoa (engl. background knowledge). Agentin ohjelma kertoo tietämyskannalle, mitä se havaitsee, ja kysyy mikä toiminto sen tulee suorittaa. Tähän kyselyyn vastaaminen saattaa vaatia monimutkaista päättelyä liittyen esimerkiksi nykyiseen ympäristön tilan selvittämiseen ja mahdollisten toimintojen tuloksiin jne. Toiminnon valittuaan agentti suorittaa sen. Oikean toiminnan kannalta agentin ei tarvitse tietää kaikkea edes itsestään. Agentin tiedon tasolla (engl. knowledge level) on agentin tieto ja tavoitteet, joita se käyttää sovittaakseen käyttäytymistään ympäristöönsä. Agentin toteutuksen tasolla (engl. implementation level) taas voi olla sellaista, mitä agentti ei tiedä. Agentin ei usein ole tarvetta tietää esimerkiksi miten sen sisäiset prosessit tapahtuvat saavuttaakseen tavoitteitaan. Esimerkiksi taksia ajavan agentin ei tarvitse tietää, onko sen maantieteellinen tieto toteutettu linkitettyinä listoina vai pikselikarttoina, tai järkeileekö se manipuloimalla merkkijonoja vai levittämällä signaaleja neuroniverkoissa. (Russell ja Norvig 2003, 195, 196, 197.)

Tiedon esittäminen (engl. knowledge representation)

Monimutkaisissa ja muuttuvissa ympäristöissä toimiakseen agentti tarvitsee yleistä ja joustavaa tiedon esittämistä. Kaiken maailmassa olevan esittämiseen ei voida käyttää menetelmää jossa kaikelle annetaan täydellinen kuvaus, koska kaiken tämän tiedon hankkiminen olisi käytännössä vaikeaa tai mahdotonta, sen säilyttäminen veisi liikaa muistitilaa ja sen käsitteleminen vaatisi liikaa prosessointitehoa. Abstrahointi (engl. abstraction) on prosessi, jossa esityksestä poistetaan tarpeettomia yksityiskohtia (Russell ja Norvig 2003, 63). Abstrahoinnin tavoitteena on poistaa esityksestä mahdollisimman paljon yksityiskohtia mutta niin että sen tulos on käyttökelpoinen. Ilman kykyä luoda käyttökelpoisia abstraktioita, älykkäät agentit joutuisivat kokonaan maailman monimutkaisuuden nielemiksi. (Russell ja Norvig 2003, 64.)

Tiedon esittämistä varten luodaan kategorioita, joihin eri kohteet luokitellaan. Kategoriat ovat käsitteitä, joita pidetään tietyn ontologian kannalta olennaisina. Kategorioista koostuvaa viitekehystä johon tiedon esittäminen perustuu, tai sellaisten muodostamista koskevaa

tutkimusta, sanotaan ontologiaksi. Ontologiat voidaan jakaa erityisiin ja yleisiin ontologioihin. Erityiset ontologiat (engl. special-purpose ontology) ovat vain tietyn, esimerkiksi tietyn tehtävän suorittamiseen tarvittavan tiedon esittämiseen luotuja käsitteellisiä viitekehyksiä, kun taas yleiset ontologiat (engl. general-purpose ontology) sopivat minkä tahansa tiedon esittämiseen. Kaksi tunnuspiirrettä erottaa yleisiä ontologioita erityisistä. Yleisen ontologian pitää soveltua käytettäväksi enemmän tai vähemmän millä tahansa erityisalueella. Toiseksi millä tahansa riittävän vaativalla määrittelyalueella (engl. domain), eri tiedon alueet on esitettävä yhtenäisesti, koska järkeilyyn ja ongelmanratkaisuun saatetaan käyttää monen eri alueen tietoa samanaikaisesti. Muun muassa filosofit ja laskentatieteilijät ovat pohtineet eri tavoin muotoiltuna kysymystä, lähenevätkö erityiset ontologiat tiedon määrän lisääntyessä yhtä yleistä ontologiaa. Vuosisatojen tutkimuksen jälkeen nykyinen vastaus on: ehkä. (Russell ja Norvig 2003, 320, 321, 322.)

Vaikka vuorovaikutus ympäristön kanssa tapahtuu yksittäisten kohteiden kanssa, suuri osa järkeilystä tapahtuu kategorioiden tasolla. Kategorioiden avulla voidaan tehdä kohteita koskevia ennusteita, kun kohteet on luokiteltu, eli sijoiteltu sopiviin kategorioihin. Agentti voi päätellä joidenkin kohteiden läsnäolon saamastaan aistisyötteestä, se voi päätellä kohteen kuulumisen tiettyyn kategoriaan havaitsemansa kohteen ominaisuuksien perusteella, ja sitten käyttää kategoriaan liittyvää tietoa tehdäkseen kohteeseen liittyviä ennusteita. (Russell ja Norvig 2003, 322, 323.) Kategoriat auttavat järjestämään ja yksinkertaistamaan tietoa, koska kategorioita voidaan järjestää niin, että alakategoriat tai alijoukot (tai osajoukot) perivät (engl. inherit) ominaisuuksia yläkategorioilta. Esimerkiksi jos kaikki kategorian *ruuat* instanssit ovat syötäviä, ja jos *hedelmät* on sen alakategoria ja *omenat* on kategorian *hedelmät* alakategoria, niin tiedämme, että kaikki omenat ovat syötäviä. Yksittäiset omenat perivät syötävyyden ominaisuuden, tässä tapauksessa koska ne kuuluvat kategoriaan *ruuat*. (Russell ja Norvig 2003, 323.)

Epävarma tieto ja päättely

Agentit eivät juuri koskaan voi tietää kaikkea tärkeää tietoa ympäristöstään, ja siksi niiden täytyy usein toimia epävarman tiedon perusteella. Ilman kaikkea relevanttia tietoa ympäristöstä agentti ei voi olla varma, johtaako tietty toiminto sen saavuttamaan tavoitteensa, eikä se myöskään tiedä varmasti, mikä sen tavoitteista olisi tällä hetkellä helpoiten saavutettavissa. Rationaalinen päätös toiminnasta riippuu tällöin erilaisten tavoitteiden suhteellisesta tärkeydestä ja todennäköisyyksistä, joilla ne voidaan saavuttaa. (Russell ja Norvig 2003, 462, 463.)

Agentin tiedon epävarmuus voi johtua ainakin kahdesta seikasta: laiskuudesta (engl. laziness) ja tietämättömyydestä (engl. ignorance). Tieto on epävarmaa laiskuuden takia, jos agentilla ei ole täydellistä tietoa koska sen hankkiminen tai käsittely on liian vaativaa joidenkin resurssien kuten energian, ajan tai muistitilan kannalta. Tieto on epävarmaa tietämättömyyden takia, kun agentilta puuttuu joko teoreettista tai käytännöllistä tietoa. (Russell ja Norvig 2003, 463, 464.) Lisäksi on mahdollista, että maailmaan itseensä kuuluu epävarmuutta, joka saattaa aiheuttaa agentin tiedon epävarmuuden.

Agentin epävarma tieto voidaan ilmaista uskomuksina, ja niiden varmuus uskomuksen asteina (engl. degrees of belief). Uskomuksen asteita voidaan käsitellä todennäköisyyslaskennan (engl. probability theory) avulla. Lauseisiin liitetään uskomuksen astetta ilmaiseva lukuarvo nollan ja yhden väliltä, ne mukaan lukien. Todennäköisyyden nolla liittäminen lauseeseen vastaa uskomusta, että lause on epätosi, kun taas todennäköisyyden yksi liittäminen lauseeseen vastaa uskomusta, että lause on tosi. Todennäköisyydet nollan ja ykkösen välillä vastaavat täydellisen epäuskon ja uskon välissä olevaa uskomuksen astetta lauseen totuuteen. (Russell ja Norvig 2003, 464.) Lauseeseen liitetty todennäköisyys kuvaa agentin uskoa, ei suoraan maailmaa. Uskomukset ja uskomusten asteet riippuvat persepteistä, jotka agentti on tähän mennessä saanut. Näistä persepteistä koostuu evidenssi (engl. evidence) eli todistusaineisto, johon todennäköisyydet perustuvat. Evidenssin lisääntyessä todennäköisyydet voivat muuttua. (Russell ja Norvig 2003, 464, 465.) (erilaisista todennäköisyyden tulkinnoista ks. Russell ja Norvig 2003, 472 ja Niiniluoto 1980).

Preferenssien esittämiseen ja niihin liittyvään päättelyyn voidaan käyttää hyötyteoriaa (engl. utility theory). Hyötyteorian mukaan joka tilalla on agentin tavoitteiden kannalta hyödyllisyyden aste, eli utiliteetti, ja että agentti suosii, eli preferoi tilaa, jolla on suurempi utiliteetti. Tilan utiliteetti on suhteellinen agentin preferensseihin, joita agentin utiliteettifunktion (engl. utility function) oletetaan kuvaavan. (Russell ja Norvig 2003, 465.) Preferenssit, utiliteetteina ilmaistuna, yhdistetään todennäköisyyksiin rationaalisen päätöksenteon yleisessä teoriassa, päätöksentekoteoriassa (engl. decision theory): Päätöksentekoteoria = todennäköisyyslaskenta + hyötyteoria. Päätöksentekoteorian keskeinen ajatus on, että agentti on rationaalinen, jos ja vain jos se valitsee toiminnon, jolla odotetaan olevan suurin utiliteetti, kun utiliteetti lasketaan siten, että toiminnon kaikkien mahdollisten tulosten utiliteeteistä lasketaan keskiarvo. Tätä sanotaan korkeimman utiliteetin odotusarvon periaatteeksi (engl. the principle of maximum expected utility, (MEU)). (Russell ja Norvig 2003, 465, 466.)

Suunnittelu (engl. planning)

Tavoitteen saavuttamiseen tarkoitetun toimintosekvenssin luomista sanotaan suunnitteluksi. (Russell ja Norvig 2003, 375.) Suunnittelun lähtökohta on maailman tila toimintosarjan alussa, jota sanotaan alkutilaksi tai lähtötilaksi. Tämä on tieto siitä, mitkä ongelman kannalta relevantit väitteet ovat tosia. Tavoitteet ja alkutila muodostavat yhdessä ongelman. Lisäksi suunnittelijalla on käytössään joukko toimenpiteitä eli toimintoja, joilla se voi muuttaa maailman tilaa tavoitteiden saavuttamiseksi. (Karanta 1993, 183.) Suunnittelija voidaan ajatella ohjelmaksi, joka etsii ratkaisun tai joka konstruktiivisesti todistaa ratkaisun olemassaolon. (Russell ja Norvig 2003, 407.) Suunnittelujärjestelmät voidaan luokitella sen mukaan, miten ne käsittelevät eri aliongelmien välistä vuorovaikutusta. Jotkut suunnittelijat osaavat ratkaista vain ongelmia, joihin pätee niin sanottu lineaarisuusoletus (engl. linearity assumption). Tällöin oletetaan, että seuraava menetelmä johtaa ratkaisuun: ratkaistaan yksi tavoite useasta, joiden pitää olla yhtä aikaa voimassa (konjunktio), ja tämän jälkeen ratkaistaan muut tavoitteet pitäen edellä mainittua tavoitetta ratkaisuna. Hyvin monet ongelmat kuitenkin vaativat useiden aliongelmien ratkomista yhtä aikaa. Useimmat nykyaikaiset suunnittelijat

ovat epälineaarisia suunnittelijoita, jotka eivät perustu lineaarisuusoletukselle. Ne osaavat käsitellä mielivaltaisesti limittyneitä ongelmia ja osaongelmia. (Karanta 1993, 184.)

Suunnittelutehtävän mielekkyyden edellytyksenä pidetään yleensä, että: 1) Maailma, jossa suunnittelu tapahtuu, on ainakin jossain määrin ennustettavissa. Kaoottisissa tai hyvin satunnaisissa ympäristöissä suunnittelu on tehotonta, ja agentti voi vain reagoida tapahtumiin. Lopputulos on sama, jos ympäristö muuttuu niin nopeasti, ettei käytettävissä oleva laskentakapasiteetti riitä suunnitelmien tekemiseen ja muuttamiseen tarpeeksi nopeasti. 2) Käsiteltävä ongelma on mielekkäästi paloiteltavissa lähes erillisiin aliongelmiin, joilla on vain vähän vuorovaikutusta keskenään. (Karanta 1993, 183, 184.)

Suunnittelussa joudutaan kontrolloimaan ns. kombinatorista räjähdystä (engl. combinatorial explosion). Jos määrittelyalueella (engl. domain) on *p* primitiivistä propositiota, silloin sillä on 2^*p* tilaa. Jos määrittelyalue on monimutkainen, *p* voi olla hyvin suuri luku. Pahimmassa tapauksessa suunnittelu on äärimmäisen vaikeata. Parhaassa tapauksessa ongelma on hajotettavissa osiin ja tämä mahdollistaa suunnittelun eksponentiaalisen nopeutumisen. Hajotettavuus häviää, jos toimintojen välillä on negatiivisia vuorovaikutuksia, eli jos toiminto sivutuotteena peruuttaa toisen toiminnon vaikutuksen. Joskus osaongelmat voidaan ratkaista sellaisessa järjestyksessä, että negatiiviset vaikutukset voidaan jättää huomiotta. (Russell ja Norvig 2003, 407.)

Korkeammilla luonnollisilla älykkäillä systeemeillä ja monilla keinotekoisilla älykkäillä systeemeillä on mielikuvitus. Mielikuvitus on työkalu, jolla voi kokeilla turvallisesti eri toimintoja ennen toimintojen käyttämistä ympäristössä. Mielikuvituksella Fritz tarkoittaa kykyä esittää nykyinen tilanne mielessä, käyttää siihen sopivaa vastesääntöä ja esittää näin syntyvä uusi nykyinen tilanne. (Fritz 2007, Mental methods and chains of response rules: habits.) Ihmisten mielikuvitus toimii seuraavasti: Mallit virtaavat kullekin aivokuoren alueelle joko aisteista tai muistin hierarkian alemmilta alueilta. Jokainen aivokuoren alue luo ennusteita, jotka lähetetään hierarkiassa takaisin alaspäin. Jotakin asiaa kuvitellessaan ihminen antaa ennusteidensa kääntyä ympäri ja muuttua syötteiksi. Ihminen pystyy seuraamaan ennusteidensa seuraamuksia tekemättä mitään fyysistä (kehollaan). "Jos tämä tapahtuu, niin

sitten tämä tapahtuu ja sitten tämäkin tapahtuu.", ja näin edelleen. Esimerkiksi shakkipelissä henkilö kuvittelee siirtävänsä ratsun tiettyyn ruutuun ja sen jälkeen visualisoi, miltä pelilaudalla näyttää siirron jälkeen. Tämän kuvan perusteella hän ennustaa, mitä vastustaja tekee ja miltä pelilauta näyttää siirron jälkeen. Lopulta hän päättää kuviteltujen tapahtumasekvenssien perusteella, oliko siirto hyvä vai huono. (Hawkins 2005, 204- 205.)

Jos ympäristössä on toisia agentteja, agentti voi sisällyttää ne omaan malliinsa ympäristöstä, mutta sen täytyy ottaa huomioon, että vuorovaikutus toisten agenttien kanssa on usein erilaista kuin muun luonnon kanssa. Moniagenttiset ympäristöt voivat olla yhteistyöllisiä (engl. cooperative) tai kilpailullisia (engl. competitive). (Russell ja Norvig 2003, 449, 450.)

Yhteistyössä agentit suorittavat jonkin tehtävän yhdessä. Ne tekevät suunnitelmia, jotka määrittävät toimintoja ja lisäksi toiminnot täytyy koordinoida (engl. coordination) keskenään. Tämä voidaan saavuttaa esimerkiksi kommunikaation (engl. communication) avulla. Kahdella agentilla, jotka pelaavat tenniksen nelinpeliä samassa tiimissä on yhteinen tavoite voittaa peli. Ratkaisu moniagenttiseen suunnitteluongelmaan on yhdistetty suunnitelma (engl. joint plan), joka sisältää yhteistyötä tekevien agenttien toiminnot. Yhdistetty suunnitelma on ratkaisu, jos tavoite saavutetaan kun agentit suorittavat oman osuutensa suunnitelmasta. Jos agenteilla on sama tietämyskanta (engl. knowledge base) ja jos tietty suunnitelma on ainoa ratkaisu, agentit voivat itsenäisesti päätellä ja suorittaa sen, jos on olemassa monta eri suunnitelmaa, jotka saavuttavat tavoitteen, niin agentit saattavat valita eri suunnitelmat ja toimia niiden mukaisesti. Tällöin agentit tarvitsevat jonkin koordinointimekanismin päätyäkseen saman yhdistetyn suunnitelman käyttämiseen. (Russell ja Norvig 2003, 450, 451.)

Yksinkertaisin tapa jolla ryhmä agentteja voi varmistaa yhteisen yhdistetyn suunnitelman käytön, on omaksua konventio (engl. convention) ennen suunnitelman suorittamista. Konventio on mikä tahansa yhdistettyjen suunnitelmien valintaa koskeva rajoite sen rajoitteen lisäksi, että suunnitelman täytyy olla ratkaisu. Jotkin rajoitteet ovat niin yleisesti käytössä, että niitä voidaan sanoa sosiaalisiksi laeiksi (engl. social law). Esimerkki sosiaalisesta laista joissakin maissa on auton ajaminen tien oikealla puolella. Konventioita voi muodostua myös evoluutioprosessien kautta. Esimerkiksi jotkin sosiaalisten hyönteisten yhdyskunnat

suorittavat hyvin yksityiskohtaisia yhdistettyjä suunnitelmia, jota edesauttaa yksilöiden yhtenäinen geneettinen koostumus. Kaikissa tapauksissa ei ole tarpeen, että agenteilla on yhdistetty suunnitelma, jossa on malli toisten agenttien toiminnasta, vaan riittää, että jokaisella on oma osuutensa yhdistetyn suunnitelman toiminnoista. (Russell ja Norvig 2003, 452, 453.) Konventiot saattavat olla puutteellisia, joustamattomia tai liian epätarkkoja. Tällaisissa tilanteissa agentit voivat käyttää kommunikaatiota (engl. communication) saavuttaakseen yhteisen tiedon sopivan yhdistetyn suunnitelman valitsemisesta. Esimerkiksi tenniksen nelinpeliä pelaavan tiimin agentit voivat huutaa "Minun!" tai "Sinun!", jos pelkän konvention perusteella on vaikeaa erottaa, kumman vastuulla sen palautus on. Agentit eivät aina tarvitse kieltä tähän, vaan voivat esimerkiksi aloittaa jonkin vaihtoehtoisen suunnitelman suorittamisen, josta toinen agentti voi päätellä, mikä suunnitelma suoritetaan ja voi lähteä siihen mukaan. (Russell ja Norvig 2003, 453, 454.)

Ympäristöt, joissa agenttien hyödyt ovat konfliktissa keskenään, ovat kilpailullisia ympäristöjä. Esimerkkejä tällaisista ovat mm. kahden pelaajan nollasummapelit (joissa toisen voitto on toisen tappio) kuten shakki ym. Kilpailullisessa ympäristössä agentin täytyy 1) tunnistaa että ympäristössä on muita agentteja, 2) laskea joitakin toisten agenttien mahdollisia suunnitelmia, 3) laskea miten toisten agenttien suunnitelmat vuorovaikuttavat sen omien suunnitelmien kanssa, ja 4) valita paras suunnitelma näiden vuorovaikutusten perusteella. (Russell ja Norvig 2003, 454.) Hyödyllinen suunnittelun väline tällaisissa tilanteissa on ns. peliteoria. Peliteoriassa tutkitaan rationaalisten pelaajien strategista vuorovaikutusta, kun ne toimivat preferenssiensä mukaan. Se on tärkeä väline, kun tutkitaan tilanteita, joissa agenttien parhaat toiminnot (niille itselleen) riippuvat toisten agenttien toimintaa koskevista odotuksista. Peliteoreettisia ongelmia ovat pohtineet jo mm. Platon ja Shakespeare, mutta sen yhä kehittyvän matemaattisen formalisaation loivat John von Neumann ja Oskar Morgenstern (1944). (Ross 2008.) Peliteoria on merkittävä tutkimuksen ala, jolla on tärkeitä käytännön sovelluksia. Sitä käytetään nykyään mm. konkurssimenettelyissä (engl. bankrupcy proseedings), tuotekehittelyssä ja hinnoittelussa, sodankäynnissä, tilanteissa joissa käsitellään miljardeja dollareita ja jotka koskevat satojatuhansia ihmishenkiä. (Russell ja Norvig 2003, 631.)

Oppiminen

Älykkään systeemin muistissa olevien käsitteiden ja vastesääntöjen lisääntymistä sanotaan oppimiseksi (Fritz 2007, Glossary: Learning). Oppimalla älykäs systeemi voi muuttaa käyttäytymistään ja saavuttaa tavoitteensa ympäristössä, joka oli sille ennen tuntematon. Tämä tarkoittaa sitä, että ongelmallisessa tilanteessa, eli tilanteessa johon sillä ei ole muistissaan sopivaa toimintoa, se kokeilee eri toimintoja, ja arvioi niiden vaikutuksia. Jos toiminto oli hyödyllinen, eli jos se siirsi älykkään systeemin lähemmäs tavoitettaan, sen aivot tallentavat tämän kokemuksen, voidakseen käyttää sitä tulevaisuudessa samankaltaisissa tilanteissa. (Fritz 2007, The ability to learn.) Oppimiskyvyissä on eroja. Ihmiset ovat yleensä parempia oppimaan kuin muut eläimet, nuoret ihmiset oppivat helpommin kuin vanhat. Joillakin ihmisillä voi olla sellaisia ongelmia aivojen toiminnassa, että he kykenevät oppimaan vain hyvin yksinkertaisia suhteita. Niidenkin ihmisten välillä, joilla on normaali aivojen toimintakyky, on jonkin verran eroja oppimiskyvyssä. (Fritz 2007, What is intelligence? Are Learning Rates A Measure Of Intelligence?) Tulevaisuudessa jotkut keinotekoiset älykkäät systeemit saattavat ohittaa ihmisen oppimiskyvyssä (Churchland 1989).

Oppiva agentti käyttää perseptejään parantaakseen käyttäytymistään tulevaisuudessa (Russel ja Norvig 2003, 649). Oppimisessa voidaan käyttää hyväksi tietoa, joka on saatu aikaisemmasta samanlaisesta tapauksesta. Koska ratkaistavat ongelmat ja kohdatut tilanteet harvoin toistuvat identtisinä, tarvitaan samankaltaisuuksien havainnointia ja yleistämistä (engl. generalization). Yleistämällä pyritään muodostamaan yleisiä periaatteita, joita voidaan soveltaa yksittäistapauksissa. Myös tietokoneohjelmille voidaan kehittää oppivuutta. Mikäli systeemi toimii paremmin kuin aiemmin ilman uudelleenohjelmointia, niin systeemiä voidaan sanoa oppivaksi. (Honkela ja Sandholm 1993, 244, 245.)

Oppimismenetelmät voidaan luokitella kolmeen luokkaan sen perusteella, miten itsenäisesti systeemi käsittelee ja tulkitsee esimerkkiaineistoa:

- Ohjattu oppiminen (engl. supervised learning) perustuu siihen, että opetusesimerkeissä on syötteen lisäksi tiedossa opettajan antama haluttu tulos (Honkela ja Sandholm 1993, 245). Ohjatun oppimisen ongelmaan sisältyy funktion oppiminen syötteistä ja vasteista.

Täysin havaittavissa (engl. fully observable environment) ympäristöissä agentti voi aina havaita toimintojensa tulokset ja voi siksi käyttää ohjatun oppimisen menetelmää ennustaakseen niitä. Osittain havaittavissa ympäristöissä (engl. partially observable environment) ongelma on vaikeampi, koska toimintojen välittömät tulokset eivät välttämättä ole havaittavissa. (Russell ja Norvig 2003, 650.) Induktiiviseen oppimiseen kuuluu konsistentin, eli esimerkkiaineiston kanssa ristiriidattoman hypoteesin löytäminen. Okkamin partaveitsi- periaatteen mukaan monien konsistenttien hypoteesien joukosta tulee valita yksinkertaisin. Laskennallinen oppimisteoria (engl. computational learning theory) analysoi esimerkkien ja hypoteesien monimutkaisuutta. (Russell ja Norvig 2003, 651- 707.)

- Vahvistusoppimisessa (engl. reinforced learning) ulkopuolinen opettaja antaa arvion siitä, kuinka hyvä systeemin antama tulos on. Tarve muutoksiin on sitä suurempi, mitä virheellisemmäksi tulos arvioidaan. (Honkela ja Sandholm 1993, 245.) Vahvistetussa oppimisessa agentti oppii saamansa vahvistuksen (engl. reinforcement) tai palkinnon (engl. reward) mukaan. Esimerkiksi jos taksia ajava agentti ei saa tippiä matkan lopuksi, ja jos sellainen on muuten tapana, niin agentti voi päätellä, että sen käytös ei ollut haluttua. Vahvistettuun oppimiseen sisältyy usein osaongelmana oppia miten maailma toimii joissakin suhteissa. (Russell ja Norvig 2003, 650.)
- Ohjaamatonta oppimista (engl. unsupervised learning) voidaan sanoa myös itsenäiseksi oppimiseksi. Systeemille syötetään esimerkkejä, joista se muodostaa jonkinlaisen mallin itselleen. (Honkela ja Sandholm 1993, 245.) Ohjaamattoman oppimisen ongelmaan sisältyy syötteiden säännönmukaisuuksien (engl. pattern) oppiminen kun mitään erityisiä vasteita ei ole niille varattuna. Esimerkiksi taksia ajava agentti voi ajan myötä oppia käsitteet "hyvä päivä liikenteessä" ja "huono päivä liikenteessä" sen mukaan, miten liikenne on sen kokemusten mukaan sujunut ilman että agentille annettaisiin valmiiksi luokiteltuja esimerkkejä kummastakaan. Pelkästään ohjaamattomasti oppiva agentti ei voi oppia mitä sen pitäisi tehdä, koska sillä ei ole informaatiota siitä mikä on oikea toiminto tai mikä on haluttava tila. (Russell ja Norvig 2003, 650.)

Älykkäiden systeemien on usein hyödyllistä kyetä unohtamaan. Jatkuvasti kasaantuva uusi informaatio alkaa jossakin vaiheessa viedä niin paljon muistitilaa, että sitä kaikkea ei voi säilyttää. Niinpä tehdäkseen tilaa uudelle, tärkeämmälle informaatiolle, älykkään systeemin täytyy hävittää eli unohtaa informaatio, jota se ei pidä tarpeeksi tärkeänä, tai jota ei ole käytetty pitkään aikaan. Usein aivot säilyttävät vain tehokasta abstraktia tai komposiitti-informaatiota ja unohtavat yksityiskohtaisen konkreetin informaation. Esimerkiksi harvat ihmiset muistavat ensimmäisen kirjastokorttinsa yksityiskohtia, mutta muistavat että heillä oli se. (Fritz 2007, Memory and forgetting.)

Tavoitteet

Kaikilla älykkäillä systeemeillä on tavoite (engl. main objective). Monet voivat myös oppia luomaan ja käyttämään alitavoitteita (engl. sub objective). Alitavoitteet ovat alemman tason ja/tai väliaikaisia tavoitteita. Alitavoitteen saavuttamalla älykäs systeemi lähenee tavoitteensa saavuttamista. (Fritz 2007, Objectives.)

Tarkoitushakuisuus tai tavoitteellisuus syntyi elämän mukana noin neljä miljardia vuotta sitten. Evoluutio kasvatti tästä pienestä alusta yhä selvempiä merkityksiä ja arvoja, preferenssijärjestelmän, jonka mukaan eliöt panevat asioita ja elämyksiä arvojärjestykseen. Evoluutio on rakentanut merkitykset ja arvot sisään vietti- ja vaistojärjestelmään ja ne voivat saada evoluutiosta alkunsa jo ennen kuin ne tiedostetaan. Yksinkertaisimmillaan tavoitteellisuus näkyy viruksissa, joilla ei ole solurakennetta, vaan lähes yksinomaan genomi. Ne käyttävät muiden eliöiden soluja monistumiseensa siten, että tunkeuduttuaan soluun ne ensiksi lamauttavat sen genomin toiminnan ja rupeavat sitten solun monistajajärjestelmällä monistuttamaan itseään ja teettämään solun lähetti- ja siirtäjä-RNA-systeemillä itselleen välttämättömiä proteiineja, kuten suojakseen oleskeluun solun ulkopuolella tarvitsemansa vaipan. Monistuneet virukset poistuvat solusta, joka saattaa kuolla, tai toipua hyökkäyksestä. Virus ei "pyri" mihinkään muuhun kuin genominsa monistamiseen, vieläpä täysin toisten kustannuksella. (Salmi 2002, 54- 55.)

Luonnollisten älykkäiden systeemien ympäristöjen vaarat ovat aiheuttaneet sen, että niiden tavoitteeksi on tullut elossa pysyminen ja lisääntyminen (Fritz 2007, Main objective). Yksilöt, joilla on tavoitteita, ominaisuuksia tai tapoja, jotka eivät edesauta tai haittaavat selviytymistä, menehtyvät jonakin vaikeana hetkenä. Tämä johtuu siitä, että ne sijoittavat selviytymiseen tarvittavan energiansa toisiin tavoitteisiin, tai niillä on muita ominaisuuksia tai tapoja, jotka haittaavat selviytymistä. Koska monet niistä menehtyvät ennen kuin ovat ehtineet lisääntyä riittävästi, lajin yksilöiden joukossa on tulevaisuudessa vähemmän tämän tyyppisiä yksilöitä. Paremmin selviytymiseen kykenevillä yksilöillä taas on keskiarvoltaan enemmän jälkeläisiä ja siten suurempi osa seuraavasta sukupolvesta on niitä. (Fritz 2007, The objectives of humans.)

Eläimillä perimän vaikutus näkyy ennen muuta motivaatiotekijöiden alueella. Millä tavoin tavoitteisiin pyritään, riippuu paljosta muustakin kuin perimästä. Vapauden lisäys aivojen kapasiteetin kasvaessa koskee ennen muuta keinojen valintaa, ei niinkään suuressa määrin perimmäisiä tavoitteita, joita ovat elossa säilyminen ja lisääntyminen. Päteekö tämä ihmisiinkin? Nykyaikana suuri osa ihmisistä ei ainakaan tietoisesti pyri tämän tavoitteen saavuttamiseen, mutta Fritzin mukaan ihmisissä voidaan havaita seuraavia toisen tason tavoitteita, jotka kaikki näyttävät auttavan päätavoitteen saavuttamisessa. Ihmiset haluavat ainakin:

1) Hengittää. Jos hengittäminen estyy, yksilö kuolee hyvin lyhyessä ajassa.
2) Kokea nautinnollisia aistimuksia. Niiden kokeminen tarkoittaa, että ruumis ja mieli toimivat hyvin, mikä tarkoittaa terveyttä. Nautinnollisiin kokemuksiin pyrkiminen johtaa usein paremmin selviytymiseen sopivaan mieleen ja ruumiiseen.
3) Syödä ja juoda. Ilman ruokaa ja juomaa yksilö kuolee.
4) Levätä ja nukkua. Tällä tavalla ihmiset palauttavat voimansa. Pitkän aikaa toimiminen ilman lepoa johtaa suorituskyvyn alenemiseen tai totaaliseen toimintakyvyttömyyteen. Tässä tilassa ei voi saavuttaa tavoitteitaan.
5) Itsen suojeleminen luonnolta ja muilta älykkäiltä systeemeiltä. Puolustautumatta jättäminen saattaa johtaa terveyden tai jopa hengen menettämiseen.
6) Rakastella. Kauniin ruumiin, luonteen tai mielen ihaileminen, seurustelu, tunteet joita ihmiset tuntevat, kaikki tämä johtaa lopulta jälkeläisten tuottamiseen.

7) Olla uteliaita. Uteliaisuus voi johtaa uusiin kokemuksiin, joka taas johtaa uusien vastesääntöjen oppimiseen ja parempaan selviytymiseen tulevaisuuden tilanteissa.
8) Olla yhdessä toisten kanssa. Ryhmä ihmisiä voi tehdä asioita, joita yksi ihminen ei. Ryhmässä yksilöllä on paremmat mahdollisuudet selviytyä.
9) Yrittää olla hyvä jossakin. Tämä johtaa jatkuvaan kehitykseen toimia tehokkaammin (Fritz 2007, The objectives of humans.), ja olla hyödyllinen ja tärkeä yhteisön jäsen. Tämä voi johtaa mm. sosiaaliseen menestykseen.

Kaiken yllämainitun tekemiseksi valitsemme vielä alemman tason tavoitteita. Esimerkiksi haluamme:
10) Keksiä uusia työkaluja kulutushyödykkeiden tuottamiseksi. Niitä käyttämällä voidaan tuottaa enemmän lyhyemmässä ajassa. Tällöin ihmiset tuottavat enemmän toisten hyvinvointiin tarvitsemia kulutushyödykkeitä, ja saavat palkaksi enemmän rahaa, mikä lisää hyvinvointia ja parantaa selviytymisen mahdollisuuksia. Ylemmän ja keskitason tuloluokkiin kuuluvat elävät alempaan tuloluokkaan kuuluvia kauemmin.
11) Kehittää yhteisöä. Paremmassa yhteisössä jäsenten on helpompaa elää ja voida hyvin.
12) Oppia (tähän kuuluu myös tieteellinen tutkimus). Oppiminen auttaa saavuttamaan vanhoja ja uusia tavoitteita.
13) Kokea taidetta. (Fritz 2007, The objectives of humans.) Taideteokset tuottavat arvokkaita, opettavia ja elämää rikastuttavia kokemuksia, joita muuten olisi vaikea saada.
14) Harrastaa liikuntaa. Se voi auttaa pysymään hyvässä kunnossa fyysisiä toimintoja varten, mikä antaa paremman kyvyn toimia ympäristössä ja muuttaa sitä.
15) Pelata älypelejä. Mielen kykyjä haastavat pelit pitävät mielen hyvässä kunnossa, mikä (älyllisten toimintojen alueella) auttaa samaan tapaan kuin fyysinen kunto. (Fritz 2007, The objectives of humans.)

Näyttää siltä, että tavoitteet, joita voidaan havaita mieleltään terveessä, keskiarvoisessa ihmisessä, ovat melkein aina jollain tavalla selviytymistä edesauttavia. (Fritz 2007, The objectives of humans.) Muutama vuosisata sitten alitavoitteiden suhde selviytymiseen oli ilmiselvä. Työskenneltiin melkein kaikki valveillaoloaika pelkästään selviytymistä varten. Nykyään ihmisten ei tarvitse työskennellä niin paljon, ja heillä on paljon vapaa-aikaa, jolle

voitaisiin periaatteessa valita mikä tahansa tavoite. Kuitenkin voidaan havaita, että enimmäkseen vapaa-ajalle valitut tavoitteet ovat yhteydessä selviytymiseen jollain tavalla. (Fritz 2007, The objectives of humans.), mikä johtuu luultavasti siitä, että tavoitteiden valintaan vaikuttavat tunteet ja motivaatio on viritetty selviytymistä varten, eikä niihin voi vaikuttaa. Kaukaisessa tulevaisuudessa voi olla niin, että suurin osa ihmisten tavoitteista ei enää perustu selviytymiseen, mikä saattaa johtaa ihmislajin sukupuuttoon kuolemiseen (Fritz 2007, The objectives of humans).

Keinotekoisten älykkäiden systeemien tavoitteet voivat olla mitä tahansa mitä ohjelmoija ohjelmoi niihin. Ihmistä palvelevan robotin tavoitteena voi olla esimerkiksi ihmisen hyväksymisen maksimointi ja moitteen minimointi. (Fritz 2007, Main objectives.) Joissakin kirjoissaan Isaac Asimov ehdotti ja kokeili kolmea tavoitetta, joista hän käytti nimitystä "Robotiikan (robottien) kolme lakia". Ne ovat:
1) Robotti ei saa toiminnallaan vahingoittaa ihmistä, eikä jättää toimimatta niin, että ihminen joutuu kärsimään.
2) Robotin täytyy totella ihmisen antamia käskyjä paitsi jos ne ovat ristiriidassa ensimmäisen lain kanssa.
3) Robotin täytyy suojella itseään paitsi jos se on ristiriidassa ensimmäisen tai toisen lain kanssa. (Seiler ja Jenkins 2004. What are the three laws of robotics anyway?)

Tavoitteiden ilmaisu

Tavoite on usein määritelty ns. menestyksen kriteeristöllä (engl. performance criteria) jonka täyttämällä agentti saavuttaa tavoitteensa. Suoritusarvo (engl. performance measure) kuvaa, kuinka hyvin toiminta täyttää menestyksen kriteerit. Tavoitteen saavuttaakseen agentti suorittaa toimintoja usein ympäristöstään saamansa palautteen perusteella. Menestyksellinen toiminta vastaa ympäristön tilan muuttumista tavoitetilaksi toiminnan avulla. Ei ole olemassa yhtä ja pysyvää menestyksen kriteeristöä, joka sopisi kaikille agenteille, vaan agenteilla on eri tavoitteita, ja menestys määräytyy niiden saavuttamisen mukaan. (Russell ja Norvig 2003, 35.) Menestyksen tunnistamisen menetelmät eivät useinkaan ole täydelliset. Esimerkiksi jos

automaattisen pölynimurin tehtävänä on pitää lattia siistinä, kuinka sen menestys voidaan määritellä ja mitata? Tämä tieto on välttämätöntä, jotta se voisi tunnistaa menestyksen ja muokata käyttäytymistään sen perusteella. Voitaisiin ehdottaa, että menestystä mitattaisiin sen kahdeksantuntisen työvuoron aikana imemän pölyn määrällä, mutta tällöin se voi maksimoida menestyksensä imemällä pölyt lattiasta, heittää ne sen jälkeen takaisin lattialle, imemällä ne uudestaan jne. Parempi ehdotus olisi palkita imuri siitä, että lattia on puhdas. (Russell ja Norvig 2003, 33- 35.) Tavoitetesti-funktio (engl. goal test) ratkaisee, onko tietty tila tavoitetila. Joskus mahdollisten tavoitetilojen joukko voidaan esittää eksplisiittisesti, jolloin tavoitetesti yksinkertaisesti selvittää onko annettu tila yksi niistä. Joskus taas tavoite on määritelty abstraktien ominaisuuksien avulla. Esimerkiksi shakissa tavoitteena on päästä tilaan nimeltä shakki matti, jossa vastustajan kuningas on uhattuna eikä voi paeta. (Russell ja Norvig 2003, 62.)

Vastaavia ongelmia kohtaavat myös luonnolliset älykkäät systeemit. Eläimillä, ihmiset mukaan lukien, menestys on joskus koodattu mielihyvän tunteeksi. "Mielihyvä on variaabeli, joka viittaa siihen, kuinka lähellä älykäs systeemi on tavoitteensa saavuttamista." (Fritz 2007, What is intelligence?, The ability to learn, objectives.), mutta se tuskin kuvaa menestystä täydellisesti. Ihmisen ja muiden eläinten päätavoitteita ovat selviytyminen ja lisääntyminen (Fritz 2007, Main objective). Molempiin tarvitaan toimivaa ruumista ja ruumiin terveyden ja toimintakyvyn ylläpitämiseksi tarvitaan ravintoa. Oletetaan että ruokahalu yrittää ilmaista tätä ravinnontarvetta. Leptiini on rasvasolujen erittämä hormoni. Se vaikuttaa hypotalamuksen neuroneihin, jotka säätävät nälän ja tyydytyksen tunteita. Normaaleilla hiirillä tehdyissä kokeissa havaittiin, että kun hiiri saa riittävästi syödäkseen, sen leptiinitaso nousee, ja se siirtyy ruoasta muihin nautintoihin. Jotkut hiiret ovat kuitenkin poikkeavia. Ne ovat lihavia ja ne jatkavat syömistä silloinkin, kun niiden leptiinitaso nousee. Geneettinen analyysi paljastaa, että leptiiniä sitovalla reseptorilla on useita mutaatioita, ja erityiset mutaatiot ennustavat, miten lihava eläin on. Yleistäen jos ihminen saa syntymässään tietyn leptiini-reseptorin mutaation, ja jos hän sen seurauksena on yhtä nälkäinen päivällisen jälkeen kuin sen alussa, tuntuu väistämättömältä, että hän syö liikaa. (Churchland 2002, 254- 258.)

Toiminnassaan ja/tai ongelmanratkaisussaan älykkään systeemin täytyy usein kyetä vertailemaan monien eri tavoitteiden tärkeyttä tietyssä tilanteessa. Yksi tällainen tapa on kiinnittää tavoitteeseen sen tärkeyttä vastaava seuraavan kaavan mukaan saatava numero nollan ja yhden väliltä: **I = w / 86400**, missä **I** on tavoitteen tärkeys (**I**mportance). **w** on sekunteja vuorokaudessa, jotka älykäs systeemi on halukas (**w**illing) käyttämään saavuttaakseen tavoitteensa. **86400** on sekuntien määrä vuorokaudessa. (Fritz 2007, Objectives.) Tavoitteen saavuttamiseksi käytetty todellinen aika on vahvasti riippuvainen ympäristön tekijöistä. Esimerkiksi monille älykkäille systeemeille juominen (ruumiin tarvitseman veden nauttiminen) on hyvin tärkeä tavoite, ja ne ovat halukkaita käyttämään paljon aikaa tämän tavoitteen saavuttamiseksi. Tavallisesti nesteen hankkiminen ja sen juominen ei vie juuri ollenkaan aikaa, mutta jos älykäs systeemi on keskellä kuumaa autiomaata, nesteen hankkimiseksi tarvittava aika on paljon suurempi. (Fritz 2007, Objectives.)

Tunteet kuuluvat ihmisten mielen perusmekanismeihin. Vaikuttaa siltä, että tunteet ilmaisevat sitä, onko ihminen saavuttanut tavoitteensa vai ei, tai että onko toinen älykäs systeemi saavuttanut tavoitteensa vai eikö. Fritzin mukaan onnellisuus tarkoittaa, että yksilö on saavuttanut omat tavoitteensa (Fritz 2007, Happiness). Näin ajattelee myös Heylighen: Onnellisuus ihmisellä viittaa siihen, että hän on biologisesti sopeutunut (lähellä optimaalista tilaa) ja kognitiivisesti kontrollissa (kykenee korjaamaan poikkeamat tästä optimaalisesta tilasta), toisin sanoen hän kykenee tyydyttämään kaikki perustarpeensa huolimatta mahdollisista häiriötekijöistä. (Heylighen, F. 1999) Joskus sanotaan, että sisialaiset kalastajat, Tyynen meren saarten asukkaat tai muut muusta ihmiskunnasta syrjässä elävät olivat onnellisimpia ihmisiä maan päällä. Mistä tämä johtuu? Näillä ihmisillä oli vähän ja helposti saavutettavia tavoitteita. He olivat eristyksissä muusta sivilisaatiosta, eivätkä tienneet kaikista mukavista asioista, joita mainoksissa, lehdissä ja televisiossa kuvaillaan. (Fritz 2007, Happiness.) Nykyajan kaupunkilaiset kuulevat monista asioista ja niinpä heillä on paljon tavoitteita, joista joitakin on vaikea saavuttaa. Niinpä he ovat jatkuvasti tyytymättömiä ja onnettomia. Tästä pitäisi ottaa opiksi ja analysoida niitä välittömiä tavoitteita, joita ihmisillä on. Onko niiden saavuttamiseen tarvittava panostus (engl. effort) järkevän suuruinen? Onko mahdollista määritellä panostuksen ja tavoitteen suhde? Pitäisikö valita muita tavoitteita? (Fritz 2007, The objectives of humans.)

Eri tunteet voidaan selittää älykkäisiin systeemeihin liittyvällä käsitteistöllä seuraavasti:
Positiiviset tunteet:
1) Mielihyvä, onnellisuus: Tavoite saavutettiin.
2) Voitto: Tavoite saavutettiin, vaikka toinen älykäs systeemi yritti estää sitä.
3) Ihailu: Toinen älykäs systeemi saavuttaa aina tavoitteensa, positiivisesti nähtynä.
4) Ystävällisyys: Toista älykästä systeemiä kohtaan, joka monta kertaa auttaa saavuttamaan tavoitteen.
5) Rakkaus: Luontoa, elämää, kaikkea sellaista kohtaan, mikä mahdollistaa tavoitteen saavuttamisen. (Fritz 2007, Emotions and energy.)

Negatiiviset tunteet:
1) Suuttumus: Tavoitetta ei saavutettu.
2) Turhautuminen: Tavoitetta ei saavutettu, monista yrityksistä huolimatta.
3) Suru: Tiedetään, että tilanne tai toinen älykäs systeemi, joka aiemmin teki tavoitteiden saavuttamisen helpommaksi, ei enää ole.
4) Viha, kauna, vihamielisyys: Toinen älykäs systeemi jatkuvasti estää saavuttamasta tavoitetta.
5) Kateus: Toinen älykäs systeemi saavuttaa aina tavoitteensa, negatiivisesti nähtynä.
6) Epätoivo: Ymmärretään, että tavoitetta ei voi saavuttaa.
7) Masentuneisuus: Viimeisimmässä menneisyydessä suurta osaa tavoitteista ei saavutettu. (Fritz 2007, Emotions and energy.)

Kaikkiin näihin tunteisiin liittyy biokemiallisia reaktioita, jotka joskus häiritsevät rauhallista ajattelua. Tunteet ovat tarttuvia. Tunteet, joita ihmiset havaitsevat ympäristössään, tuottavat heille samanlaisia tunteita. Fritzin mukaan aivojen optimaalisen toiminnan kannalta vaikuttaa siltä, että pitkällä aikavälillä positiivisia tunteita pitäisi olla negatiivisia enemmän. Jos ihminen on jatkuvasti ympäristössä, jossa negatiiviset tunteet ovat vallalla, hän ei voi toimia hyvin vaan kärsii ympäristön tunnesaasteesta. (Fritz 2007, Emotions and energy.)

Toiminnot

Älykkään systeemin näkökulmasta toiminto on pitkän prosessin viimeinen osa. Tämä prosessi alkaa kun se saa informaatiota aistiensa kautta. Sitten se luo nykyisen tilanteen, jonka jälkeen se valitsee siihen sopivan vastesäännön, ja viimeiseksi suorittaa vastesäännön vaste-osan (toiminnon), jonka tuloksena ympäristö muuttuu. Se saa aistiensa kautta informaatiota muuttuneesta ympäristöstä, ja prosessi alkaa uudelleen. (Fritz 2007, Actions, As the end of a process.)

Toimiminen tarkoittaa, että älykäs systeemi käyttää energiaa muuttaakseen ympäristöään jonkin tavoitteen saavuttamiseksi. Toiminto on älykkään systeemin vaste, joka on suhteessa sen saamaan aistisyötteeseen tai tietoon ympäristön nykyisestä tilanteesta. Toimiminen edellyttää tavoitetta, koska ilman tavoitetta systeemillä ei ole syytä valita syötteeseen liittyvää toimintoa. (Fritz 2007, What is intelligence?, Acting on the environment, Action.) Jotkut objektit kuten kivet, ilmamolekyylit tai vesi eivät toimi, vaan niiden liike johtuu niihin kohdistuneesta ympäristöstä tulevasta voimasta. (Fritz 2007, What is intelligence?, Acting on the environment.) Jotkin objektit toimivat, vaikka sitä ei heti tule ajatelleeksi. Jääkaappi toimii kun se pitää lämpötilan tietyllä tasolla. Lentokoneen automaattiohjaus toimii säätäessään lentokoneen korkeutta, kurssia ja muita joidenkin muuttujien arvoja. Ihmiset ovat sisällyttäneet tavoitteita näihin systeemeihin, ja ne toimivat niiden mukaisesti. Kasvit toimivat tietyissä tilanteissa sopeutuakseen paremmin ympäristöönsä. Kasvin runko ja oksat kasvavat pituutta päästäkseen muiden kasvien tai kivien ja puiden varjosta ja kääntyvät valoa kohti. Sen juuret kasvavat kosteutta kohti. Kasveilla on sisään rakentuneita, biologisen evoluution ohjelmoimia tavoitteita, joiden mukaan ne toimivat. (Fritz 2007, What is intelligence?, Acting on the environment.)

Suhteessa muihin älykkäisiin systeemeihin toiminta voidaan luokitella yhteistyöksi, neutraaliksi tai hyökkääväksi. Toiminta on yhteistyötä, kun älykäs systeemi auttaa toista saavuttamaan tavoitteensa. Tämä on yleensä vastavuoroista. Toiminta on neutraalia, jos se ei vaikuta muihin älykkäisiin systeemeihin, tai jos se ei auta tai estä niitä. Toiminta on hyökkäävää, jos se estää toista tai toisia älykkäitä systeemejä saavuttamasta tavoitteitaan. (Fritz 2007, Actions, Classification of actions.)

Toiminto muuttaa nykyisen tilanteen tulevaisuuden tilanteeksi. Tämä voidaan esittää täsmällisellä tavalla matemaattisesti. Esitys voi näyttää esimerkiksi seuraavanlaiselta:

A(St) = S(t+1)

Missä A on toiminto (engl. action), t on aika (engl. time) ja S tilanne (engl. situation). (Fritz 2007, Actions, The action "causes " a transformation.)

Toimintaa eli tavoitteen saavuttamiseen tähtäävän toiminnon (tai toimintosekvenssin) suorittamista voidaan sanoa käyttäytymiseksi. Esimerkiksi yksisoluinen *Euglena*-suvun siimalevä aistii solunsa valoherkällä jyväsellä valoa ja hakeutuu sitä kohti. Tosin se ei tunne valon tulosuuntaa mutta tunnistaa valon voimakkuuserot. Se lähtee uimaan siimansa avulla sattumanvaraiseen suuntaan. Jos se sattuu uimaan väärään suuntaan ja valon määrä alkaa vähetä, se muuttaa suuntaansa sattumanvaraisesti niin monta kertaa, että valon määrä alkaa lisääntyä. Tällä menetelmällä se pääsee veden pintakerrokseen, jossa on eniten valoa sen lehtivihreähiukkasen toimintaa varten. Siimalevän toimintaa voidaan kutsua käyttäytymiseksi. Se otti ympäristöstään informaatiota yksisoluisen ruumiinsa tähän tehtävän erikoistuneella osalla ja reagoi sitten tähän tietoon ainoan solunsa liikuntaan erikoistuneiden soluelinten toiminnalla mahdollisimman yksinkertaisesti yrityksen ja erehdyksen avulla. Siimaeliöllä on 1) aistinelin, 2) toimijaelin sekä 3) kyky siirtää tieto aistinelimeltä toimijaelimelle. Nämä osat täytyy olla kaikilla käyttäytyvillä organismeilla, ja Ilman näitä kolmea ei voida puhua käyttäytymisestä. (Viitala 2004, 23.)

3.3 Älykkäiden systeemien vuorovaikutus, etiikka ja yhteisöt

Päätöksentekoa, toimintaa, vuorovaikutusta ja yhteisöllistä käyttäytymistä voidaan analysoida älykkäiden systeemien käsitteistölllä. Tämä voi auttaa ymmärtämään yhteisöjen toimintaa, tehtäviä ja sitä, millainen merkitys yhteisöllä on yksilön elämänlaadulle. Nykyään yhteisöjen tutkimus (yhteiskuntatieteet) ovat usein ns. "pehmeää" tiedettä. Sillä tarkoitetaan tutkimusta tai oppialaa, jonka oletetaan olevan tieteellistä, mutta joka ei kuitenkaan perustu toistettavista

kokeista saatavaan tietoon, ja/tai tämän tiedon matemaattiseen selittämiseen ja järjestelyyn. Sen vastakohdaksi mainitaan usein ns. "kova tiede", kuten fysiikka tai biologia, jolla on edellä mainitut ominaisuudet. Yhteisöjen tutkimuksen koventamiseksi tai tieteellistämiseksi yritetään keksiä ratkaisuja (ks. esim. Black 2000a ja Black 2000b). Tätä kehitystä voidaan auttaa muodostamalla niin täsmällisesti määriteltyjä käsitteitä, että niitä voidaan käyttää matemaattisissa kaavoissa. Hyvä ehdokas tähän on älykkäisiin systeemeihin liittyvää käsitteistö. (Fritz, 2007 Human societies.)

3.3.1 Vuorovaikutus

Älykäs systeemi voi olla ympäristössä, jossa ei ole muita älykkäitä systeemejä (yksiagenttinen ympäristö) tai sellaisessa, jossa on (moniagenttinen ympäristö). Vuorovaikutus toisten älykkäiden systeemien kanssa on erilaista kuin muun luonnon kanssa, koska muu luonto on välinpitämätön sen pyrkimyksistä, kun taas toiset älykkäät systeemit useinkaan eivät ole. Ne voivat reagoida monin eri tavoin. Vuorovaikutus toisten älykkäiden systeemien kanssa voi olla hyödyllistä tai haitallista, koska se voi auttaa älykästä systeemiä saavuttamaan tavoitteensa tai estää sitä (Fritz 2007, Ethics of the artificial intelligent system). Tältä kannalta etiikan tutkimus ja sen tulokset koskevat sitä, miten älykkään systeemin on hyödyllistä käyttäytyä moniagenttisissa ympäristöissä (Fritz 2007, Glossary: Ethics, Ethics of the artificial intelligent system), suhteessa toisiin agentteihin.

- Jos älykäs systeemi on eristyksissä, ympäristössä, jossa ei ole muita älykkäitä systeemejä, sillä ei ole tarvetta etiikalle. Alitavoitteiden ja niitä vastaavien vastesääntöjen valinta on suoraviivaista, koska ympäristössä ei ole reaktioita muilta älykkäiltä systeemeiltä. (Fritz 2007, Ethics of the artificial intelligent system.)
- Monet älykkäät systeemit ovat ympäristössä, jossa niillä on kontakteja muiden älykkäiden systeemien kanssa, jotka reagoivat sen toimintoihin. (Fritz 2007, Ethics of the artificial intelligent system.) Vuorovaikutus toisten älykkäiden systeemien kanssa voi olla hyödyllistä tai haitallista, koska se voi auttaa älykästä systeemiä saavuttamaan tavoitteensa tai estää sitä.

Älykäs systeemi voi yrittää tehdä vuorovaikutuksesta hyödyllistä (eli saada toinen tai toiset käyttäytymään halutulla tavalla) mm. seuraavilla tavoilla:

- Suostuttelu

Suostuttelussa yritetään saada toinen käyttäytymään halutulla tavalla ilman voimankäyttöä. Koska toinen on älykäs systeemi, se yrittää saavuttaa omia tavoitteitaan. Jotta se voitaisiin saada käyttäytymään halutulla tavalla, näitä tavoitteita täytyy muuttaa. Tämä voidaan tehdä esittämällä toiselle uusia tavoitteita, jotka ovat (tai joiden hän uskoo olevan) itselleen nykyisiä hyödyllisempiä, ja jotka ovat samalla suostuttelijalle hyödyllisiä. Suostuttelussa voidaan vedota tunteeseen tai järkeen (riippuen siitä, mikä on tehokkainta). Usein tunteet ovat järkisyitä tehokkaampia välineitä ihmisten suostutteluun. Ne ovat tarttuvia, joten kun ilmaisemme tunteemme, tämä todennäköisesti tuottaa saman tunteen toisessa. Ilo tuottaa iloa, pelko pelkoa, ja innostus innostusta. Jos tunteiden käyttäminen epäonnistuu, käytetään järkiperusteluja, jolloin toisen mielenkiinto täytyy herättää antamalla yksityiskohtaista tietoa kyseessä olevasta asiasta, ja kertomalla miksi haluttu toiminta on hyödyllistä. (Fritz 2007, Interaction between intelligent systems.)

- Voimankäyttö

Voimankäytössä toiselle luodaan niin haitallisia tilanteita, että hän pitää hyödyllisempänä käyttäytyä halutulla tavalla. Tämä voidaan tehdä vahingoittamalla häntä joko henkisesti tai fyysisesti. Sitä ennen toista voidaan uhkailla, eli kommunikoida hänelle, millä tavalla ja miten paljon häntä vahingoitetaan, jos hän ei toimi halutulla tavalla. Tilanne, jossa älykäs systeemi toimii niin, että se estää toista saavuttamasta tavoitettaan, on hyökkäys. Normaalisti tämä aiheuttaa sen, että toinenkin hyökkää. Molemminpuolinen hyökkääminen on taistelu. Taistelun tulos on usein toisen hyödyksi ja toisen haitaksi. Usein lopuksi siitä on molemmille haittaa tuhlattujen resurssien ja vaihtoehtoiskustannusten (menetetyn hyödyllisemmän käyttäytymisen) takia. Jos molemmat ovat saman yhteisön jäseniä, yhteisö yrittää yleensä estää taistelun ja selvittää konfliktin. (Fritz 2007, Using force, Ethics of the artificial intelligent system.)

- Yhteistyö

Yhteistyössä älykkäät systeemit toimivat auttaen toisiaan saavuttamaan tavoitteensa. Yhteistyö tulee mahdolliseksi, jos toisella on ylimäärä tietyn tyyppistä hyödykettä, taitoja, työkaluja tai tietoja, joita se voi vaihtaa toisen tyyppisiin, joita toisella on ylimäärä. Yhteistyön tulos on molemmille hyödyllinen, mutta ei aina yhtä paljon; normaalisti toinen hyötyy enemmän kuin toinen. (Fritz 2007, Cooperation.) On kaksi tapaa tehdä yhteistyötä:

- Molemmat toimivat saavuttaakseen yhteisen tavoitteen.
- Molemmat toiminnallaan auttavat toista saavuttamaan oman tavoitteensa.

Usein älykäs systeemi saavuttaa tavoitteensa helpommin yhteisön jäsenenä. Älykkäät systeemit voivat perustaa yhteisön, jos ne saavuttavat siten paremmin tavoitteensa kuin muuten. Yhteisö syntyy kun monet älykkäät systeemit toimivat saavuttaakseen yhteisen tavoitteen. (Fritz 2007, A scientific philosophy.) Ihmisille yhteisö on hyvin tärkeä, koska hänen elämänlaatunsa on suuresti siitä riippuvainen. Ilman yhteisöään energinen ja älykäskin ihminen eläisi "sivistymättömänä villinä", jonka elintaso olisi paljon alempi. (Fritz 2007, Human societies.)

- Vallan käyttö

Vallankäyttö on sitä, että älykäs systeemi saa toisen toimimaan oman tavoitteensa mukaisesti. Se voi muuttaa tai muokata toisen systeemin tavoitteita. Valtaa voidaan mitata esimerkiksi tunteina kuukaudessa, jotka toinen on valmis käyttämään älykkään systeemin tavoitteiden saavuttamiseen. Jotkut ihmiset ovat huomanneet, että heille on hyödyksi organisoida yhteistyötä. Tätä tehdään esimerkiksi taloudellisissa (tuotanto ja kaupankäynti) ja yhteiskunnallisissa (hallinto, uskonto, järjestöt) asioissa. Organisointi on hyödyksi sekä organisoijalle että organisoitaville, mutta antaa valtaa organisoijalle. Tätä valtaa käytetään kaikkien hyödyksi, mutta omia tavoitteitaan saavuttaakseen organisoija voi myös väärinkäyttää sitä enemmän tai vähemmän. Jos yhteistyötä ohjaava organisaatio on hyvin suuri, yksi vallassa oleva ei kykene käsittelemään kaikkia yksityiskohtia. Hän joutuu valtuuttamaan muita hoitamaan organisaation eri yksityiskohtia. (Fritz 2007, Power.)

Valtaa voidaan hankkia ainakin seuraavilla tavoilla:

- Antamalla tietoa, joka muuttaa toisen tavoitteita. Tämä toimii parhaiten silloin, kun toinen ei huomaa sitä, kuten mainoksissa tehdään.

- Lupaamalla toimia hyödyksi toiselle tulevaisuudessa.
- Taistelemalla, joko sanoilla tai teoilla.
- Käyttämällä kysymyksiä, jotka saavat toisen ajattelemaan ja muuttamaan tavoitteitaan.
- Luomalla toiselle tavan totella, mikä tekee vallankäytön helpommaksi tulevaisuudessa. Se joka tottelee, tottuu siihen ja luovuttaa helposti vallan toiselle. Tottelija voi olla oppinut menneisyyden kokemuksista, että omien tavoitteidensa saavuttamiseksi sen on parempi toimia niin kuin vallan omistaja haluaa. (Fritz 2007, Power.)

3.3.2 Etiikka

Älykkäät systeemit -maailmankatsomuksen käsitteistöllä etiikan tutkimus ja sen tulokset koskevat sitä, miten älykkään systeemin on hyödyllistä käyttäytyä moniagenttisissa ympäristöissä. Etiikkaa tarvitsevat kaikki moniagenttisissa ympäristöissä toimivat älykkäät systeemit. Etiikan tutkimuksen tulokset ovat eettisiä teorioita ja toimintaohjeita.

Yleensä ajatellaan, että etiikkaan kuuluu oikeaa ja väärää käyttäytymistä koskevien käsitysten systematisointi, puolustaminen ja suositteleminen. Nykyään etiikka jaetaan yleensä kolmeen osaan: metaetiikkaan, normatiiviseen etiikkaan ja soveltavaan (käytännölliseen) etiikkaan:

- Metaetiikassa tutkitaan sitä mistä eettiset käsitykset tulevat, ja mitä ne tarkoittavat sekä onko olemassa mielipiteistä riippumatonta eettistä systeemiä jota voitaisiin soveltaa joka tilanteessa riippumatta ajasta ja paikasta (Newall 2005). Ovatko eettiset käsitykset sosiaalisia keksintöjä? Ilmaisevatko ne muuta kuin yksilöllisiä tunteita? Mikä on järjen rooli eettisissä arvostelmissa? Mistä oikeudet ja velvollisuudet tulevat? (Fieser 2006).
- Normatiivisessa etiikassa etsitään, ja tutkitaan oikeaa tai väärää käyttäytymistä sääteleviä periaatteita. Tavoitteena on löytää normi, testi tai kriteeri, joka ilmaisee, mitä on oikea käyttäytyminen (Newall 2005). Siihen kuuluu oikean käyttäytymisen, velvollisuuksien ja oikeuksien sekä niiden seurausten artikulointi (Fieser 2006).
- Soveltavassa etiikassa yritetään ratkaista tiettyjä eettisiä ongelmia käyttämällä metaetiikan ja normatiivisen etiikan tarjoamia käsitteellisiä työkaluja. Siinä tutkitaan

tiettyjä kiistanalaisia moraalisia aiheita, kuten nykyään abortti, lapsenmurha, eläinten oikeudet, ympäristöasiat, homoseksuaalisuus, kuolemanrangaistus tai sotaoikeus. (Fieser 2006.)

Etiikka ja moraali erotetaan yleensä toisistaan. Moraalisuus tarkoittaa teon, elämäntavan tai päätöksen oikeellisuutta tai vääryyttä, joka riippuu siitä, noudattaako se tiettyä ohjetta. Etiikka on tällaisten ohjeiden tai standardien tutkimusta. (Newall 2005.)

Etiikan tieteellisyys

Eettiset periaatteet selitetään tulevan yleensä joistakin seuraavista lähteistä:

- Jumalalta (mm. eri uskonnot, kuten kristinusko)
- Abstraktista maailmasta, jossa käsitteet ovat olemassa jollakin tavalla (mm. Platon)
- Ihmisten sopimuksesta (mm. Rousseau)
- Vastuun pohdinnasta (mm. Pufendorf, Locke, Kant (Fieser 2006))
- Hyveen pohdinnasta (mm. Aristoteles, MacIntyre (Fieser 2006))
- Eri toimintojen tuloksia koskevasta arvioinnista (mm. Bentham, Mill, Moore (Fieser 2006)). (Newall 2005.)

Uskontoon, perinteisiin ja tunteisiin perustuvia etiikoita voidaan kritisoida kankeudesta (oppimiskyvyttömyydestä); ne eivät yleensä ole muutettavissa tarpeen vaatiessa (Churchland 1996, 291- 292). Ne ovat myös perusteina ongelmallisia: Mihin uskonto perustuu? Miten ja miksi se on syntynyt? Onko varma että juuri käytössä oleva uskonto on oikea? Miksi tiettyihin (usein perustelemattomiin tai huonosti perusteltuihin) uskomuksiin nojautuen voitaisiin saada käytännön elämässä paras tulos? Mistä voidaan tietää, että perinteen uskomukset ja käytännöt sopivat uusin tilanteisiin? Mistä tunteet ovat tulleet? Jos tunteet ovat evoluution kulussa kehittyneitä päätöksentekomekanismeja, mistä tiedetään, että ne sopivat uusiin tilanteisiin, joihin ei evoluution aikana ole vielä jouduttu, ja ovatko ne parhaita edes niihin tilanteisiin, joihin ovat kehittyneet? Eettiset toimintaohjeet perustuvat usein uskonnollisille opetuksille. Ne voivat olla hyvin vanhoja, ne eivät ole aina selkeitä ja täsmällisiä, joten niitä ei voida soveltaa

nykyajan olosuhteissa. Toisaalta etiikkaa perustellaan joskus moraalisilla tunteilla tai yhteisön perinteillä. Kuitenkin teknologia muuttuu hyvin nopeasti, ja moraaliset tunteet ja yhteisön perinteet hyvin hitaasti. Niinpä ne eivät ole päivitettyjä nykyajan oloihin. (Fritz 2007, Create a scientific ethics.) Jumalaa tai abstraktia maailmaa voidaan pitää ongelmallisina etiikan lähteinä, koska ei tiedetä onko niitä olemassa, miten niihin saadaan yhteys, minkälaista etiikkaa ne ajavat, miten ne liittyvät ihmisten tai muiden älykkäiden systeemien elämään ja vuorovaikutukseen, ja miksi meidän pitäisi harjoittaa niiden ajamaa etiikkaa. Sopimukseen, vastuuseen ja hyveeseen perustuvat etiikat ovat ongelmallisia, koska niitä esitettäessä ei yleensä selitetä, mihin ne itse perustuvat ja miksi juuri ne ovat parhaita. Voi olla tyhmiä, huonoja, jne. sopimuksia ja vastuita, tai hyveitä, jotka voivat olla hyödyksi vain vähän, tai haitaksi. Miten selvitetään parhaat?

Jos omaksutaan käsitys, että etiikan tutkimus ja sen tulokset koskevat sitä, miten älykkään systeemin on hyödyllistä käyttäytyä moniagenttisissa ympäristöissä, niin etiikan perusta voi olla vain eri toimintojen tuloksia koskeva arviointi. Kaikilla älykkäillä systeemeillä on tavoite. Hyödyllistä on se mikä auttaa tavoitteen saavuttamisessa. Vuorovaikutus toisten älykkäiden systeemien kanssa voi olla hyödyllistä, neutraalia tai haitallista. Vuorovaikutuksessa älykkäät systeemit ottavat huomioon toisten käyttäytymisen omia toimintojaan valitessaan. Vuorovaikutuksessa toimintojen tulokset ovat osaksi toisten reaktioita. Etiikka on sen kohteena olevien älykkäiden systeemien käyttäytymistä koskeva sääntöjen kokoelma, jonka tavoitteena on ohjata niiden välistä vuorovaikutusta mahdollisimman hyödylliseksi. Kun yksittäiset älykkäät systeemit vuorovaikuttavat moniagenttisissa ympäristöissä, ne saattavat joissain tilanteissa kokea (omalta kannaltaan) hyödyllisimmäksi omaksua osittain tai kokonaan yhteinen alitavoite ja tehdä yhteistyötä. Voi käydä niin että yhteistyön tekijöiden muodostama ryhmä (ehkä yhteisö) koetaan niin hyödylliseksi (edelleen jokaisen omalta kannaltaan), että niiden on hyödyllistä myös puolustaa yhteisön jäseniä ja sitä kokonaisuudessaan. Näin omasta hyödystään kiinnostuneet yksilöt voivat ajautua tilanteeseen, jossa ne auttavat muita. Ei ole selvää, millaisten periaatteiden avulla vuorovaikutus olisi yksilöiden ja yhteisön kannalta kaikkein hyödyllisintä.

On olemassa yleinen ja enimmäkseen julkilausumaton oletus, että etiikan, moraalin ja politiikan periaatteet ovat erilaisia kuin tieteen. Ajatellaan että tieteelliset periaatteet esittävät objektiivisia faktoja, kun taas moraaliset ja poliittiset periaatteet ilmaisevat vain subjektiivisia tunteita, toiveita, keinotekoisia sääntöjä, tai kunkin hetken tyrannin sortoa tai hyväntahtoisuutta. Tämä vastakkainasettelu on pinnallista. Mm. USA:lla ja Englannilla on kolme tai neljä sataa vuotta vanha perustuslaillinen hallitus. Valtioilla voi olla keskeytyksiä perustuslaillisen hallituksensa historiassa, tai se voi olla lyhyt, mutta niiden kehityksen tapa on sama. Asiaan kuuluvat lakia säätävät elimet muotoilevat ja luovat uusia sosiaalisia toimintaperiaatteita - kieltäen joitakin, rajoittaen ja kannustaen toisia käyttäytymisiä - vasteena ympäristön muutoksiin ja jo käytössä olevien sääntöjen vaikutuksiin. Pysyvät poliittiset instituutiot ovat olemassa vastatakseen muokatuilla tai uusilla toimintaperiaatteilla tai laeilla huomattuun tyytymättömyyteen ja epäoikeudenmukaisuuteen. Tätä voidaan pitää oppimisprosessina: Voimassa olevat lait ovat yleensä menneisyyden kokemusten perusteella tehtyjen keksintöjen ja korjausten tulosta. Sekä tieteessä että etiikassa ihmiset yrittävät ennustaa ja manipuloida kohteita joidenkin periaatteiden avulla. Yritykset eivät yleensä onnistu täydellisesti, ja periaatteita korjataan saatujen kokemusten valossa. Molemmissa tapauksissa tuloksena on parempi käsitys siitä, miten maailma toimii ja miten tavoitteita voidaan saavuttaa. (Churchland 1996, 286- 288.)

Askel tieteellisen etiikantutkimuksen suuntaan olisi yhteisen käsitteellisen viitekehyksen luominen sosiaalisen käyttäytymisen tutkimukseen. Eri käyttäytymistieteet - joihin kuuluvat mm. taloustiede, biologia, antropologia, sosiologia, psykologia ja politiikan tutkimus, sekä niiden osat, mm. neurotiede, arkeologia ja paleontologia sekä osaksi samoja ilmiöitä tutkivat historia, oikeustiede ja filosofia - nykyään mallintavat ihmiskäyttäytymistä erilaisilla ja ainakin osittain yhtyeensopimattomilla tavoilla. Kuitenkin viimeisimmät teoreettiset ja empiiriset kehitykset ovat tuoneet mahdollisuuden yhteisen viitekehyksen luomiseen. (Gintis 2007, 1.) Evoluutioteoria, johon kuuluvat sekä geeni- että kulttuurievoluutio on käyttäytymistieteiden yhdistävä periaate (Gintis 2007, 1). Homo sapiens on evoluutiossa kehittynyt laji, jonka piirteet ovat sen evoluutiohistorian tulosta. Koska kulttuuri vaikuttaa ihmisten geneettiseen koostumukseen, ihmisten kognitiiviset, affektiiviset ja moraaliset kyvyt ovat geeni-kulttuurievoluution dynamiikan tulosta. Geeni-kulttuurievoluutiossa geenit

sopeutuvat ympäristöön, jossa kulttuuri on tärkeä tekijä. Tämän sopeutumisen tuloksena olevat muuttuneet genotyypit ovat tulevan kulttuurievoluution perusta. Yleiset evoluution periaatteet viittaavat siihen, että eliöiden yksilöllinen päätöksenteko voidaan käsittää rajallisten informaatio- ja materiaaliresurssien rajoittamaksi preferenssifunktion optimoinniksi. Luonnonvalinta varmistaa, että preferenssien sisältö heijastaa biologista kelpoisuutta, ainakin ympäristöissä, joissa preferenssit ovat kehittyneet. Tämä yhteisevoluutioprosessi on tuottanut ihmisiin preferenssejä myös yhteistyöhön, reiluuteen, rangaistukseen, empatian kykyyn, sekä kykyyn arvostaa rehellisyyttä, työteliäisyyttä, erilaisuuden sietämistä, ja uskollisuutta. (Gintis 2007, 1-2.) Peliteoria tarjoaa peruskäsitteistön valintojen analysointiin strategisessa vuorovaikutuksessa, kunhan jotkut sen klassiset mm. oletukset agenttien itsekkyydestä (lukuun ottamatta tiettyjä tilanteita, kuten nimettömät vuorovaikutukset markkinoilla), ja rationaalisuudesta hylätään. Peliteoria tutkii strategista vuorovaikutusta, joka on mm. elävien systeemien sosiaalisen käyttäytymisen perusta, kun oletetaan, että ne valitsevat kelpoisuuteensa nähden relevantteja toimintoja vasteena toisten yksilöjen toiminnoille. Jos päätöksentekoteoriaa ja peliteoriaa laajennetaan käsittämään myös agenttien toisia huomioivia, ns. other-regarding -preferenssejä, niiden avulla voidaan mallintaa kaikkia päätöksenteon aspekteja, mukaan lukien myös normaalisti "psykologisina", "sosiologisina" tai "antropologisina" pidetyt. (Gintis 2007, 2-3; Skyrms ja harms, 1-2.)

Varmin menetelmä selvittää eettisten periaatteiden hyödyllisyyttä on kokeellinen tutkimus. Tuottaako tietty periaate muihin verrattuna parhaan tuloksen? Jos sitä ei jostain syystä voida kokeilla käytännössä, voidaan käyttää erilaisia teoreettisia apuvälineitä, simulaatioita, malleja, spekulaatioita, jne. Etiikka voidaan luoda tutkimalla, mitkä ovat älykkäiden systeemien tavoitteet, millä tavalla toimimalla ne saavuttavat parhaiten tavoitteensa, ja minkälaisia älykkäiden systeemien vuorovaikutusta koskevia toimintaa ohjaavia sääntöjä kannattaa luoda, jotta tämä onnistuisi parhaiten. (Fritz 2007, Create a scientific ethics.) Etiikka joka ei perustu auktoriteeteille, vaan kokeelliselle tieteelle, jota voidaan verifioida ja kehittää, voisi johtaa vakaampiin ja hyödyllisempiin tuloksiin kuin mitä etiikan historiassa on saatu aikaan. Eettisiä kokeita ei usein voida tehdä ihmisillä, mutta mm. tietokonesimulaatiot ovat mahdollisia, ja niitä on tehty. Tietokonesimulaatiot voivat antaa paljon nopeammin tuloksia kuin kokeet luonnollisella populaatiolla, jos niitä ylipäätänsä olisi mahdollista suorittaa. Jos jotain menee

pieleen, kukaan ei kärsi siitä. Kokeellinen etiikan tutkimus on vasta alussa, mutta siitä tulee luultavasti erittäin hyödyllistä ennen pitkää, koska se voi tuottaa selkeitä tuloksia. (Fritz 2007, Experimental ethics, Experimental sosiology, Create a scientific ethics.) Seuraavaa voidaan pitää tieteellisen etiikan tutkimuksen esimerkkinä: Kun tutkittiin tietokonesimulaatiolla yhteistyön evoluution malleja, osoittautui kaikissa tilanteissa parhaaksi taktiikaksi tit for tat -strategia. Sen mukaa yksilön kannattaa aina ensin olla altruistinen ja sen jälkeen tehdä samoin kuin vastapuoli. Strategiaan kuuluu, että ollaan myös anteeksiantavaisia. Jos kerran pettänyt yksilö käyttäytyykin taas altruistisesti, niin parhaan hyödyn saa, jos sitä kohtaan käyttäytyy tämän jälkeen taas altruistisesti. Tit for tat -strategiaa tukevia havaintoja on tehty monista lajeista kaloista kädellisiin. (Viitala 229- 230.) Esimerkkejä etiikan tutkimuksesta empiiriseen dataan perustuen: (Henrich, Boyd, Bowles, ja muut 2005 ja Orr, 1999), ja tietokonesimulaatioiden avulla: (Danielson, P. 1998), (Jaffe, K. 2002), (Macy ja Willer 2002), (Saam, N. J. and Harrier, A. 1999), (Younger, S. M. 2002 ja 2003).

Etiikan universaalius

On luotu monia etiikoita, moniin eri tarkoituksiin. Niitä voi olla yksilöillä, pienemmillä ja suuremmilla yhteisöillä. Ne voi olla tarkoitettu vain hyvin rajattuun käyttöön, tai äärimmillään voivat koskea kaikkia. Etiikoita luodaan, koska niistä on ollut hyötyä niitä käyttäville älykkäille systeemeille. Ne säätelevät älykkäiden systeemien välisiä vuorovaikutuksia, ja niiden hyödyllisyys johtuu niitä käyttävien vuorovaikutusten hyödyllisyydestä. Laajempi etiikka olisi luultavasti hyödyllisempi, koska sen hyödylliset vaikutukset ovat laajempia. Se mahdollistaisi laajemman yhteistyön tavoitteiden saavuttamiseksi ja haitallisten asioiden välttämiseksi.

Universaali, kaikkien älykkäiden systeemien käyttämä etiikka olisi hyödyllinen mm. seuraavista syistä:

- Yhteinen etiikka parantaisi yhteistyön mahdollisuuksia. Ehkä hyödyllisin älykkäiden systeemien välinen vuorovaikutussuhde on hyvin organisoitu yhteistyö. Siksi voidaan ajatella, että hyvä eli hyödyllinen etiikka tuottaa hyvin organisoitua yhteistyötä. Mm.

liikemiesten ja poliittisten johtajien täytyy usein toimia ihmisten kanssa, joilla on erilainen uskonto ja erilainen etiikka. Maailman globalisoituessa tarvitaan entistä enemmän etiikkaa, joka ei ole vain jonkun yksittäisen ryhmän kannattamaa, vaan jonka kaikki voivat hyväksyä. Muuten maailmanpolitiikka ja talous eivät voi toimia tehokkaasti ja kaikille hyödyllisellä tavalla, ja seurauksena voi olla esimerkiksi sotia, köyhyyttä, sairauksia ja ympäristöongelmia. (Fritz 2007, Create a scientific ethics.)

- Yhteinen etiikka estäisi tai vähentäisi hyökkäyksiä ja taisteluja. Se voisi myös parantaa kommunikaatiota, toiminnan perusteluja (kun kaikki joutuvat perustelemaan toimintansa samalla kielellä ja samoista lähtökohdista), ja oikeuden käytäntöönpanoa.
- Universaali etiikka välttäisi negatiiviset ulkoisvaikutukset. Kun kaikki kuuluvat saman etiikan piiriin, ei yhden osan etua voida jättää huomiotta.

Tarvitaan universaali älykkäiden systeemien etiikka, joka sopii kaikille älykkäille systeemeille ja niiden yhteisöille. Niihin kuuluisivat ihmiset, korkeammat eläimet, robotit, alkuasukasheimot, yritykset, valtiot ja mahdolliset vierailevat ulkoavaruuden oliot. (Fritz 2007, Create a scientific ethics.) Universaali etiikka täytyy perustaa sille, mikä on kaikille yhteistä ja pätee siten kaikkiin. (Fritz 2007, Ethics for all intelligent systems.) Paras menetelmä universaalin etiikan luomiseksi on kokeellinen menetelmä, eli tieteellinen tutkimus. (Fritz 2007, Create a cientific ethics.)

Hyvän käsite

Eetiikan tutkimuksessa hyvän käsite on keskeinen. Yleensä hyvällä tarkoitetaan jotain tavoittelemisen arvoista. Älykkäisiin systeemeihin liittyvän käsitteistön avulla hyvä voidaan määritellä seuraavasti: Älykkäillä systeemeillä on tavoitteita (Fritz 2007, Definition of the intelligent system). Periaatteessa tavoitteet voivat olla mitä tahansa. Älykäs systeemi oppii käyttäytymään niin, että se saavuttaa tavoitteensa: Se aistii ympäristöään, ja kokemustaan käyttäen valitsee toiminnon ja suorittaa sen. (Fritz 2007, Create a scientific ethics.) Seuraava ohjaava periaate pätee kaikkiin älykkäisiin systeemeihin, ihmiset mukaan lukien:

- Älykäs systeemi toimii aina saavuttaakseen tavoitteensa. (Fritz 2007, Interaction between intelligent systems.)

Tämä koskee myös tilanteita, joissa se on vuorovaikutuksessa muiden älykkäiden systeemien kanssa. Ainakaan nykyään ei tiedetä, onko olemassa esim. universaalia, transsendentaalia, absoluuttista, ikuista, jne. hyvää ja miten siitä voidaan tietää. Ympäristössään toimiessaan älykäs systeemi saa kokemuksia ja vähitellen oppii, mikä auttaa sitä saavuttamaan tavoitteensa. (Fritz 2007, Ethics of the artificial intelligent system.) (Fritz 2007, A scientific philosophy.)

- Älykkäälle systeemille se, mikä auttaa sitä saavuttamaan tavoitteensa, on hyvää, ja se, mikä estää, on pahaa.

Yksiagenttisissa ympäristöissä tilanne on yksinkertaisempi kuin moniagenttisissa, joissa toiset älykkäät systeemit saattavat reagoida sen toimintoihin, ja älykkään systeemin täytyy ottaa tämä huomioon arvioidessaan oman toimintansa hyödyllisyyttä.

Ihmiset pitävät usein hyvänä ainakin sitä, mistä on hyötyä heidän tai heidän sukulaistensa tai liittolaistensa kelpoisuuden kannalta. Luonnonvalintaan perustuvan evoluution toimintaperiaate on eri geenien vaihteleva lisääntyminen erilaisten synnynnäisten ominaisuuksien seurauksena. Perinnölliset variantit jotka johtavat kilpailijoitaan menestyksellisempään lisääntymiseen edustavat suurempaa osuutta seuraavasta sukupolvesta. Luonnonvalinta on arvovapaa siinä mielessä, että mikä tahansa menestyksekkäämpään lisääntymiseen johtava ominaisuus siirtyy todennäköisesti jälkeläisille. (Duntley ja Buss 2004, 113.) Toiminnan vaikutus toisen yksilön kelpoisuuteen voi vaihdella. Jos vaikutus kelpoisuuteen on positiivinen, sitä pidetään hyvänä. Negatiivisella puolella esim. toiseen törmääminen koulun käytävällä tai varpaille astuminen ovat todennäköisesti vain pieniä harmeja, kun ääripäässä ovat mm. ryöstö, raiskaus, kidutus ja murha. (Duntley ja Buss 2004, 114.) Sen lisäksi että ihmiset ovat kehittyneet tuottamaan hyötyä toisilleen, heille on myös kehittynyt adaptaatioita haitan tuottamiseen, mm. valehteluun, pettämiseen, varastamiseen, vahingoittamiseen ja murhaan. Koska ne ovat yleisiä ja usein evolutiivisesti hyödyllisiä, niitä voidaan pitää perustavina ja universaaleina ihmisluonnon osina, eikä niitä perimmiltään voida pitää esim. tietyn median, vanhempien, opettajien, kapitalismin tai kulttuurin aiheuttamina. Toisaalta ihmisille on kehittynyt myös adaptaatioita näiden haittojen välttämiseen ja estämiseen, kuten tuntemattomien pelko, stereotypioiden käyttö, muukalaisviha, valheen tai

huijauksen huomaaminen, ryhmästä poissulkeminen. Haittojen tuottamisen ja estämisen kykyjen välillä on varustelukilpa. (Duntley ja Buss 2004, 120.)

Lisäksi ihmisillä on kyky omaksua spesifimpiä käsityksiä tavoitteista ja hyvästä yhteisönsä kulttuurista. Myyttiin perustuvissa riittimenoissa (nykyään esim. uskonnolliset ja valtiolliset tilaisuudet, urheilukilpaillut ja sota) ihmiset voidaan kiihottaa ekstaasiin, jossa he ovat valmiit omaksumaan yhteisön edustamat käsitykset hyvästä ja pahasta. Samalla riittiin osallistuvat vahvistavat yhteisöön kuulumisen tunnettaan ja välittävät eteenpäin käsityksiään oikeasta käyttäytymisestä ja yhteiskunnan järjestyksestä. Yhteinen tavoite, esim. todellinen tai kuviteltu vaara voi olla samanlainen yhdistävä tekijä. Ihmisen geeneihin ei ole ohjelmoitu tiettyä yhteiskuntajärjestystä, vaan vain alttius riiteissä tapahtuvalle joukkosuggestiolle ja sitä kautta tapahtuvalle sosiaalistumiselle ja lauman ohjaukseen alistumiselle: eettisen normin ja yhteiskunnan jäsenyyden ja yhteisön käyttäytymismallien omaksumiselle. Tämä alttius on evoluution tuotetta, vaikka itse myytit ja riitit ovat kulttuurista perimää. (Viitala 2004, 230-232.) Toinen ainakin ihmisillä esiintyvä ryhmän tapa painostaa yksilöä käyttäytymään tietyllä tavalla on ns. moralistinen rangaistus. Mm. pelkurit, karkurit ja petkuttajat voivat joutua aikaisempien tovereidensa hyökkäysten kohteiksi, yhteiskuntansa hyljeksimiksi, juoruilun kohteiksi, tai he eivät pääse alueilleen tai saa puolisoita. Jos moralistinen rankaiseminen on yleistä ja rangaistukset riittävän ankaria, yhteistyö kannattaa. Moralistinen rangaistus voi vakiinnuttaa kaikenlaisen mielivaltaisen käyttäytymisen, esimerkiksi solmion pitämisen, eläinten ystävällisen kohtelun tai kuolleiden sukulaisten aivojen syömisen. Merkitystä ei ole sillä, tuottaako käyttäytyminen etuja ryhmälle. Tärkeää on vain, että kun moralistiset rankaisijat ovat yleisiä, rangaistuksi tuleminen maksaa enemmän kuin moitteeton käyttäytyminen, mitä se sitten onkaan. (Richerson ja Boyd 2006, 252, 253.)

Eettinen toiminta ja toiminnon eettisyyden arvioiminen

Etiikkaa voidaan tarkastella älykkäiden systeemien näkökulmasta seuraavasti: Koska ihminen on älykäs systeemi, ja voidaan ainakin jossain määrin selvittää, mitä mahdollisimman älykäs systeemi tekisi tietyssä tilanteessa, tämä on se miten ihmisenkin tulisi toimia. (Fritz 2007,

Extensive summary.) Kun älykäs systeemi toimii hyödyllisesti (tavalla mikä auttaa sitä saavuttamaan tavoitteensa) moniagenttisissa ympäristöissä, sen toimintaa voidaan sanoa eettiseksi. Fritzin mukaan seuraava periaate on ainakin nykyään paras ohje toimintaan moniagenttisissa ympäristöissä (ohje, joka käytännössä tuottaa hyödyllisen tuloksen), ja niinpä eettisellä toiminnalla voidaan myös tarkoittaa tämän ohjeen seuraamista. On myös mahdollista, että ohjetta voidaan kehittää tai että löytyy parempi.

Moniagenttisissa ympäristöissä toiset älykkäät systeemit saattavat reagoida sen toimintoihin, ja älykkään systeemin täytyy ottaa tämä huomioon arvioidessaan toimintansa hyödyllisyyttä. Fritzin mukaan hyödyllisten reaktioiden todennäköisyyden lisäämiseksi ja haitallisten vähentämiseksi moniagenttisissa ympäristöissä älykkäiden systeemien kannattaa käyttää toimintoja, jotka:
1) auttavat sitä saavuttamaan tavoitteensa, ja
2) ovat samalla enemmän hyödyksi kuin haitaksi kaikille, joihin toiminta vaikuttaa. (Fritz 2007, Ethics as a science.)

Kohtaan (2) on kaksi pääsyytä: Sitä noudattamalla älykäs systeemi voi (ehkä) välttää muiden haitalliset reaktiot ja tehdä yhteistyötä. Jos toiminnosta on jollekin toiselle enemmän haittaa kuin hyötyä, tämä saattaa reagoida ei-toivotulla tavalla. Älykäs systeemi ei voi tehdä laskelmia niiden osalta, joihin se ei tiedä toiminnon vaikuttavan. Toiminto on melkein varmasti eettinen, jos siitä on hyötyä sen suorittavalle älykkäälle systeemille, ja enemmän hyötyä kuin haittaa kaikille muille, joihin se tietää toiminnon vaikuttavan. (Fritz 2007, Ethics of the artificial intelligent system.)

Jos toiminto on enemmän haitallinen kuin hyödyllinen vain joillekin muille, älykäs systeemi voi yrittää selvittää, onko mahdollista muuttaa toimintoa tuottamaan lisää hyötyä näille ja korjata näiden kokeman haitan ja hyödyn suhdetta. Esimerkiksi hallitus voi maksaa tietyn summan rahaa korvaukseksi ihmisille, joiden omistama maa joutuu veden alle padon takia. (Fritz 2007, Ethics of the artificial intelligent system.) Yksityiset ihmisetkin usein yrittävät vakuuttaa ystävyyttään, apua tai hyötyä tulevaisuudessa eli luvata kompensoida toiminnasta toiselle aiheutunutta haittaa tulevaisuuden hyödyllä. Esimerkiksi äiti voi sanoa lapselleen: "Jos

teet tämä nyt niin kun tulemme kotiin, annan sinulle makeisen." (Fritz 2007, Ethics of human beings.)

Arvioidessaan vaihtoehtoisten toimintojen eettisyyttä älykkään systeemin täytyy ottaa huomioon välittömien toiminnan tulosten lisäksi myös toisten reaktiot ja lyhyen sekä pitkän aikavälin vaikutukset, niin hyvin kuin se voi. Älykkäillä systeemeillä on yleensä rajoitetusti tietoa ympäristöstään, eivätkä ne yleensä tiedä kaikkia toiminnan vaikutuksia. Myös aika oikean vastesäännön löytämiseen ja valitsemiseen on rajoitettua. Niinpä se ei voi useinkaan löytää parasta vaan mahdollisimman hyvän ja eettisen toiminnon. (Fritz 2007, Ethics of the artificial intelligent system.)

Toiminnon eettisyys voidaan laskea seuraavalla kaavalla:

E1 = (A1-D1)
E2 = (A2-D2)
E3 = (A3-D3)
...
En = (An-Dn)

E tarkoittaa toiminnon eettisyyttä. Toiminnon aiheuttama hyöty A (engl. Advantage) ja haitta D (engl. Disadvantage) sijoitetaan yhtälöön kunkin sen vaikuttaman älykkään systeemin (1, 2, 3, jne.) kohdalle. Jos E:n arvo on jako tapauksessa positiivinen, niin toiminto on varmasti eettinen, muussa tapauksessa se voi olla epäeettinen. A ja D mitataan tavoitteiden tärkeyden mukaan, joka taas mitataan sekunteina päivässä, jotka älykäs systeemi on halukas käyttämään tavoitteen saavuttamiseksi. (Fritz 2007, Ethics of the artificial intelligent system.) Normaalisti ihmisten tavoitteet voidaan tietää iän, sukupuolen, työn, ympäristön tai muiden tietojen perusteella. Tällainen laskeminen on melko työlästä, ja sitä tarvitaan luultavasti vain hyvin tärkeiden päätösten kohdalla, yrityksen tai valtion päätöksenteossa. Päivittäisessä elämässä ihmisten välillä kokemukseen perustuvan arvion pitäisi olla riittävä. (Fritz 2007, Ethics of human beings.)

Kaikki älykkäät systeemit eivät välttämättä osaa tai halua käyttäytyä eettisesti. Tällöin toiset älykkäät systeemit tai yhteisö voi yrittää muuttaa sen käytöstä. Jos älykäs systeemi ei osaa käyttäytyä eettisesti, sitä voidaan opettaa antamalla tietoa eri tilanteihin liittyvästä eettisestä käyttäytymisestä. Ehkä tulevaisuudessa mm. eettistä käyttäytymistä voidaan harjoitella tietokonesimulaatioiden avulla. Jos älykäs systeemi ei halua käyttäytyä eettisesti, eli se uskoo saavuttavansa tavoitteensa paremmin toiminnalla, josta on enemmän haittaa kuin hyötyä toisille, tämä voidaan yrittää estää osoittamalla, että sen usko ei pidä paikkaansa. Tämä voidaan tehdä joko sanoin tai teoin, uhkaamalla tai rankaisemalla. Mahdollinen rangaistus vähentää todennäköisyyttä, että älykäs systeemi valitsee tietyn vasteen tulevaisuudessa tekemällä siitä todennäköisesti vähemmän hyödyllisen.

Rankaisun käyttäminen ihmisyhteisöissä on kehittynyt luultavasti seuraavien strategisten vaiheiden kautta: Kaikkeen yhteistyöhön liittyy aina tyypillinen ongelma: jos kaikki muut yhteisössä tekevät yhteistyötä, silloin suurin voitto odottaa sitä yksilöä, joka ryhtyy pettämään. Evoluution odottaisi tuottavan sitä tukevan genotyypin. Evolutiivisesti tällainen systeemi ei olisi tasapainossa, sillä pettäminen väistämättä yleistyisi, koska siihen ei sisälly suurta panostusta. Lauman pettävään yksilöön kohdistama rangaistus voisi torjua pettämistä. Rankaisusta on aina kustannusta sen suorittajalle, koska siitä on vaivaa ja rankaisijakin voi vahingoittua, jos petturi puolustautuu niin kuin se todennäköisesti tekee. Niinpä parhaan hyödyn saisivat yksilöt, jotka eivät osallistu rankaisuun, joten palataan lähtökohtaan. Mahdollinen ratkaisu on tulkita kaikki rankaisuun osallistumattomat myös pettureiksi. (Viitala 2004, 229; Skyrms ja Harms, 7-10.) Ihmisillä onkin voimakas taipumus rankaista sellaisia, jotka eivät tee yhteistyötä (Richerson, Boyd, Gintis ja Bowles 2003).

3.3.3 Yhteisöt

Yhteisöjen evoluutio

Useimpien lajien yksilöt elävät melko yksinäistä elämää ja kohtaavat lajikumppaneitaan yleensä vain paritellakseen tai joskus hoitaakseen jälkeläisiään yhdessä. Sosiaalisillakin

lajeilla yhteistyö on yleensä rajoittunut sukulaisiin ja ehkä pieniin vastavuoroisiin ryhmiin. Lyhyen hoiva-ajan jälkeen yksilöt hankkivat itse kaiken ravintonsa. Ei ole työnjakoa, vaihtokauppaa, eikä laajoja konflikteja. Kommunikaatio on rajoittunut pieneen repertuaariin signaaleja. Kukaan ei hoida sairaita, vammaisia tai ruoki nälkäisiä. Vahvat ottavat heikoilta ilman pelkoa kolmannen osapuolen väliintulosta. (Richerson ja Boyd 2006b, 1-2. ja Le, S. ja Boyd, R. 2007, 8) Monimutkaista sosiaalisuutta voidaan odottaa syntyvän vain kun tavoitteiden yhtyminen tuottaa hyödyn, joka ylittää tavoitteiden ristiriidoista koituvat haitat. Eläinten sosiaalisen käyttäytymisen syynä ovat usein tarve yhteiseen puolustukseen ja yhteisten jälkeläisten tuottamiseen. (Workman 2004, 194.)

Monimutkaista sosiaalisuutta voi syntyä ainakin kahdella tavalla. Kun ryhmät koostuvat geneettisistä sukulaisista, valinta voi suosia käyttäytymistä, joka heikentää yksilön kelpoisuutta, kunhan siitä seuraa riittävä ryhmän kelpoisuuden lisäys. Valinta suosii sukulaisiin kohdistuvaa altruismia, koska sukulaiset ovat geneettisesti samanlaisia. Täydet sisarukset voivat olla varmoja siitä, että heidän geeneistään 50 prosenttia on yhteisiä, koska heillä on yhteiset vanhemmat, ja siksi heillä on varaa auttaa sisarusta lisääntymään, kunhan kelpoisuuden tuotto on kaksi kertaa suurempi kuin sen kustannukset. Kaukaisemmat sukulaiset edellyttävät suurempaa hyödyn ja hinnan suhdetta. Tätä periaatetta sanotaan usein Hamiltonin säännöksi, ja se selittää suuren määrän käyttäytymisiä ja morfologiaa hyvin monenlaisissa eliöissä. (Richerson ja Boyd 2006, 249- 250.) Monimutkainen sosiaalisuus mahdollistuu myös silloin, kun vuorovaikutus on toistuvaa, vaikka vuorovaikuttajat eivät olisikaan sukulaisia. Oletetaan että eläimet elävät sosiaalisissa ryhmissä ja sama yksilöpari vuorovaikuttaa pitkän ajan. Usein parin toisella jäsenellä on mahdollisuus auttaa toista jollain itselleen koituvalla hinnalla. Oletetaan, että ryhmässä on kahta tyyppiä: petkuttajia jotka eivät auta ja vastavuoroisia, joiden strategia on: "Auta heti tilaisuuden tullen. Sen jälkeen auta kumppaniasi niin kauan kuin se auttaa sinua, mutta jos se ei auta, älä auta sitä enää." Aluksi kumppanit valitaan satunnaisesti, joten ensimmäisen tilaisuuden tullen vastavuoroiset auttavat todennäköisemmin kuin petkuttajat. Kuitenkin ensimmäisen vuorovaikutuksen jälkeen vain vastavuoroiset saavat apua. Jos vuorovaikutukset kestävät riittävän pitkään, vastavuoroisten suuri kelpoisuus tällaisessa parinmuodostuksessa riittää nostamaan vastavuoroisten keskimääräisen kelpoisuuden suuremmaksi kuin petkuttajien. (Richerson ja Boyd 2006, 250.)

Jos yhteistyö perustuu pelkkään vastavuoroisuuteen, sen täytyy rajoittua varsin pieniin ryhmiin, koska petkutus on suurissa ryhmissä liian helppoa. Oletetaan, että yksilöt elävät ryhmissä ja jokainen auttava teko hyödyttää kaikkia ryhmän jäseniä. Auttava käyttäytyminen voi olla esimerkiksi varoitushuuto, joka varoittaa ryhmän jäseniä lähestyvästä saalistajasta, mutta paljastaa huutajan ja näin lisää tämän mahdollisuutta tulla syödyksi. Oletetaan, että ryhmässä on petkuttaja, joka ei koskaan huuda. Jos vastavuoroiset käyttävät sääntöä, jonka mukaan yhteistyössä ollaan vain kun kaikki ovat yhteistyössä, petkuttaja saa vastavuoroiset lopettamaan yhteistyön. Tällainen petkutus aiheuttaa yhä enemmän petkutuksia. Viattomat yhteistyön tekijät kärsivät yhtä paljon kuin syylliset petkuttajat, kun petkuttamisen ainoa vastatoimi on yhteistyön lopettaminen. Toisaalta jos vastavuoroiset sietävät petkuttajia, niin pitkällä aikavälillä petkuttajat hyötyvät. (Richerson ja Boyd 2006, 250- 251.) Suurimittaista yhteistyötä tukee paitsi yhteistyöstä koituva hyödy, myös ns. moralistinen rankaiseminen ja mukautuva painotus (Richerson ja Boyd 2006, 256.)

Moralistinen rankaiseminen (Trivers 1971) on monimuotoista, esimerkiksi aseman alentumista, vähemmän ystäviä ja vähemmän seurustelutilaisuuksia. Se on kahdesta syystä tehokkaampi kuin vastavuoroisuus suurimittaisen yhteistyön tukemisessa. Ensinnäkin rangaistus voidaan kohdentaa, mikä tarkoittaa sitä, että petkuttajia voidaan rangaista aiheuttamatta laajamittaista petkuttamista, joka seuraa siitä että vastavuoroiset kieltäytyvät yhteistyöstä petkuttajien kanssa. Toiseksi vastavuoroisuudessa rangaistuksen vakavuutta rajoittaa yhden yksilön yhteistyön vaikutus muihin ryhmän jäseniin, ja vaikutus pienenee, kun ryhmä kasvaa. Moralistiset rangaistukset voivat olla paljon kalliimpia petkuttajalle, joten suurissa ryhmissä yhteistyötä tekevät voivat saada muut yhteistyöhön, vaikka heitä itseään on vähän. Pelkurit, karkurit ja petkuttajat voivat joutua aikaisempien tovereidensa hyökkäysten kohteiksi, yhteiskuntansa hyljeksimiksi, juoruilun kohteiksi tai he eivät pääse alueilleen tai saa puolisoita. (Richerson ja Boyd 2006, 251- 252.) Moralistinen rangaistus voi vakiinnuttaa kaikenlaisen mielivaltaisen käyttäytymisen, esimerkiksi solmion pitämisen, eläinten ystävällisen kohtelun tai kuolleiden sukulaisten aivojen syömisen. Merkitystä ei ole sillä, tuottaako käyttäytyminen etuja ryhmälle. Tärkeää on vain, että kun moralistiset rankaisijat ovat yleisiä, rangaistuksi tuleminen maksaa enemmän kuin moitteeton käyttäytyminen, mitä se sitten onkaan. (Richerson ja Boyd 2006,

252, 253.) Jos moralistinen rankaiseminen on yleistä ja rangaistukset riittävän ankaria, yhteistyö kannattaa. Useimmat ihmiset voivat elää koko elämänsä joutumatta rankaisemaan kovin paljon, mikä vuorostaan tarkoittaa että alttius rankaisemiselle voi olla halvempaa verrattuna alttiuteen yhteistyöhön (kun rangaistuksia ei ole). Moralistinen rangaistus voi olla välttämätön kestävässä suurimittaisessa yhteistyössä, mutta se ei ole riittävä selittämään, miksi suurimittaista yhteistyötä ilmenee. (Richerson ja Boyd 2006, 253.) Luonnonvalinta voi suosia psykologista taipumusta jäljitellä yleistä tyyppiä tai menestyviä naapureita. Tämä taipumus on evolutiivinen voima, joka johtaa siihen, että yleiset kulttuuriset muunnokset yleistyvät ja harvinaiset harvinaistuvat entisestään. Jos tämä vaikutus on vahva verrattuna muuttoliikkeeseen, niin ryhmien välinen muuntelu säilyy. Ihmiset omaksuvat vallitsevan uskomuksen, koska siitä saa suurimman tuoton helposti mitattavissa valuutoissa, jotka sisältävät myös rangaistuksi joutumisen hinnan (Richerson ja Boyd 2006, 258- 259, 263, 264.)

Instituutioiden synty

Uusi nopean kulttuurisen adaptaation synnyttämä sosiaalinen ympäristö saattoi ajaa geenievoluution tuottamaan uusia sosiaalisia vaistoja sukuperäämme. Tällaiset ympäristöt suosivat ryhmissä elämiseen sopivia vaistoja kuten psykologiaa, joka on rakentunut oppimaan ja sisäistämään moraalinormeja. Kehittyi uusia tunteita, kuten häpeä ja syyllisyys, jotka lisäävät todennäköisyyttä että normeja seurataan. Yksilöt, joilta puuttui uusia sosiaalisia vaistoja, rikkoivat useammin vallitsevia normeja, ja kokivat epäsuotuisaa valintaa. Heitä on saatettu syrjiä, heiltä on voitu kieltää yleinen hyöty, kuten jaettava ruoka, tai he ovat voineet menettää pisteitä pariutumispelissä. (Richerson ja Boyd 2006b, 16- 18.) Nämä uudet sosiaaliset vaistot ovat kehittyneet ihmispsykologiaan ilman että vanhoja itseä, sukulaisia ja ystäviä suosivia on eliminoitu. Nämä ns. heimovaistot tukevat ryhmäidentifikaatiota ja yhteistyötä suurissa ryhmissä, ja ovat usein ristiriidassa itsekkyyden, nepotismin ja välittömän vastavuoroisuuden kanssa. (Richerson ja Boyd 2006b, 16- 18, Richerson ja Boyd 2006, 268, 269.) Miksi valinnan pitäisi suosia uusia, sosiaalisuutta lisääviä motiiveja? Ihmiset ovat älykkäitä, joten eikö heidän kannattaisi pelkästään laskea paras yhteistyön ja petkutuksen sekoitus ja ottaa rankaisemisen riski? Luultavasti ihmiset eivät ole niin älykkäitä, että

evoluutio luottaisi siihen, että he osaavat tehdä välttämättömät laskelmat oikein. (Richerson ja Boyd 2006, 269, 270.)

Ihmisen yhteiskuntien koon ja monimutkaisuuden kasvuun ei luultavasti ole liittynyt merkittäviä sosiaalisten vaistojemme muutoksia. Vaikka luonnonvalinta voi johtaa merkittäviin geneettisiin muutoksiin vain muutamassa tuhannessa vuodessa, useimmat biologit uskovat, että monimutkaisten piirteiden tärkeät muutokset kestävät paljon kauemmin. Niinpä olemme luultavasti saaneet synnynnäisen sosiaalisen psykologiamme pleistoseenisiltä esi-isiltämme. (Richerson ja Boyd 2006, 289.) Monimutkaisten yhteiskuntien instituutiot ovat selkeästi rakentuneet muinaisille ja heimon vaistoille ja niissä on ennustettavia epätäydellisyyksiä, jotka johtuvat kulttuurievoluution prosesseista. (Richerson ja Boyd 2006, 295- 296.) Niinpä nykyiset yhteiskunnat, joissa esiintyy hierarkiaa, vahvaa johtajuutta, epätasa-arvoisia sosiaalisia suhteita ja äärimäistä työnjakoa suosivia instituutioita, on rakennettu alkujaan heimoyhteiskunnissa elämiseen sopeutuneen sosiaalisen "kieliopin" päälle. Tullakseen toimeen ihmiset rakentavat sosiaalisen maailman, joka muistuttaa sitä jossa sosiaaliset vaistomme ovat kehittyneet. Samalla suuri yhteiskunta ei toimi, elleivät ihmiset pysty käyttäytymään aivan toisin kuin he käyttäytyisivät pienissä heimoyhteiskunnissa. Työ on jaettava hienosyisesti. Kuri on tärkeää ja johtajilla on oltava muodollista valtaa vaatia kuuliaisuutta. Suuret yhteiskunnat edellyttävät toisilleen vieraiden ihmisten rutiininomaista rauhallista vuorovaikutusta. Nämä vaatimukset ovat ristiriidassa muinaisten ja heimotason sosiaalisten vaistojen kanssa ja siksi ne aiheuttavat tunneperäisiä konflikteja, sosiaalisia häiriöitä ja tehottomuutta. (Richerson ja Boyd 2006, 289.)

Ihmisen sosiaaliset vaistot rajoittavat ja painottavat hyvin samalla tavalla yhteiskuntatyyppejä, joita rakennamme, ja ne jättävät tärkeitä yksityiskohtia paikallisen kulttuurisyötteen täytettäväksi. Kun kulttuuriset muuttujat on asetettu, vaistojen ja kulttuurin yhdistelmä tuottaa toimivia sosiaalisia instituutioita. (Richerson ja Boyd 2006, 271.) Suuren yhteiskunnan mahdollistavilla sosiaalisilla innovaatioilla, jotka samanaikaisesti tehokkaasti simuloivat heimoyhteiskunnan elämää, on taipumus levitä. Ihmiset suosivat tällaisia ratkaisuja ja valinnan mahdollisuuden saadessaan omaksuvat ne. Jos yhteiskunnissa on tällaisia instituutioita, niissä

on vähemmän sisäisiä konflikteja ja muun ollessa ennallaan tällaiset yhteiskunnat ovat kilpailussa tehokkaampia kuin muut ryhmät. (Richerson ja Boyd 2006, 290.)

Jotta monimutkainen yhteiskunta toimisi, heimoyhteiskuntien moralistiseen rangaistukseen on liitettävä institutionaalinen pakottaminen. Muuten yksilöt, järjestyneet saalistajajoukot sekä luokat ja kastit, joilla on erikoisvaltuudet pakottamiseen, takavarikoivat kokonaisuudessaan yhteistyön, koordinaation ja työnjaon edut. Institutionaalinen pakottaminen kuitenkin luo tehtäviä, luokkia ja alakulttuureja, joilla on valta kääntää pakottaminen omaksi kapeaksi edukseen. Jonkinlaisten sosiaalisten instituutioiden on valvottava valvojia, jotta he todella toimivat yleisen edun nimissä. Tällainen valvonta ei yleensä ole täydellistä. Eliittien tapa antaa aina itselleen etuja osoittaa, että yksilölliseen itsekkyyteen, sukulaisuuteen ja usein eliitin heimotason solidaarisuuteen perustuvat kapeat edut nostavat ennustettavasti päätään. (Richerson ja Boyd 2006, 290.)

Toimintansa perustasolla instituutiot luovat tunteen, että lait ja tavat ovat reiluja. Järkevästi hallinnoidut byrokratiat, toimivat markkinat, sosiaalisesti hyödyllisten omistusoikeuksien puolustaminen, laaja osallistuminen julkisiin asioihin ja vastaavat auttavat tuottamaan julkisia ja yksityisiä hyödykkeitä tehokkaasti, ja tietyssä määrin suojaavat yksilöllisiä vapauksia ja aliyhteisön autonomiaa. Yleensä instituutioita voidaan pitää toiminnassa niin kauan kuin olemassa olevilla instituutioilla on useimpien yksilöiden mielestä kohtuullinen oikeutus, ja uudistuksia saadaan aikaan tavallisella poliittisella toiminnalla, mm. tarkoituksellisella uusien sosiaalisten instituutioiden kehittämisellä. (Richerson ja Boyd 2006, 294.) Toisaalta monimutkaisten yhteiskuntien kehittyvien instituutioiden monet väistämättömät puutteet johtavat siihen, että oikeutusta on vaikea pitää yllä. Yksilöt, jotka eivät hyväksy nykyisen institutionaalisen järjestyksen oikeutusta, saattavat ryhmittyä vastarintajärjestöiksi, kuten ovat tehneet esimerkiksi nykyiset fundamentalistit ja heimoryhmät, joiden mielestä maallistunut modernismi on laitonta. (Richerson ja Boyd 2006, 294- 295.)

Yhteisöjen tärkeys

Ihmiselle yhteisön on tärkeä ainakin seuraavista syistä:

1) *Mahdollisuus käyttää yhteisön yhteisiä resursseja, energiaa, tavaroita ja tietoa.* Ihmiset syntyvät yhteisöön, joka on rakennettu ennen heitä. Jos ihminen joutuisi elämään ilman tietoa ja tavaroita, jotka yhteisö on tuottanut, hän eläisi esimerkiksi ilman sähköä ja vettä, koska hän ei ole keksinyt sähkön tuotantoa ja käyttöä, eikä rakentanut vedenjakelujärjestelmää. Hänellä ei olisi työkaluja, koneita tai kirjoja, eikä voisi edes tehdä monia työkaluja, koska ei olisi rautaa tai terästä, eikä taitoa tuottaa näitä aineita. Monet edelliset sukupolvet ovat keränneet tietoa ympäristöstä, ja tämä tieto on nykyisen sukupolven käytettävissä. Luultavasti suurin hyöty yhteisöön kuulumisesta on mahdollisuus tämän tiedon käyttämiseen. Koska yhteisön olemassaolon aika on paljon pidempi kuin yksilön, sen keräämän tiedon määrä on merkittävä. (Firtz 2007, The importance of societies.)

2) *Mahdollisuus tehdä yhteistyötä, työn jakaminen*. Jokainen yhteisön jäsenen tarvitsee oppia vain osan olemassa olevasta tiedosta ja voi käyttää oppimiseen vähemmän aikaa, erikoistua, ja olla tuottava suuremman osan elämästään tuottaen itselleen ja toisille yhteisön jäsenille työnsä tuloksen. Työn tulosten vaihdon yhteisö tekee mahdolliseksi standardisoimalla ajan, painon, pituuden jne. mittauksen ja asettamalla rahan. Tämä mahdollistaa yhteisön jäsenien paremman elämänlaadun, kuin mitä he yksinään voisivat saavuttaa. Esimerkiksi suurimmalla osalla ihmisistä ei ole tietoja, joita arkkitehti tarvitsee rakennuksia suunnitellessaan ja rakennusyhtiö taloja rakentaessaan. Niinpä ihmiset antavat niiden tuottaa talonsa sen sijaan että rakentaisivat ne itse. Yhteistyöllä puolustautuminen on helpompaa. (Firtz 2007, The importance of societies.)

Yhteisön käsite

Yhteisö on monimutkainen systeemi. Tällä tarkoitetaan, että sen osat ovat vuorovaikutuksessa keskenään, mistä johtuu, että yhteisön muuttamisen tulos on vaikeasti ennustettavissa. (Fritz 2007, Governing subsociety.) Fritz rakentaa yhteisön käsitteen seuraavalla tavalla:
Yhteisö on systeemi, joten sillä on:

- Raja.
- Ympäristö.
- Osia, jotka vaikuttavat enemmän toisiin yhteisön osiin kuin yhteisön ulkopuolella oleviin.
- Vahvoja kommunikaatioita yhteisössä ja tätä varten tarvittavan median (materian, energian ja informaation vaihto) olemassaolo.
- Rajoitettu olemassaolo ajassa ja tilassa. (Fritz 2007, Concept of a society.)

Yhteisö on älykäs systeemi, joten sillä edellisten lisäksi myös seuraavia ominaisuuksia:

- Se oppii olemassaolonsa aikana.
- Se toimii jatkuvasti saavuttaakseen tavoitteensa ja onnistuu siten tässä useammin kuin satunnaisesti.
- Se käyttää energiaa toimintaan ja sisäisiin prosesseihinsa. (Fritz 2007, Concept of a society.)

Yhteisö on systeemi, joka koostuu monien älykkäiden systeemien aliyhteisöistä, joten sillä on lisäksi seuraavia ominaisuuksia:

- Työnjako.
- Kulutushyödykkeiden tuotanto.
- Kulttuuri (tieto, joka on yhteistä suurimmalle osalle yhteisön jäsenistä).
- Yhteisön tavat.
- Hallinnoiva alisysteemi (engl. governing subsociety), joka johtuu tarpeesta kontrolloida systeemiä.
- Lait (toiminnan ohjeet).
- Puolustus systeemin ulkopuolisten ja sen jäsenten hyökkäyksiä vastaan.
- Uusien jäsenten koulutus. (Fritz 2007, Concept of a society.)

Yhteisön määritelmä voidaan esittää tiiviimmässä muodossa:

Yhteisö:

- On systeemi.
- Koostuu monista osista, joita sanotaan jäseniksi ja jotka ovat älykkäitä systeemejä tai itsekin yhteisöjä, joita tällöin sanotaan aliyhteisöiksi (engl. subsocieties).

- Koska, yhteisöjen perusrakennuspalikka on älykäs systeemi, niin yhteisöillä on ainakin älykkään systeemin ominaisuudet. Yhteisöillä voi olla muitakin ominaisuuksia, koska se koostuu monista älykkäistä systeemeistä.
- Yhteisön tavoitteet ovat ne, jotka ovat sen jäsenille yhteisiä.
- Jäsenten elinaika on yleensä huomattavasti lyhyempi kuin yhteisön tai aliyhteisön, johon he kuuluvat. (Fritz 2007, Concept of a society.)

Joitakin esimerkkejä yhteisöistä ovat: valtiot, provinssit, (engl. municipalities), kaupalliset yritykset, poliittiset puolueet, uskonnolliset organisaatiot ja järjestöt. Myös susilauma ja porolauma koostuvat älykkäistä systeemeistä ja ovat siten tämän määritelmän mukaan yhteisöjä. Muurahaiskeot ja mehiläispesät ovat yhteisöjä jos muurahaiset ja mehiläiset määritellään älykkäiksi systeemeiksi. Yhteisö koostuu monista jäsenistä, miljoonista monissa valtioissa, tai noin viidestäkymmenestä, kuten keskiaikaisessa maatalousyhteisössä. Määritelmä on täydellisesti sovellettavissa yhteisöihin, joissa on esimerkiksi miljoona jäsentä, mutta vain osittain sovellettavissa yhteisöihin, joissa on vähäinen jäsenmäärä, koska esimerkiksi usean tason aliyhteisöjä on olemassa vain yhteisöissä, joissa on hyvin monta jäsentä. (Fritz 2007, Concept of a society.)

Tällaisen systeemiteoreettisen yhteisön käsitteen etuja ovat yksinkertaisuus, täsmällisyys ja muokattavuus. Jos tulevaisuudessa huomataan, että jotain yhteisön tärkeää aspektia ei voida käsitellä nykyisen käsitteen avulla, niin sitä täytyy muuttaa. Käsitteen avulla voidaan tutkia monia kiinnostavia ihmisyhteisöjen aspekteja ja sitä voidaan käyttää myös tietokonemallinnuksen lähtökohtana. (Fritz 2007, Concept of a society.)

Yhteisön ominaisuuksia

- Yhteisön kommunikaatio

Kommunikaatioilla tarkoitetaan materian tai energian vaihtoa universumin eri osien välillä (Fritz 2007, Glossary), ja tämä sisältää myös sen materian tai energian, jota käytetään informaation lähettämiseen. Ajan kuluessa kommunikaatio yhteisössä voi lisääntyä tai

vähentyä. Kommunikaation määrä voidaan laskea, esimerkiksi jos se on informaatiota, megabitteinä, jos se on tavaraa, kiloina, tai jos se on energiaa, kilowattitunteina. Informaatiota voidaan lähettää materian tai energian struktuurina, tai materian tai energian virtauksen vaihteluina. Kommunikaation määrän kehitystä ajan kuluessa voidaan kuvata graafisesti ja sitä voidaan ennustaa esimerkiksi ekstrapoloimalla. (Fritz 2007, "A society is a system. . .") Yhteisöllä on enemmän tai vähemmän kommunikaatioita toisten yhteisöjen kanssa. Yhteisön avoimuutta voidaan mitata seuraavan kaavan avulla: O = Ci/Ce, missä O on systeemin avoimuus. Ci (internal communications) on yhteisön sisäisten kommunikaatioiden summa ja Ce (external communications) yhteisön ulkoisten kommunikaatioiden summa. Jos O on pieni, yhteisöä voidaan sanoa suljetuksi tai lähes suljetuksi ja jos se on suuri, avoimeksi. Luonnollisesti tämä arvo voi vaihdella. (Fritz 2007, "A society is a system. . .")

- Rajat

Jokaisella systeemillä on rajoitettu olemassaolo tilassa ja ajassa, eli raja, joka erottaa sen ympäristöstään. Yhteisöt ovat usein sijoittuneet jollekin maa-alueelle, ja niillä on usein hyvin määritelty raja, ja rajavalvonta. (Firtz 2007, A limited extension in space.) Systeeminä yhteisöllä on myös rajoitettu olemassaolo ajassa. Se syntyy, kasvaa ja tuhoutuu. Jotkin yhteisöt muodostetaan tyydyttämään jotain tiettyä tarvetta. Monet näistä yhteisöistä lakkaavat olemasta, kun niiden tavoite, tämän tarpeen tyydyttäminen on saavutettu. Toisilla yhteisöillä on pitkä olemassaolon aika. Usein yksi tällaisten yhteisöjen tavoite on niiden oma selviytyminen. (Fritz 2007, A limited extension in time.)

- Syntyminen

Fritz ajattelee yhteisön synnyn seuraavasti. Joku älykäs systeemi uskoo, että se voi saavuttaa tavoitteensa paremmin toisten avustuksella kuin yksin ja kommunikoi ajatuksensa toisille älykkäille systeemeille. Tämä kommunikointi voidaan tehdä sanoin tai teoin. Se saa ne toimimaan yhdessä tavoitteen saavuttamiseksi, joten ryhmällä on yhteinen tavoite. Oletetaan vielä, että ryhmän tavoite on pitkäaikainen. Tällöin, koska se on koostunut monista osista jotka ovat älykkäitä systeemejä, ja joilla on yhteisiä tavoitteita ja systeemin olemassaolon aika ylittää sen yksittäisen jäsenen elinajan, sitä voidaan sanoa yhteisöksi edellä annetun määritelmän mukaan. (Fritz 2007, Creation of a society.)

- Kasvaminen

Yhteisöt kasvavat kun niihin liittyy älykkäitä systeemejä, joko vapaaehtoisesti tai pakotetusti. Edellisessä tapauksessa älykkäät systeemit havaitsevat, että yhteisöllä on niiden kanssa samoja tavoitteita ja että tavoitteet saavutetaan, ja niinpä ne liittyvät jäseniksi, jos voivat. Toinen tapa on että yhteisön jäsenet saavat jälkeläisen, ja se liittyy tai liitetään yhteisöön. Yhteisöt voivat myös pakottaa yksittäisiä älykkäitä systeemejä tai niiden yhteisöjä liittymään itseensä, esimerkiksi ottamalla orjia tai valloittamalla heimoja, kansoja tai valtioita. Monesti pakotetusti liitetyillä ei ole kaikkia samoja oikeuksia kuin muilla yhteisön jäsenillä. Tietysti yhteisön jäsenmäärä voi myös vähentyä ajan kuluessa. (Fritz 2007, Growth.)

- Tuhoutuminen

Usein yhteisön tuhoutuminen johtuu sen ulkopuolelta tulevista syistä, esimerkiksi sodasta toisen yhteisön kanssa tai epidemioista. Usein se tuhoaa itsensä populaation kasvaessa niin suureksi, että yhteisön valtaama alue ei riitä ravitsemaan sitä ja käsittelemään sen tuottamaa jätettä. Voi käydä myös niin, että jos riittävän pitkän aikaa yhteisö ei saavuta tavoitteitaan, jotkut sen jäsenet uudelleenorganisoivat sen (vallankumous), tai suurin osa yhteisön jäsenistä lopettaa kuulumisen siihen ja yhteisö hajoaa aliyhteisöihinsä. (Fritz 2007, Destruction of a society.) Joskus osa yhteisöstä eroaa siitä. Oletetaan, että osalla yhteisön jäsenistä on tavoitteita, jotka ajan mittaan alkavat erota muun yhteisön tavoitteista. Voidaan ennustaa, että jossain tulevaisuuden vaiheessa suurin osa tavoitteista, ei enää vastaa muun yhteisön tavoitteita. Silloin seuraa epävakaa tilanne. Yksittäinen tapaus, joka näyttää kaikille, että tavoitteet eivät enää ole samat, voi toimia eroamisen laukaisijana. (Fritz 2007, Destruction of a society.) Syy yhteisön hajoamiseen voi olla jopa sen jäsenten liiallinen hyvinvointi. He tuntevat saavuttaneensa tavoitteensa eivätkä enää tue yhteisöä. Voi olla, että he eivät ymmärrä voivansa säilyttää elämänlaatunsa tason vain yhteisöön kuulumalla. (Fritz 2007, Destruction of a society.) Yhteisö voi hitaasti yhdistyä toiseen. Vähitellen yhteisön jäsenten välinen vuorovaikutus voi vähentyä ja toisen yhteisön osien kanssa lisääntyä. Tietyssä vaiheessa nämä ulkoiset vuorovaikutukset alkavat dominoida ja yhteisö lakkaa olemasta; sen osat ovat nyt toisen yhteisön osia. Esimerkiksi oletetaan vierekkäin nykyaikaisessa mielessä kehittynyt ja kehittymätön yhteisö. Kehittymättömän yhteisön jäsenet näkevät toisen televisio-ohjelmia ja

näkevät sen mainoksia. Hitaasti sen kulttuuri kehittyy samanlaiseksi nykyaikaisen kanssa, ja menettää oman identiteettinsä. Jossain vaiheessa se liittyy toiseen poliittisesti, ja lakkaa kokonaan olemasta. (Fritz 2007, Destruction of a society.)

- Yhteisön jäsenet

Yhteisön osia sanotaan sen jäseniksi. Usein yhteisön jäsenet ovat itsekin yhteisöjä. Tässä tapauksessa yhteisö on yhteisöjen hierarkia. Jokaisella tasolla yliyhteisöksi (engl. super society) sanotaan yhteisöä, jolla on yhteisöjä jäseninään, joita voidaan sanoa niiden aliyhteisöiksi. (Fritz 2007, members.) Yhteisön jäsenellä on enemmän ja tärkeämpiä kommunikaatioita toisten yhteisön jäsenien kuin yhteisön ulkopuolisten kanssa. Voi olla hetkiä, jolloin jäsenellä on enemmän kommunikaatioita yhteisönsä ulkopuolisten kanssa, ja se on siten tänä aikana tämän toisen yhteisön jäsen. Usein on hyödyllistä, että yhteisö voi tunnistaa jäsenensä tehokkaasti ulkopuolisista. Monet ihmisyhteisöt identifioivat jäsenensä jäsenkorteilla, passilla tms. (Fritz 2007, members.) Aliyhteisöjä on ainakin kahdenlaisia: alueellisia ja funktionaalisia, joista edellinen rajoittuu selvimmin tiettyyn alueeseen ja jälkimmäinen tiettyyn tehtävään. Esimerkiksi kunta on alueellinen ja posti funktionaalinen aliyhteisö. Myös aliyhteisön olemassaolon aika on usein sen yksittäisten jäsenten elinajasta riippumatonta. Jotkut jäsenet lähtevät ja toisia tulee, mutta yhteisö pysyy tavoitteensa pysyessä. (Fritz 2007, members.)

- Hallinnoiva aliyhteisö

Koska yhteisön jäsenet ovat älykkäitä systeemejä eivätkä useinkaan ole täydellisesti ohjelmoidut toimintojaan varten, niiden täytyy oppia ennen kuin ne voivat toimia järkevästi. Jäsenten olemassaolon aika on rajallinen, joten on tehokasta että jotkin jäsenet opettelevat joitain taitoja ja toiset toisia. Tämä koskee myös hallinnointia. Yhteisön suorittamista vasteista päättävää aliyhteisöä sanotaan hallinnoivaksi aliyhteisöksi (engl. governing subsociety (GS)). Jokainen jäsen, joka vaikuttaa suorasti tai epäsuorasti yhteisön päätöksentekoprosessiin, on osa hallinnoivaa aliyhteisöä. Jäsenen vaikutuksen voimakkuus voi vaihdella. Esimerkiksi valtiossa hallituksen jäsenillä, kansanedustajilla, uutistoimittajilla, kirjailijoilla ja kansalaisilla on erisuuruiset ja erilaiset vaikutusmahdollisuudet. (Fritz 2007, Governing subsociety.)

Myös yhteisöllä voidaan ajatella olevan älykkyyttä. Yhteisöjen kyseessä ollessa älykkyys voidaan määritellä samalla tavoin kuin yksittäisten älykkäiden systeemien kohdalla. Älykkyyden mitta on se, kuinka hyvin systeemi saavuttaa tavoitteensa. Älykkyys on älykkään systeemin kykyä oppia toimimaan siten, että se saavuttaa tavoitteensa. Tätä määritelmää vastaavasti yhteisöjen kohdalla älykkyyttä mitataan sen perusteella, kuinka tehokkaita hallinnoivan aliyhteisön valitsemat vasteet ovat. Tämä riippuu kahdesta tekijästä: hallinnoivan aliyhteisön jäsenten älykkyydestä ja hallinnoivan aliyhteisön käyttämästä päätöksentekomenetelmästä. Jos hallinnoiva aliyhteisö koostuu yhteisön älykkäämmistä jäsenistä, silloin yhteisön älykkyys on korkeampi kuin sen keskiarvoisen jäsenen. (Fritz 2007, Governing subsociety.)

Hallinnoiva aliyhteisö voi käyttää ainakin kolmea eri päätöksentekomenetelmää:
1) Hallinnoivan aliyhteisön kaikkein vaikutusvaltaisin jäsen valitsee vasteen, jolloin käytetään vain tämän yhden jäsenen tietoa.
2) Kaikki hallinnoivan aliyhteisön jäsenet voivat äänestää vasteesta välittömästi, jolloin käytetään kaikille jäsenille yhteistä tietoa.
3) Päätös perustuu keskusteluun ja suostumukseen, jolloin yksilöiden erityistieto voi vaikuttaa. Hallinnoivan aliyhteisön älykkyys on usein suurempi kuin yksittäisten jäsenten: sen älykkyys on sen jäsenten tiedon summa. (Fritz 2007, Governing subsociety.)

Hallinnoiva aliyhteisö valitsee vasteet sen ja joidenkin sen jäsenten tavoitteiden mukaisesti. Nämä tavoitteet eivät välttämättä ole yhteisön tavoitteita. Hallinnoivan aliyhteisön jokainen jäsen toimii saavuttaakseen oman tavoitteensa (älykkään systeemin määritelmän mukaisesti), joten yhteisö tulee suunnitella sellaiseksi, että hallinnoivan aliyhteisön jäsenelle on henkilökohtaista hyötyä siitä, että se saavuttaa yhteisön tavoitteita. Yhteisön ja sen hallinnoivan aliyhteisön tavoitteiden tulee olla yhteensopivat. Historiassa vain hyvin harvoin hallinnoivan aliyhteisön tavoitteena on ollut yhteisön kaikkien jäsenten hyvinvointi. Yleensä tavoitteena on ollut yhteisön vahvimman aliyhteisön hyvinvointi. Kaikkien yhteisön jäsenten täytyy voida kontrolloida hallinnoivan aliyhteisön vasteita, jotta varmistetaan koko yhteisön tavoitteiden saavuttaminen. (Fritz 2007, Governing subsociety.)

Kuten kaikki älykkäät systeemit, yhteisö valitsee vasteensa saavuttaakseen tavoitteensa tietyssä tilanteessa ja menneisyyden kokemuksiin perustuen. Jos tilanteen arviointi on väärä, tai saadut kokemukset yksipuolisia, valitut vasteet ovat kaukana optimista. Valtiossa hallinnoiva aliyhteisö valitsee vasteet, jotka koskevat sen suhdetta toisiin yhteisöihin, esimerkiksi se voi päättää hyökkäyksestä tai yhteistyöstä. Virheet tilanteen arvioinnissa voivat johtaa esimerkiksi siihen, että yhteisö julistaa sodan toiselle yhteisölle, koska ennustaa itselleen hyötyä tästä. Usein kun hallinnoiva aliyhteisö päättää sodasta, se tuottaa tarvittavan mielialan yhteisön jäsenille (esim. propagandan keinoin) niin että yhteisön jäsenet tukevat sitä ja sotaa. (Fritz 2007, Governing subsociety.)

Yhteisössä on paljon korrelaatioita, mikä tarkoittaa, että sen jäsenten toiminta vaikuttaa toisiin hyvin paljon. Jos jäsenillä on normaalisti tapana hyökätä, he voivat joissakin tapauksissa hyötyä näistä hyökkäyksistä, mutta toisissa tapauksissa kokevat haittaa toisten jäsenten hyökätessä heitä itseään kohtaan. Koska näistä vastavuoroisista hyökkäyksistä on normaalisti pitkällä aikavälillä enemmän haittaa kuin hyötyä, niin yhteisön jäsenet eivät saavuta tavoitteitaan niin hyvin kuin silloin jos hyökkäyksiä ei käytettäisi. Tästä syystä on yhteisön hyödyksi pitää hyökkäyksien määrä niin alhaisena kuin mahdollista. (Fritz 2007, Internal organization.) Suuret yhteisöt estävät sisäisiä hyökkäyksiä luomalla lakeja ja panemalla ne käytäntöön. Tätä varten ne tarvitaan joitakin aliyhteisöjä:
1) Aliyhteisön, joka luo lait ja niiden yksityiskohtaiset kuvaukset.
2) Aliyhteisön, joka dokumentoi rikkomukset ja pidättää niiden tekijät. (poliisi)
3) Aliyhteisö, joka määrittelee rikkomuksesta koituneet haitat, jotka rikkomuksen tehneen täytyy korvata niille, joihin rikkomus on vaikuttanut, ja jos tämä ei ole mahdollista, estää rikkomuksen uusiutumisen pidättämällä tai karkottamalla tekijän.
Lakien ja säädösten lisäksi hallinnoiva aliyhteisö toimii myös suoraan hallinto-organisaatioiden kuten armeijan ja keskuspankin kautta. (Fritz 2007, Internal organization.)

- Yhteisön Tavoitteet

Yhteisö syntyy, kun älykkäillä systeemeillä on yhteisiä tavoitteita, ja näistä tavoitteista tulee yhteisön tavoitteita. Joskus niitä mainitaan esimerkiksi valtioiden perustuslaeissa tai

yhdistysten säännöissä. Yhteisöillä voi myös olla julkilausumattomia tavoitteita, jotka ovat olemassa vain sen jäsenten mielissä. (Fritz 2007, A society is an intelligent system: objectives.)

Yhteisö luodaan siksi, että sen jäsenet voisivat saavuttaa joitakin tavoitteita, omia tai yhteisiä. Yksi yhteisistä tavoitteista on yhteisön selviäminen. Toinen on että kaikki sen jäsenet voisivat saavuttaa omat tavoitteensa mahdollisimman hyvin. Käytännössä tämä tarkoittaa sitä, että jokainen yhteisön jäsen voi toimia vapaasti, kunhan se ei estä yhteisön muita jäseniä tai yhteisöä saavuttamasta tavoitteitaan. (Fritz 2007, Internal organization.) Joillekin jäsenille tulee tilanteita, joissa hänestä näyttää itselleen hyödyllisimmältä, ainakin lyhyellä aikavälillä ja kenties virhearviointiin perustuen tai sitten aivan oikein, suorittaa toimintoja joiden sivutuotteena yhteisön olemassaolo saattaa vaarantua tai häiriintyä. Samalla yhteisön muille jäsenille, jotka eivät ole vastaavassa tilassa, on hyödyllistä suojella yhteisöä. Yhteisön selviytyminen sen jäsenten tavoitteena on ongelmallinen, koska ei ole selvää, että jokaisessa tilanteessa jäsenen etu on sama kuin yhteisön, vaikka ainakin periaatteessa jäsen kuuluu yhteisöön juuri oman etunsa takia. Niinpä yhteisö voi käyttää erilaisia keinoja jäsenten uskollisuuden lisäämiseksi: kasvatusta, uskontoa tai traditiota, propagandaa tai rangaistuksia. Ilmeisesti yhteisön selviytyminen on niin vaikeaa, että ilman sitä, että jäsenillä on tämä tavoite, yhteisö hajoaisi jossakin vaikeassa tilanteessa, niin kuin usein tapahtuu. Joskus jäsenten täytyy tukea yhteisöä, vaikka tästä olisikin niille (ainakin lyhyellä aikavälillä) haittaa.

Kaikkien jäsenten suurimman hyödyn kannalta vaikuttaa järkevältä, että yhteisö tulisi suunnitella siten, että kaikki sen jäsenet voisivat saavuttaa omat tavoitteensa niin paljon kuin mahdollista. Tietysti on selvää, että kaikkia maailman historian aikana olleita yhteisöjä ei ole suunniteltu tällaisiksi. Edellä mainittu vaatimus tarkoittaa myös sitä, että yhteisön täytyy ylläpitää jokaisen jäsenen mahdollisimman suurta vapautta. Kaikkien jäsenten mahdollisimman suuren toiminnan vapauden turvaaminen vaatii joitakin toiminnan vapauden rajoituksia, lakeja. Lait antavat jokaiselle jäsenelle täydellisen vapauden toimia, kunhan se ei vaikuta merkittävästi ja negatiivisesti toisten jäsenten toimintaan, kun he pyrkivät omiin tavoitteisiinsa. (Fritz 2007, A society is an intelligent system: objectives.)

Usein poliittisessa keskustelussa tulee esille eri muodoissa seuraavan kysymys: Kummat ovat tärkeämpiä, yhteisön vai sen jäsenen tavoitteet? Yhteisölle sen tavoitteet ovat tärkeämpiä, ja jokaiselle jäsenelle hänen omansa, mutta jos jäsenet eivät tue yhteisöään, se voi hajota, mikä voi johtaa siihen, että jäsenen on vaikeampi saavuttaa omia tavoitteitaan. Joten eräs yhteisön tärkeä tavoite on luoda olosuhteet joissa sen jäsenet voivat saavuttaa omat tavoitteensa, ja toisaalta jokaisen jäsenen tärkeä tavoite on yhteisön tukeminen. Esimerkiksi jos yhteisö rakennuttaa jäsenen omistaman maan läpi autotien, josta lähes kaikki yhteisön jäsenet hyötyvät mutta maan omistava jäsen kokee haittaa, niin yhteisö korvaa jäsenen kokeman haitan jollakin tavalla. (Fritz 2007, A society is an intelligent system: objectives.)

Ympäristö muuttuu ajan myötä ja siten myös hallinnoivan aliyhteisön jäsenten mielissä oleva nykyisen tilanteen kuvaus. Muuttunut tilanne vaatii yhteisön alitavoitteiden muuttamista sen päätavoitteen, yhteisön selviytymisen saavuttamiseksi. (Fritz 2007, A society is an intelligent system: Change of objectives.)

Tavoitteiden tärkeyttä voidaan mitata aikana, jonka älykäs systeemi on halukas käyttämään tavoitteensa saavuttamiseen. Jotkut yhteisöjen jäsenet elävät mieluummin leväten suhteellisessa mukavuudessa kuin työskentelevät saadakseen paremman elämänlaadun joskus tulevaisuudessa ja toisin päin. Jotkut yhteisöjen jäsenet haluavat mieluummin sotilaallista elämää, johon kuuluu taistelua, kunniaa, kuuluisuutta ja voittoja kuin rauhallista elämää työskennellen, keräten omaisuutta, turvaa ja hyvinvointia (Fritz 2007, A society is an intelligent system: On the selection of objectives.)

Yhteisöt käyttävät eri prosenttiosuuksia resursseistaan koulutukseen. Mikä on tulos kun koulutukseen sijoitetaan paljon resursseja? Ensiksi seuraa joitakin vuosia kestävä alentunut elintaso, koska vähemmän resursseja on käytettävissä. Sitten seuraa korkeampi elintaso, mikä johtuu jäsenten suuremmasta tehokkuudesta kun he nyt samalla määrällä työtunteja voivat tuottaa suuremman määrän resursseja. Korrelaatio koulutuksen ja tuotannon välillä on vahva. Tämä osoittaa, että parempi koulutus tuottaa korkeamman elintason. (Fritz 2007, Assignment of resources.) Fritzin mukaan nykyinen koulutus kaikkialla on vanhentunutta. Suunnilleen samoja aineita opetetaan nyt kuin 100 vuotta sitten ja vaikka sisällöt on uudistettu, tämä ei riitä.

Tulisi selvittää kyselyin ihmisiltä elämänsä keskiosassa, mitkä aineet osoittautuivat heille kaikista eniten ja vähimmin hyödyllisimmiksi elämässään. Sitten aineen opetukseen käytettävissä oleva aika tulisi suhteuttaa sen hyödyllisyyteen. Uusia hyödyllisiä aineita tulee lisätä ja vähiten hyödyllisiä vähentää. Näyttää siltä, että hyödyllisiä aineita olisivat esimerkiksi: 1) Kuinka hankitaan tietoa, 2) Kuinka selvitetään tiedon todenmukaisuus, 3) Kuinka valita hyödyllisiä tavoitteita, 4) Kuinka valita toiminto, 5) Kuinka saavuttaa tavoitteet. (Fritz 2007, The members of a society have a limited life span: Source of knowledge.)

Yhteisöt käyttävät eri prosenttiosuuksia resursseistaan työkalujen valmistamiseen. Mitä tästä seuraa? Jäsen, joka tuottaa työkaluja, ei tuota hyödykkeitä, joten tämän ajan hänen tuotantonsa ja siten yhteisön tuotanto on alhaisempaa, mutta työkalun tehtyään hän voi tuottaa sen avulla paljon enemmän. Ehkä hän tuottaa samassa ajassa enemmän, vaikka tähän lasketaan mukaan työkalun valmistamiseen käytetty aika. Ylimääräisen ajan hän voi käyttää useampien työkalujen suunnitteluun ja valmistamiseen. Tämä on kiihtyvä prosessi; hän tuottaa enemmän ja enemmän työkaluja ja tuottaa enemmän ja enemmän hyödykkeitä. (Fritz 2007, Assignment of resources.) Näin selittyy teknologisen edistyksen tehokkuutta lisäävä vaikutus.

Yhteisöt käyttävät eri prosenttiosuuksia resursseistaan terveydenhuoltoon. Miten tämä vaikuttaa? Resurssien käyttö alentaa elintasoa. Toisaalta terveydenhuolto johtaa jäsenten parempaan terveyteen ja pidempään elinaikaan. Terveemmät jäsenet voivat tuottaa enemmän, mikä voi korvata terveydenhoitoon sijoitetut resurssit ja johtaa korkeampaan elintasoon. Kuitenkin he elävät kauemmin, vähemmän tuottava tai tuottamaton aika elämän loppuosassa pitenee, mikä vähentää tuottavuutta. Ennen suurempi osa jäsenistä kuoli geneettisistä syistä johtuviin sairauksiin. Paremman terveydenhoidon avulla he voivat elää ja saada enemmän jälkeläisiä. Jotkut jälkeläisistä perivät geneettisen vian. Niinpä terveydenhoidon kustannukset kasvavat. Ellei yhteisö keksi keinoja, joilla geneettisten vikojen leviäminen tuleviin sukupolviin estetään, paremmasta terveydenhoidosta voi siis johtua, että yhteisön jäsenten keskiarvoinen terveystaso laskee. (Fritz 2007, Assignment of resources.) Toisaalta riittävän kehittynyt geenimanipulaatio voi korjata tätä kehitystä.

Yhteisön tieto, tavat ja kulttuuri

Kun älykäs systeemi toistaa usein toimintojen sarjan, koska se on ollut hyödyllinen, se muodostaa näin yhdistetyn toiminnon, jota voidaan sanoa tavaksi. Kun tietyt tavat yhteisössä ovat eniten käytössä, yhteisöä voidaan kuvailla niiden mukaisesti. Suuri kommunikaation määrä yhteisössä johtaa sen jäsenten yhteiseen tietoon ja tapoihin, jota voidaan sanoa yhteisön kulttuuriksi. Nykyisen kaltainen maailmanlaajuisen kommunikaation lisääntyminen voi johtaa yhteiseen maailmanlaajuiseen kulttuuriin muutaman sukupolven kuluessa. (Fritz 2007, members.)

Jotkin tavat ovat hyödyllisiä ja jotkin haitallisia yhteisölle tai aliyhteisölle. Hyödyllisenä tapana voidaan pitää esimerkiksi sitä, että ihmiset työskentelevät yhdessä toisiaan auttaen ja haitallisena esimerkiksi sitä, että epäeettinen toiminta annetaan anteeksi usein. Tämä johtaa epäeettisen toiminnan lisääntymiseen. Sosiaalinen oppiminen mahdollistaa ihmispopulaatioiden adaptiivisen informaation kasaamisen monien sukupolvien ajan, mikä voi johtaa hyvin adaptiivisten käyttäytymisten ja teknologian kulttuurievoluutioon. Koska tämä prosessi on paljon geenievoluutiota nopeampaa, ihmispopulaatioille voi kehittyä kulttuurisia adaptaatioita paikallisiin ympäristöihin, mikä on erittäin hyödyllistä pleistoseenikauden kaoottiseen, nopeasti muuttuvaan maailmaan. Kuitenkin sama psykologinen mekanismi joka mahdollistaa tämän hyödyn aiheuttaa sen, että ihmiset ovat herkkäuskoisia ja useimmiten pitävät yhteisönsä tapoja järkevinä ja oikeina, mikä mahdollistaa maladaptiivisten käyttäytymisten leviämisen. (Richerson ja Boyd 2006b, 16.)

On keksittävä keinoja, joilla kulttuurisia maladaptaatioita voidaan tunnistaa ja poistaa. Tähän ei nykyään ole mitään systematisoitua keinoa, vaan se tapahtuu mm. kulttuurikritiikin, huumorin, taiteen, sananvapauden, demokratian ja kautta.

4 YHTEENVETO

Tutkimuksessa on
4.1 Määritelty, mitä maailmankatsomuksilla tarkoitetaan
4.2 Miten niitä rakennetaan
4.3 Miten niitä arvioidaan
4.4 Esitetty mahdollisimman hyvä maailmankatsomus.

Tutkimuksen tavoite on parantaa mahdollisuuksia luoda mahdollisimman hyvä maailmankatsomus. Voidaksemme luoda maailmankatsomuksia, oli tarpeellista ensin määritellä, mitä maailmankatsomuksilla tarkoitetaan, sekä selvittää, miten niitä rakennetaan ja arvioidaan. Tämän jälkeen esitin maailmankatsomuksen, joka nykyään parhaiten soveltuu tarpeisiimme. Esitän yhteenvedon ja joitain huomioita joka kohdasta.

4.1 Maailmankatsomuksen määritelmä

Maailmankatsomuksen käsite (alun perin Immanuel Kantin käyttämä *Weltanschauung* (Kant 1987, 111- 112)) on määritelty monella eri tavalla (katso: Naugle 2002). Yleensä maailmankatsomuksia pidetään enemmän tai vähemmän yhtenäisinä ajatuskokonaisuuksina, joiden avulla ihmiset ovat yrittäneet ymmärtää itseään ja ympäristöään. Täysin tyydyttävää tai yhtä yleisesti hyväksyttyä määritelmää ei ole olemassa. Yleisimmät puutteet ovat epätäsmällisyys, ei-tieteellisyys, määrittelyalueen kapeus ja liian vähäinen yksityiskohtaisuus.

Maailmankatsomuksen käsitteelle on luotava täsmällinen, tieteellinen ja yleinen määritelmä, koska nämä ominaisuudet lisäävät käsitteet hyödyllisyyttä ja käyttökelpoisuutta. Ihmisen, kuten muidenkin tunnettujen agenttien älyllinen kapasiteetti on rajallinen, joten liian monimutkaiset määritelmät ovat käyttökelvottomia, ja epätäsmälliset eli liian monitulkintaiset, sumeat tai ristiriitaiset antavat liian monia tai ei yhtään mahdollista tulkintaa. Mitä yleisempi määritelmä on, sitä useammassa tapauksessa käsitettä voidaan käyttää ja sen tarjoamaa hyötyä, mm. laskennallisten resurssien säästöä tapahtuu. Määritelmien tulee olla yksinkertaisia, jotta

niihin perustuva päättely olisi helppoa. Niiden tulee olla täsmällisiä ja ilmaistu aikaisemmin määritellyillä sanoilla, jotta päättely olisi eksaktia, ja jotta määritellyt käsitteet muodostaisivat yhtenäisen järjestelmän. Määritelmien tulee olla käyttökelpoisia, jotta niistä voidaan tehdä kiinnostavia ja käyttökelpoisia johtopäätöksiä. Mielestäni uskottava maailmankatsomuksen määritelmä voidaan tehdä vain perustamalla se tieteellisiin käsitteisiin. Tämä johtuu siitä, että pidän tieteellisen menetelmän avulla hankittua tietoa luotettavimpana. Tiede ei ole ennen ollut tarpeeksi edistynyt tarjotakseen tarvittavia käsitteitä, mutta vaikuttaa siltä, ns. agentti-ympäristö -viitekehyksen sekä tekoälytutkimuksen käsitteistön avulla tieteellisiin käsitteisiin perustuva maailmankatsomuksen määritelmä on mahdollinen. Määritelmä voi myöhemmin avata mahdollisuuden tutkia maailmankatsomuksia tieteellisesti.

Tutkimuksen lähtökohtana on ns. agentti-ympäristö -viitekehys. Tekoälytutkimuksessa sitä käytetään mm. vahvistusoppimisen viitekehyksenä (Sutton, R. ja Barto, A. 1998), ja eri tavalla nimettynä kontrolliteoriassa (Bertsekas, D. P. ja Tsitsiklis, J. N. 1996). Viitekehys tuli tunnetuksi, kun Norbert Wiener (1894- 1964) kollegoidensa Arturo Rosenbluethin ja Julia Bigelowin kanssa haastoivat behavioristisen ajattelutavan tarkastellen käyttäytymistä nykyisen ja tavoitetilan välistä virhettä säätelevänä systeeminä. Wienerin kirjasta Kybernetiikka (1948) tuli bestseller, joka myös kiinnitti yleisön huomion mahdollisuuteen luoda älykkäitä koneita (Russel ja Norvig 2003, 15). Viitekehys sisältää kolme olennaista komponenttia: agentin, ympäristön ja tavoitteen. Agentti on mikä tahansa, mitä voidaan tarkastella sensoreillaan ympäristöään havainnoivana ja siinä aktuaattoreillaan toimivana (Russell ja Norvig 2003, 32). Agentin ja ympäristön tulee voida olla vuorovaikutuksessa toistensa kanssa. Agentin tulee voida lähettää signaaleja ympäristöön ja ottaa vastaan ympäristöstä tulevia signaaleja ja ympäristön tulee voida saada signaaleja ja lähettää niitä agentille. Agentin ympäristöön lähettämiä signaaleja voidaan sanoa toiminnoiksi (engl. actions), ja sieltä vastaanottamia havainnoiksi (engl. perceptions). (Legg ja Hutter 2007b, 15- 16.) Agentilla täytyy olla ainakin yksi tavoite. (Legg ja Hutter 2007b, 15- 16.) Tavoite on tietty tilanne, johon se pyrkii. (Fritz 2007, What is intelligence?, Acting on the environment, Objectives.) Agentti voi olla älykäs ilman tavoitetta, jonka saavuttamiseen se voi käyttää älykkyyttään, ja siinäkin tapauksessa, että agentti ei halua käyttää älyään tavalla joka vaikuttaa sen ympäristöön, mutta kummassakaan tapauksessa sen älykkyyttä ei voida havaita. Älykkyys voidaan havaita kun

sillä on tavoite, jota se aktiivisesti ympäristöönsä vaikuttamalla yrittää saavuttaa. (Legg ja Hutter 2007b, 15- 16.) Ilmeisesti mikä tahansa peli, haaste, ongelma tai testi voidaan ilmaista tämän yksinkertaisen viitekehyksen avulla ilman suurta vaivaa. Viitekehys ei sano mitään siitä, miten agentti tai ympäristö todella toimii, vaan kuvaa niiden roolit. (Legg ja Hutter 2007b, 17.) Määritelmä esitetään formalisoidussa muodossa lähteessä: (Legg ja Hutter 2007b, 17- 24). Käytän viitekehystä, koska se on yleinen, eri aloilla menestyksellä käytetty, sekä täsmällisesti ja formaalisti määriteltävissä. Siten se on käyttökelpoinen perusta, jolle on suhteellisen turvallista rakentaa uusia käsitejärjestelmiä. Käsitykseni maailmankatsomuksista, niiden rakentamisesta ja arvioinnista sekä ehdotus nykyiseksi maailmankatsomukseksi perustuvat tähän viitekehykseen.

Tämän viitekehyksen näkökulmasta maailmankatsomus liittyy siihen, miten agentti päättää toimia ympäristöstään saamansa informaation ja tavoitteidensa perusteella. Maailmankatsomusta voidaan pitää agentille kuuluvana informaatiojärjestelmänä, jonka tehtävänä on tuottaa oikea vaste syötteeseen. Oikea vaste määräytyy viime kädessä agentin tavoitteen, resurssien ja ympäristön mukaan. Oikea vaste on se, joka parhaiten auttaa agenttia saavuttamaan tavoitteensa. Tietty tavoite, tietyt resurssit ja tietty ympäristö muodostavat mahdollisuudet ja ehdot menestyksekkäälle käyttäytymiselle. Sopiva maailmankatsomus on funktio, jonka lähtöjoukkoon tavoite, resurssit ja ympäristö kuuluvat, ja jonka maalijoukkoon kuuluu menestyksekäs käyttäytyminen. Tässä merkityksessä kaikki agentit tarvitsevat maailmankatsomuksen vuorovaikutukseen ympäristönsä kanssa.

Ensimmäiset tunnetut maailmankatsomukset ovat olleet biologisilla organismeilla. Ympäristön mallin rakentamisen välttämättömyys johtuu elävän organismin tarpeesta sopeutua siihen. Niinpä se johtaa myös maailmankatsomuksen ja sen kehittämisen tarpeisiin (Vidal 2007, 9). Tämän mallin rakenteeseen vaikuttavat sen tekijän tarpeet ja kyvyt (Aerts, Apostel ja muut 2007, 18- 19.) sekä ympäristö. Myös mm. robotit ja ohjelma-agentit tarvitsevat maailmankatsomuksia, ja niiden rakentamisesta ja ohjelmoinnista saadut kokemukset ovat hyödyllisiä maailmankatsomusten tutkimukselle, koska niitä voidaan muutella, niiden osat tunnetaan ja rakenne voidaan pitää yksinkertaisena. Nykyään ihmisten kannalta tärkeä tarve on kehittää maailmankatsomuksia laajan yhteistyön mahdollistamiseksi. Yksilöiden ja yhteisöjen

elämänsuunnitelmat tai toimintasuunnitelmat edellyttävät maailmankatsomusta. Hyvien suunnitelmien tarve lisääntyy ympäristön monimutkaisuuden kasvaessa. Kulttuurienvälinen vuorovaikutus, ihmiskunnan suurempi yhdistyminen sekä tieteen ja teknologian edistyminen johtavat siihen, että yksilön elämänsuunnitelma ja aliyhteisöjen suunnitelmat ovat voimakkaammin yhteydessä globaaliin kokonaisuuteen. Elämänsuunnitelman muodostaminen on samalla tullut vaikeammaksi huomioonotettavan kokonaisuuden koon ja monimutkaisuuden lisääntymisen vuoksi. Esimerkiksi ihmiskunnan selviytymiseen tällä planeetalla liittyvät ekologiset ongelmat tulevat enenevästi olemaan jokaisen ongelma. (Aerts, Apostel et al. 2007, 8, 11.) Vastuullinen toiminta, ympäristön muuttaminen sekä yhteistyö edellyttävät maailmankatsomusta. Saadaksemme tietoa itsestämme, tarpeistamme, päämääristämme ja arvoistamme, rakennamme malleja ympäristöstämme ja itsestämme. Meidän täytyy rakentaa malleja ihmisistä, historiasta, arvoista ja toimintastrategioista ja yhdistää kosmosta, maata ja biosfääriä koskevan tietomme kanssa. Ilman näiden käsitysten integraatiota vastuullinen toiminta näyttää olevan mahdotonta. Koska emme voi vain antaa asioiden kehittyä omalla painollaan, vaan on otettava vastuu maailmastamme, uusi näiden elementtien integraatio on välttämätön. Tämän yrityksen tulee olla kollektiivinen, koordinoitu ja tietoinen. Eettisiä ja poliittisia valintoja valaiseva tieto ihmisistä, luonnosta, historiasta ja yhteiskunnista mahdollistaa kenties tulevaisuudessa ihmisten ottaa kohtalonsa omiin käsiinsä. (Aerts, Apostel et al. 2007, 11.) Tarvitaan viitekehys, joka sitoo kaiken yhteen ja mahdollistaa yhteiskunnan, maailman ja ihmisten paikan maailmassa ymmärtämisen, koska se auttaisi ihmisiä ja ihmiskuntaa tekemään tärkeitä päätöksiä, jotka muovaavat tulevaisuuttamme. (Heylighen 2000: "What is a world view?") Maailmankatsomuksen kehittäminen voi parantaa elämänhallintaa, ja mahdollistaa uusia kykyjä ja kokemuksia.

Mitkä ovat maailmankatsomusten välttämättömät ominaisuudet? Toisin sanoin: Mitä toimintoja agentin kognitio tarvitsee, jotta se voi muodostaa vasteita ympäristöstään saamansa kommunikaation perusteella? Clement Vidal (2007), Aerts ja kumppanit (2007) sekä Francis Heylighen (2000) tarjoavat uskottavan maailmankatsomuksen määritelmän. Heidän mukaansa kokonaiseen maailmankatsomukseen kuuluu ainakin seuraavat kuusi osaa: mallit maailmasta, menneisyydestä ja tulevaisuudesta sekä teoriat tiedosta, arvosta ja toiminnasta. Ne vastaavat osia koskeviin kysymyksiin. Osat täytyy yhdistää koherentiksi ja käyttökelpoiseksi

kokonaisuudeksi. Kaikkein yksinkertaisimmilla agenteilla osat eivät ole välttämättä kovin kehittyneitä.

1) Malli maailmasta/ympäristöstä (Ontologia): Mitä on olemassa? Miksi on olemassa jotain? Millainen maailma/universumi on? Mikä on maailman rakenne ja kuinka se toimii? (Vidal 2007, 8.), (Aerts, Apostel ja muut. 2007, 14.), (Heylighen 2000, What is a worldview?)

2) Malli menneisyydestä (Historia): Mistä kaikki on tullut? Miksi maailma on sellainen kuin on? Miksi olemme tällaisia kuin olemme emmekä erilaisia? Mikä on universumin alkuperä? Millaisia yleisiä selittäviä periaatteita voidaan käyttää?(Vidal 2007, 8.) Maailmankatsomuksen selitysvoima on sitä suurempi, mitä yleisempiä ja käyttökelpoisempia lakeja tai säännönmukaisuuksia se löytää todellisuudesta. Käytännössä selitys tarkoittaa vähemmän itsestään selvien tosiasioiden johtamista yleisistä ja hyväksytyistä laeista. (Aerts, Apostel et al. 2007, 14- 15.)

3) Malli tulevaisuudesta (Futurologia): Minkälaisia tulevaisuuksia me ja lajimme voi saavuttaa? Millä kriteereillä valitsemme näistä mahdollisista tulevaisuuksista? Tulevaisuuden kehitystä koskevan mallin tulisi antaa lista enemmän tai vähemmän todennäköisistä tulevaisuuden kehityksistä. Näistä voitaisiin sitten valita, mihin tulisi pyrkiä ja mitä tulisi välttää. (Heylighen 2000, What is a worldview?). Kuinka kulttuurit vuorovaikuttavat tulevaisuudessa? Mikä on tieteen ja talouden rooli tulevaisuudessa? Mikä taho tekee tai tulee tekemään ihmiskunnan tulevaisuuteen vaikuttavia päätöksiä? Leviävätkö ihmiset maapallon ulkopuolelle tulevaisuudessa? (Aerts, Apostel et al. 2007, 17.)

4) Teoria tiedosta (Epistemologia): Mitä tieto on ja kuinka sitä voidaan hankkia? Millä tavalla voidaan rakentaa käsitys maailmasta, jotta voitaisiin vastata maailmankatsomuksen osia koskeviin kysymyksiin? Kuinka voidaan hankkia tietoa? Mikä on totta ja mikä ei? (Vidal 2007, 9.)

5) Teoria arvoista (Aksiologia): Mikä on arvokasta ja miten se saavutetaan? Mihin meidän tulisi pyrkiä? Mitä on hyvä ja paha, tai hyvä ja huono? Mikä on elämän tarkoitus? Miksi

tunnemme niin kuin tunnemme ja miten suhtaudumme todellisuuteen ja mikä on roolimme siinä? (Vidal 2007, 8.) Miten muodostetaan moraali tai etiikka, sääntöjen joukko, joka kertoo millä tavalla tulisi tai ei tulisi käyttäytyä. (Heylighen 2000, What is a worldview?) Onko/voidaanko muodostaa arvojen ja päämäärien hierarkiaa? Onko yleisiä arvoja tietyille agenttien joukoille kuten kaikille ihmisille? Miten muodostetaan etiikka, tiettyjen agenttien vuorovaikutusta koskevien sääntöjen joukko, joka auttaa toteuttamaan tietyt arvot/tavoitteet?

6) Teoria toiminnasta (Prakseologia): Millä tavalla toimimalla tavoitteet voidaan saavuttaa? Mitkä ovat yleiset periaatteet, joiden mukaan toiminta pitäisi organisoida? Tällainen tieto auttaa suunnitelmien toteuttamista käytännön ongelmien ratkaisemiseksi. (Vidal 2007, 8.) Sopeutua voi monella tavalla. Voidaan muuttua itse tai muuttaa ympäristöä. Ihmislajin tyypillinen piirre on muuttaa ympäristöä toteuttamaan ihmisten tavoitteita. Maailmankatsomuksen tulee sisältää organisoitu näkemys tosiasiallisista ja mahdollisista vaikutuksista, joita ihmisillä voi olla ympäristöönsä. Esimerkiksi jos halutaan rakentaa globaali maailmankatsomus, sille on hyväksi yleinen prakseologia, yleinen päätöksenteko- ja ongelmanratkaisuteoria ja strategisen tutkimuksen ja suunnittelun yksikkö, jotta sen toiminta olisi mahdollisimman tehokasta. (Aerts, Apostel et al. 2007, 19- 20.) Prakseologia kertoo millä tavalla toimimalla tavoitteet voidaan saavuttaa. (Heylighen 2000, What is a worldview?)

Tekoälytutkimuksen käsitteistön avulla voidaan ymmärtää, miten maailmankatsomus liittyy kognitiiviseen arkkitehtuuriin. Käsitteistö on täsmällinen, tieteellinen, formalisoitu ja yleinen (toisin kuin psykologian, joka keskittyy tiettyihin toteutuksiin, yleensä ihmisten tai eläinten aivoihin), joten se on mahdollisimman hyvä tämän tutkielman tavoitteiden kannalta. Tekoälyn käsitteistöllä maailmankatsomuksen osat ovat kognitiivisen arkkitehtuurin osia. Kognitiiviset arkkitehtuurit ovat agentin arkkitehtuurien osajoukko. Ne ovat agentin kognitiossaan tarvitsemien prosessien vaatimia rakenteita. Maailmankatsomuksen osat ovat laskennallisia osarakenteita, ja osaprosesseja joita agentin kognitiivinen koneisto käy läpi kun se yrittää saavuttaa tavoitteensa.

1) Malli maailmasta (ympäristöstä) ja 2) malli menneisyydestä (Historia) muodostavat agentin tietokannan. Agentin täytyy kyetä erottelemaan ympäristöstä tulevia relevantteja

kommunikaatioita epärelevanteista. Toisin sanoin agentilla täytyy olla ontologia. Ontologia voi joko olla agentilla valmiina tai se voi rakentaa sitä saamiensa kokemusten perusteella. Agentin perseptisekvenssissä voi jo sinällään olla hyödyllistä tietoa. Kaikkia agentin kokemuksia ei yleensä voida säilyttää muistitilan puutteen vuoksi. Siksi niillä täytyy olla jokin mekanismi, jolla muistojen säilytyksen pituus päätetään. Lisäksi monet agentit kykenevät etsimään persptisekvenssistään säännönmukaisuuksia, jotka mahdollistavat käsitteiden ja niistä ontologian rakentamisen. Malli maailmasta on ontologia, jonka avulla agentti hahmottaa ympäristöään. Ontologia tarkoittaa teoriaa siitä, mitä ja millaisia asioita on olemassa (Russell ja Norvig 2003, 261). Ontologia on käsitteistö, johon kaikki ympäristön kohteet luokitellaan. Ontologiat koostuvat kategorioista, jotka ovat tietyn ontologian kannalta olennaisia käsitteitä. Ontologiat voidaan jakaa erityisiin ja yleisiin ontologioihin. Erityiset ontologiat (engl. special-purpose ontology) ovat vain tietyn, esimerkiksi tietyn tehtävän suorittamiseen tarvittavan tiedon esittämiseen luotuja käsitteellisiä viitekehyksiä, kun taas yleiset ontologiat (engl. general-purpose ontology) sopivat minkä tahansa tiedon esittämiseen. Kaksi tunnuspiirrettä erottaa yleisiä ontologioita erityisistä. Yleisen ontologian pitää soveltua käytettäväksi enemmän tai vähemmän millä tahansa erityisalueella. Toiseksi millä tahansa riittävän vaativalla määrittelyalueella (engl. domain), eri tiedon alueet on esitettävä yhtenäisesti, koska järkeilyyn ja ongelmanratkaisuun saatetaan käyttää monen eri alueen tietoa samanaikaisesti. (Russell ja Norvig 2003, 320, 321, 322.) Ontologian tehtävänä on mahdollistaa kohteiden luokittelu, jonka ansiosta niiden ominaisuuksia voidaan päätellä. Melkein kaikilla yleistyksillä on poikkeuksia. Esimerkiksi vaikka "tomaatit ovat punaisia" on hyödyllinen sääntö, jotkut tomaatit ovat vihreitä, keltaisia tai oransseja. Se, miten poikkeuksia tai epävarmaa tietoa käsitellään, on hyvin tärkeää. Relevanttien kohteiden luonne, agentin resurssit sekä se, millaisia suunnitelmia halutaan luoda (tavoitteet), määräävät sen, millainen menestystä edistävän ontologian täytyy olla. Juuri tämän enempää ei voida sanoa yleisellä tasolla.

3) Malli tulevaisuudesta (Futurologia) ja 6) teoria toiminnasta (Prakseologia) ovat suunnitelmia, joita agentti muodostaa tietokannan ja päättelyn avulla, tavoitteensa mukaisesti. Tavoitteen saavuttamiseen tarkoitetun toimintosekvenssin luomista sanotaan suunnitteluksi. (Russell ja Norvig 2003, 375.) Suunnittelun lähtökohta on maailman tila toimintosarjan alussa, jota sanotaan alkutilaksi tai lähtötilaksi. Tämä on tieto siitä, mitkä ongelman kannalta

relevantit väitteet ovat tosia. Tavoitteet ja alkutila muodostavat yhdessä ongelman. Lisäksi suunnittelijalla on käytössään joukko toimenpiteitä eli toimintoja, joilla se voi muuttaa maailman tilaa tavoitteiden saavuttamiseksi. (Karanta 1993, 183.) Suunnittelutehtävän mielekkyyden edellytyksenä pidetään yleensä, että: 1) Maailma, jossa suunnittelu tapahtuu, on ainakin jossain määrin ennustettavissa. Kaoottisissa tai hyvin satunnaisissa ympäristöissä suunnittelu on tehotonta, ja agentti voi vain reagoida tapahtumiin. Lopputulos on sama, jos ympäristö muuttuu niin nopeasti, ettei käytettävissä oleva laskentakapasiteetti riitä suunnitelmien tekemiseen ja muuttamiseen tarpeeksi nopeasti. 2) Käsiteltävä ongelma on mielekkäästi paloiteltavissa lähes erillisiin aliongelmiin, joilla on vain vähän vuorovaikutusta keskenään. (Karanta 1993, 183, 184.) Korkeammilla luonnollisilla älykkäillä systeemeillä ja monilla keinotekoisilla älykkäillä systeemeillä on mielikuvitus. Mielikuvitus on työkalu, jolla voi kokeilla turvallisesti eri toimintoja ennen toimintojen käyttämistä ympäristössä. Mielikuvitus tarkoittaa kykyä esittää nykyinen tilanne mielessä, käyttää siihen sopivaa vastesääntöä ja esittää näin syntyvä uusi nykyinen tilanne. Älykäs systeemi toistaa tätä prosessia, kunnes tavoitetilanne on saavutettu. (Fritz 2007, Mental methods and chains of response rules: habits.)

4) Teoria tiedosta (Epistemologia) sisältää keinot joilla agentti voi hankkia tietoa ympäristöstään. Tieto voidaan määritellä informaatioksi, jonka avulla voidaan tehdä oikeaan osuvia ennusteita. Yleensä agenteilla on jonkinlainen tietokanta, jossa on agentin tieto. Useissa tapauksissa tietokantaan voidaan lisätä tai ottaa pois tietoa. Agentin toteutuksessa täytyy ratkaista mm. se, millä tavalla tietoa hankitaan, millä periaatteilla perseptisekvenssistä etsitään säännönmukaisuuksia, mitkä muistot säilytetään. Nämä seikat vaikuttavat tiedon määrään ja esitysmuotoon, jotka taas vaikuttavat agentin menestykseen. Agentin epistemologian täytyy olla yhteensopiva sen resursseihin, ympäristöön ja tavoitteisiin.

5) Teoria arvoista (Aksiologia) sisältää agentin tavoitteen ja menestyksen määritelmän (ja mitan). Agentilla täytyy olla jokin mekanismi, jolla se voi päättää, onko tietty tilanne tavoitetila. Kaikilla älykkäillä systeemeillä on tavoite (engl. main objective). Monet voivat myös oppia luomaan ja käyttämään alitavoitteita (engl. sub objective). Alitavoitteet ovat alemman tason ja/tai väliaikaisia tavoitteita. Alitavoitteen saavuttamalla älykäs systeemi

lähenee tavoitteensa saavuttamista. (Fritz 2007, Objectives.) Tavoite on usein määritelty ns. menestyksen kriteeristöllä (engl. performance criteria) jonka täyttämällä agentti saavuttaa tavoitteensa. Suoritusarvo (engl. performance measure) kuvaa, kuinka hyvin toiminta täyttää menestyksen kriteerit. Tavoitteen saavuttaakseen agentti suorittaa toimintoja usein ympäristöstään saamansa palautteen perusteella. Menestyksellinen toiminta vastaa ympäristön tilan muuttumista tavoitetilaksi toiminnan avulla. Ei ole olemassa yhtä ja pysyvää menestyksen kriteeristöä, joka sopisi kaikille agenteille, vaan agenteilla on eri tavoitteita, ja menestys määräytyy niiden saavuttamisen mukaan. (Russell ja Norvig 2003, 35.) Menestyksen tunnistamisen menetelmät eivät useinkaan ole täydelliset. Tämä tieto on välttämätöntä, jotta se voisi tunnistaa menestyksen ja muokata käyttäytymistään sen perusteella. (Russell ja Norvig 2003, 33- 35.) Tavoitetesti-funktio (engl. goal test) ratkaisee, onko tietty tila tavoitetila. Joskus mahdollisten tavoitetilojen joukko voidaan esittää eksplisiittisesti, jolloin tavoitetesti yksinkertaisesti selvittää onko annettu tila yksi niistä. Joskus taas tavoite on määritelty abstraktien ominaisuuksien avulla. (Russell ja Norvig 2003, 62.)

Määritelmä on ns. stipulatiivinen; se esittää uuden merkityksen jo käytössä olevalle termille, toisin kuin ns. deskriptiivinen määritelmä, joka ilmaisee termin yleisesti käytössä olevaa merkitystä. Stipulatiivinen määritelmä ei sitoudu aikaisemmin käytössä olleeseen merkitykseen, vaan termi määritellään kuin se parhaaksi nähdään (Gupta 2008). Määritelmän hyödyllisyys on kuitenkin hyvä perustella tai osoittaa, jotta muutkin haluaisivat käyttää sitä.

Yleisen määritelmäni mukaan maailmankatsomus on kokonaiskäsitys (teknisesti: informaatiosysteemi) sen omaajan ympäristöstä, tavoitteista ja menetelmistä tavoitteiden saavuttamiseksi. Jokaisella agentilla ja siksi ainakin useimmilla elävillä oliolla ja joillakin elottomilla, kuten roboteilla, on maailmankatsomus, vähintään implisiittisenä ja tiedostamattomana. Yksityiskohtaisempi määritelmä on esitetty tekstin kappaleessa (2.1 Maailmankatsomukset).

Määritelmä tarjoaa hyötyjä, joita ei muissa määritelmissä ole:

1) Yleisyys. Yleisyyden hyödyt ovat selvät: Mitä yleisempi käsite, sitä useammassa tapauksessa sitä voidaan käyttää. Tietenkään pelkkä yleisyys ei riitä siihen että käsite olisi hyödyllinen. Käsitteen yleistyksessä tavoitteena on epäolennaisten asioiden huomiotta jättäminen. Ainakin kaksi epäolennaista komponenttia löytyy usein maailmankatsomuksen määritelmistä.

Monien määritelmien mukaan maailmankatsomus on ihmisen tietoisen ajatustyön tulosta (mm. Ketonen 1981). Nykyään tietoisuuden luonnetta ei tunneta tarpeeksi tiedostaen ja tiedostamatta suoritetun ajattelun erottamiseksi. Uskon että suuri osa ihmisten maailmankatsomuksista - niistäkin joiden väitetään olevan tietoisesti rakennettuja - on rakentunut tiedostamatta. Suuri osa ajattelua on tiedostamatonta. Tietysti maailmankatsomukset, jotka on saatu ilmaistua esimerkiksi kirjallisesti, ovat ainakin kirjoitusvaiheessa luultavasti olleet tiedostettuja. Joka tapauksessa tietoisuuden vaatimus on mielestäni turha ja ongelmallinen. Jos maailmankatsomuksen käsite varattaisiin vain tietoisille olioille, täytyisi keksiä lisäksi jokin käsite kaikkia muita agentteja - joilla kaikilla näyttää olevan jotain hyvin samankaltaista - varten. Kaikilla agenteilla täytyy olla jokin keino hahmottaa ympäristöään ja joissakin kin tapauksissa jossakin määrin itseään, ja jäsentää ympäristöstä tulevia kommunikaatioita. Niillä täytyy olla jokin tavoite ja keino keksiä ja päättää, miten ne toimivat ympäristössään. Tätä voidaan mielestäni kutsua maailmankatsomukseksi, se on kaikilla agenteilla, ja tämä käsitys on mielestäni hyvin intuitiivinen.

Joidenkin määritelmien mukaan maailmankatsomuksia voisi olla vain ihmisillä. Tämä on vanha ajattelutapa, jolle ei nykytiedon valossa ole mitään erityistä syytä. Maailmankatsomus on älykkään käyttäytymisen edellytys. Nykyään ajatellaan, että ihminen ei ole ainoa älykäs olio. Kaikki eliöt ovat jossakin määrin älykkäitä, koska ne hyödyntävät muistia ja ennustamista lisääntyäkseen tehokkaammin. "Kaikki käyttäytyminen, olipa se sitten ihmisen, etanan, yksisoluisen eliön tai puun käyttäytymistä, on keino hyödyntää maailman struktuuria lisääntymiseen." (Hawkins 2005, 181- 182.) Biologiset oliot eivät ole ainoita, jotka voivat käyttäytyä älykkäästi, eikä lisääntyminen ole ainoa mahdollinen tavoite. Universaalin älykkyyden määritelmän mukaan älykkyyden määrä kuvaa agentin kykyä saavuttaa

tavoitteitaan eri ympäristöissä. Kyky oppia ja ymmärtää, ratkaista ongelmia, suunnitella jne. sisältyvät implisiittisesti tähän määritelmään. (Legg ja Hutter 2007a, 12, Legg ja Hutter 2006, 73- 80.) Nykyään on käytössä paljon koneita, jotka tämän määritelmän mukaan ovat älykkäitä. Tässä tutkielmassa esitettyjen älykkyyden ja maailmankatsomuksen määritelmien mukaan maailmankatsomuksia on hyvin monilla erilaisilla olioilla, koska maailmankatsomus on älykkään käyttäytymisen edellytys.

2) Täsmällisyys ja yksityiskohtaisuus. Käsitteet joita määrittelyssä usein käytetään, eivät ole täsmällisiä. Niinpä niiden perusteella on vaikea luoda täsmällisiä määritelmiä. Määritelmien komponentteina käytetään usein käsitteitä kuten; tietoisuus mm. Ketosella (Ketonen 1981, 2.), näkökulma ja kulttuuri mm. Nietzschellä (Naugle 2002, 101- 102, 106.), absoluuttinen henki Hegelillä (Naugle 2002, 68- 73.), joita ei ole täsmällisesti eikä yksityiskohtaisesti määritelty. Tekoälytutkimuksessa käytetty käsitteistö on täsmällistä ja yksityiskohtaista. Maailmankatsomuksen osien tarvitsemat laskennalliset struktuurit ja prosessit on määritelty siinä täsmällisesti ja yksityiskohtaisesti (kts. kpl. 3.2.3). Jos maailmankatsomus määritellään näitä komponentteja käyttäen ja jos komponenttien suhteet on selkeästi ilmaistu, tulos on täsmällinen ja yksityiskohtainen.

3) Formaalisuus. Tässä tutkielmassa ei saavuteta täysin formaalista määritelmää. Agentti-ympäristö -viitekehys on formaalisti määritelty lähteessä: (Legg ja Hutter 2007b, 17- 24). Maailmankatsomuksen välttämättömät ja riittävät osat esitetään tässä tekstissä ja lähteessä (Vidal 2007, 8-9). Osien yhteys kognitiiviseen arkkitehtuuriin esitetään tässä tekstissä. Maailmankatsomuksen osat ja siten maailmankatsomus koostuu kognitiivisen arkkitehtuurin osista. Nämä osat voidaan erottaa niiden tehtävien perusteella, jotka ovat kognitiivisen arkkitehtuurin tehtävän vaatimia osaprosesseja. Osat voidaan esittää tekoälytutkimuksen kielellä ja ne ovat formaalisti määriteltävissä. Formaalisuudesta on hyötyä mm. koska se usein vähentää erilaisten tulkintojen mahdollisuutta, mikä selkeyttää määritelmän käyttöä tutkimuksessa ja kommunikaatiossa ja helpottaa tiedon yhdistämistä.

4) Tieteellisyys. Tekoälytutkimuksen käsitteistöä voidaan pitää tieteellisenä. Tämä tarkoittaa, että niitä käytetään teorioissa ja malleissa, jotka ovat testattavia. Käsitteiden tieteellisyys, jos

niitä käyttävät teoriat ja mallit ovat saaneet empiiristä tukea, kertoo, että ne ainakin jollakin tavalla ovat käyttökelpoisia, ja lisää niiden uskottavuutta verrattuna muihin vähemmän tukea saaneisiin käsitteisiin. Käsitteiden tieteellisyys voi myös parantaa niiden integroimista muuhun tieteelliseen tietoon, sekä edesauttaa niiden tutkimusta.

4.2 Maailmankatsomusten rakentaminen

Usein agentit käyttävät maailmankatsomustaan sitä elämänsä aikana muuttamatta, mutta jotkut voivat muuttaa sitä spontaanisti tai suunnitelmallisesti. Yhden asian tai ilmiön muutosta sanotaan sen kehittymiseksi, ja jos se sen ansiosta myös auttaa enemmän jonkin tavoitteen saavuttamisessa, voidaan puhua tältä kannalta edistymisestä. Maailmankatsomuksen muutos on usein osittainen ja koskee useimmiten ympäristön mallia. Ympäristön muutoksesta tai samanlaisena pysyvää ympäristöä koskevien uusien havaintojen perusteella agentti voi muodostaa mielessään uusia käsitteitä ja vastesääntöjä, jolloin maailmankatsomus muuttuu joiltain osin. Yleensä sitä yritetään muuttaa paremmaksi ympäristöön sopeutumisen eli siinä agentin tavoitteiden saavuttamisen kannalta, jolloin maailmankatsomus edistyy. Usein jotkin osat pysyvät muuttumattomina koko elämän ajan. Eliöillä perimmäiset tavoitteet, selviytyminen ja lisääntyminen eivät luultavasti voi muuttua. Toisaalta esimerkiksi robotti voidaan ohjelmoida kokonaan uudestaan perimmäisiä tavoitteita myöten.

Suuri osa maailmankatsomuksista ihmisilläkin rakentuu osittain spontaanisti (eli siihen tietoisesti puuttumatta). Kun aivomme ovat kehittyneet riittävästi, saamme kokemuksia ympäristöstä, havaitsemme suhteita, kategorisoimme, erottelemme ja yleistämme sitä, mitä aistimme havaitsevat. Korvaamme aistikokemukset ja muistot abstrakteilla yleistetyillä käsitteillä ja niiden rakennelmilla. Sovitamme monia käsitteitä yhtenäisiin skeemoihin, ja rakennamme näistä skeemoista käsitteellisiä viitekehyksiä. Uusien käsitteiden muodostaminen ja käsitteellisten viitekehysten muovautuminen hidastuu ikääntymisen myötä (Project Worldview 2007, Worldviews-An Introduction), mutta monissa tapauksissa jatkuu koko elämän ajan. Elämänsä aikana ihminen rakentaa hitaasti elementaarikäsitteistä alkaen korkeamman tason käsitteitä ja vastesääntöjä. Vastesäännöt muodostavat hierarkian konkreeteimmista abstrakteimpiin, jotka vaikuttavat hyvin moniin vasteisiin. (Fritz 2007,

Mindscapes.) Käyttäjä testaa ja kehittää sitä jatkuvasti ympäristöstään saamansa palautteen perusteella. Yksilö- ja aliyhteisökohtaisten kokemusten ohella monet tieteen, tekniikan ja etiikan löydöt ja keksinnöt sekä yhteiskuntien ja luonnon tapahtumat sekä kulloinkin vallitseva kulttuuri (yhteisön jäsenten yhteinen tietämys ja tavat (Fritz 2007, members)) vaikuttavat ihmisten maailmankatsomusten muotoutumiseen

Maailmankatsomuksia voidaan myös rakentaa suunnitelmallisesti. Suunnitelma edellyttää jotain tavoitetta, jonka perusteella maailmankatsomuksen suoritusarvoa voidaan arvioida tai mitata. Jotkin maailmankatsomusten rakentamiseen erikoistuneet yksilöt tai ryhmät voivat tarjota maailmankatsomuksia halukkaille. Mahdollisesti rakennettavien maailmankatsomusten laatua voidaan ja tulee arvioida ja valvoa joidenkin haittojen vähentämiseksi. Tämä on kuitenkin avoin kysymys.

Kenen tai minkä tahon tehtävänä on rakentaa maailmankatsomuksia? Periaatteessa kuka tahansa kykenevä voi sen tehdä. Mille taholle tämä tehtävä on ensisijainen, ja/tai mikä tuottaa parhaan tuloksen? Tiede ja filosofia ovat hyviä ehdokkaita, koska ihanteellisessa tapauksessa niissä ylläpidetään kokeellisen tutkimuksen ja objektiivisen, sekä hyvin järjestetyn tiedon ihanteita. Nämä ihanteet eivät ole satunnaisesti eivätkä tietyn edun takia valittuja, vaan perustuvat niiden avulla tuotetun tiedon ylivertaisen tehokkaaseen ja laajaan käyttökelpoisuuteen. Tällainen tieto auttaa mahdollisimman hyvin pääsemään tavoitteeseen ja on monikäyttöistä.

Tieteen ja filosofian tavoite on sama: hankkia tietoa. Maailmankatsomusten kannalta tieto on tärkeää, koska sen avulla voidaan rakentaa parempia maailmankatsomuksia. Perimmältään tietoa hankitaan jotta voitaisiin tehdä parempia ennusteita. Ennusteita tehdään jotta voitaisiin selviytyä, ja ylipäänsä menestyä (saavuttaa tavoitteet). Tiedon pätevyyden kannalta varmin tapa hankkia tietoa on kokeellinen tutkimus, koska tällöin tieto on jo aluksi ainakin jossain yhteydessä havaintoaineistoon. Aina ja kaikissa tapauksissa tämä ei ole mahdollista vaan täytyy turvautua muihin menetelmiin. Toiseksi paras tapa on kokeellisen tutkimuksen tuloksiin ja hallittuun ajatteluun (joihinkin ajattelua koskeviin sääntöihin, esim. johonkin monista nykyisistä logiikoista) perustuva spekulaatio. Suuri osa filosofiasta käyttää tätä

menetelmää. Kolmanneksi paras on hallittu ajattelu, joka ei perustu kokeellisen tutkimuksen tuloksiin. Usein filosofit sortuvat tähän. Huonoin on perusteettomien ja koettelemattomien uskomusten ylläpitäminen. Uskonto kuuluu tähän. Tästäkin voi joskus olla hyötyä, erityisesti jos uskomukset ovat jostain syystä sopivia ja ympäristö ei muutu.

Tällä hetkellä ei ole tieteenalaa, jonka tehtävänä on muodostaa maailmankatsomusten kaltaisia laajoja tietojärjestelmiä, mutta asiantila voi muuttua ajan kuluessa. Suurta osaa maailmankatsomuksen rakentamiseen tarvittavasta tiedosta ei tällä hetkellä voida hankkia pelkästään tieteellisen tutkimuksen menetelmällä, mikä kuitenkin on ihanne, ja joudutaan käyttämään myös toiseksi parasta eli filosofian menetelmää. Siksi nykyään maailmankatsomusten rakentaminen on ennen kaikkea tieteen ja filosofian yhteinen tehtävä. Jos tieteilijät ja filosofit eivät rakenna maailmankatsomuksia, muut kulttuurin tekijät ottavat hyödyn tästä tilanteesta ja tuottavat sekä tarjoavat omia vastauksiaan. Näitä voivat olla esimerkiksi uskonnot, vaarallisemmassa muodossa kultit, äärimmäisiä mielipiteitä kannattavat ideologiat tai fundamentalistisia tulkintoja uskonnoista levittävät tahot (Vidal 2007, 9.)

Eri aikoina kokonaisten maailmankatsomusten luomista pidetään eri määrässä mahdollisena ja mahdottomana ja ylipäänsä haluttavana. Yrittäjiä näyttää kuitenkin riittävän, joten voidaan olettaa, että tälle on olemassa tarve. Voidaan olettaa, että kaikki agentit haluavat menestyä, ja sopiva maailmankatsomus on tässä avuksi, koska se tuotttaa hyödyllisiä ja käyttökelpoisia ratkaisuja niiden ongelmiin. Kokonaiset ja yhtenäiset tietojärjestelmät ovat usein sekä monikäyttöisempiä että tehokkaampia (tiiviimmin ilmaistavissa, helpommin käytettävissä) kuin hajanaiset. Yhtenäinen kokonaiskäsitys on hyödyllinen tavoite. Vaikka emme tällä hetkellä tiedä, missä määrässä se on mahdollista, on hyvä tavoitella sitä ja jatkaa tätä projektia mahdollisuuksien mukaan. Tiedon yhtenäistyminen on joka tapauksessa hyödyllistä riippumatta siitä, saadaanko kokonaiskäsitystä koskaan valmiiksi.

Luultavasti tiede jo olemuksestaan johtuen lähestyy yhtenäistä kaikenkattavaa mallia tai teoriaa maailmasta, riippumatta siitä, tavoitellaanko tätä. Muutamat tieteen ihanteet tai usein hyvänä pidetyt ominaisuudet ohjaavat käytännössä tieteellisen tiedon esitystä tiiviiseen ja yhtenäiseen muotoon. Eri variaatioissaan usein ajatellaan, että teoria T2 on parempi kuin T1,

jos siitä voidaan päätellä useampia päteviä havaintolauseita (Lakatos ja Musgrave 1970, Kuipers 2000). Tämä tavoite tuntuu johtavan laajempiin teorioihin ja koska pyritään yhtenäisyyteen, niin myös laajempiin tietokantoihin. Tieteellinen selittäminen on eri ilmiöiden välisten yhteyksien kuvailua. Menestyksekkäät selitykset paljastavat ennen erillisinä pidettyjen ilmiöiden välisen yhteyden. Tiede edistää ymmärrystämme luonnosta näyttämällä, kuinka johtaa monien ilmiöiden kuvauksia käyttämällä samoja päättelysääntöjä uudestaan ja uudestaan, ja vähentää näin perustavimmiksi hyväksyttävien faktojen määrää (Kitcher 1989, 423). Joidenkin mukaan tieteen edistys tarkoittaa lisääntyvää ongelmanratkaisua ja ilmiöiden kontrollia tieteen mahdollistamien sovellusten kautta (Recher 1977). Usein ongelmien ratkaisemiseksi tarvitaan tiedon yhdistämistä. Tiede edistyy ja karkeasti tämä voidaan todeta siitä, että sen avulla voidaan nykyään tehdä enemmän asioita kuin ennen. Tieteessä jos teoriat ovat havaintoaineiston alideterminoimia, tällöin usein kehoitetaan valitsemaan yksinkertaisin evidenssin kanssa yhteensopiva teoria (Foster ja Martin 1966). Yksinkertaisuus voi olla paitsi esteettinen kriteeri, myös kognitiivinen; se voi auttaa ymmärtämään maailmaa kognitiivisesti ekonomisesti (Niiniluoto 2008). Käsitykset tieteen yhtenäisyydestä ovat merkityksellisiä tieteessä ja filosofiassa. Tieteessä ne tuottavat heuristista ja metodologisia opastusta ja oikeutusta hypoteeseille, projekteille, ja tavoitteille sekä rahoitukselle ja koulutukselle. Filosofiassa oletukset yhdistämisestä auttavat valitsemaan kysymyksiä ja tutkimusaiheita. (Cat 2008.)

Yleinen ohje maailmankatsomuksen rakentamiseen:
1. Tee synopsis kaikesta, mikä saattaa olla hyödyllistä maailmankatsomuksen kysymyksiin vastaamiseen.
2. Valitse valmiita tai tee itse parhaat käsitteet synteesin luomiseksi tästä synopsiksesta.
3. Ehdota synteesiä systemaattisen filosofian muodossa.
4. Vertaile tuloksena olevaa maailmankatsomusta toisiin maailmankatsomuksiin edellä esitettyjen arviointikriteerien avulla, ja osoita, miksi se on parempi kuin muut.
5. Osoita, kuinka se voi ratkaista aikamme ongelmat.
6. Levitä tätä maailmankatsomusta. (Vidal 2007, 26.)

4.3 Maailmankatsomusten arviointi

Monista eri maailmankatsomuksista voidaan valita paras tai parhaat erilaisten kriteerien perusteella. Kohdetta voidaan arvioida vain jonkin kriteerin tai kriteeristön perusteella, ja jos halutaan arvioida maailmankatsomuksia, on siis valittava arvioinnin kriteeristö, joka on myös perusteltava. On esitettävä, mitä kriteereillä yritetään saavuttaa, miten ne sen tekevät, ja miksi ja kenelle tämä on tavoittelemisen arvoista.

Mihin maailmankatsomusten arviointikriteerit perimmiltään ja yksityisistä toteutuksista riippumatta perustuvat? Maailmankatsomukset ovat agenttien informaatiosysteemejä, jolla ne käsittelevät ympäristöstään saamaansa tietoa ja yrittävät muodostaa sen ja tavoitteensa perusteella toimintoja joiden suorittaminen auttaa niitä saavuttamaan perimmäisen tavoitteensa. Jos maailmankatsomusten tehtävä on tällainen, maailmankatsomusten hyvyyttä voidaan arvioida sen perusteella, täyttävätkö ne tämän tehtävän, eli auttavatko ne agenttia saavuttamaan tavoitteensa. Siten tavoitteellisen agentin kannalta maailmankatsomus on hyvä, jos se auttaa sitä saavuttamaan perimmäisen tavoitteen. Maailmankatsomus on paras, jos se tekee tämän paremmin kuin muut, ja sen hyvyys on yhtä kuin sen hyödyllisyys käyttäjälleen.

Lisää kriteerejä voidaan muodostaa sillä perusteella, että niiden täyttäminen auttaa maailmankatsomuksen perimmäisen tavoitteen saavuttamisessa. Agentin menestyksen kannalta tärkeitä tekijöitä ovat sen resurssit, ympäristö ja tavoitteet. Ilman tietoa näistä tekijöistä, oikeaa toimintoa ei ole mahdollista määrittää. Jos ei tiedetä agentin resursseja, ei tiedetä, mitä toimintoja agentti voi suorittaa, mitä se voi havaita ja miten paljon ja miten monimutkaista tietoa se voi käsitellä. Ilman tietoa ympäristöstä ei tiedetä, mihin tilaan ympäristö siirtyy tietyn toiminnon suorittamisen seurauksena, joten ei voida selvittää, mikä toiminto siirtää ympäristön tavoitetilaan. Ilman tietoa tavoitteista ei voida tietää, mikä ympäristön tila on tavoitetila, joten sitä ei voida löytääkään. Yleisesti agentin resursseja ovat kaikki sen tavoitteen saavuttamisen kannalta hyödylliset asiat. Agentin resurssit voidaan luokitella sisäisiin, välittäviin ja ulkoisiin resursseihin. Sisäiset resurssit ovat hyödyllisiä agentin sisäisiä rakenteita tai prosesseja. Välittävät resurssit ovat agentin ja ympäristön kommunikaatioita välittävien agentin osien (sensorien ja aktuaattorien) hyödyllisiä kykyjä.

Ulkoiset resurssit ovat hyödyllisiä ympäristön osia. Agentin maailmankatsomus kuuluu sen sisäisiin resursseihin. Parempi maailmankatsomus on suurempi/arvokkaampi resurssi. Mitkä maailmankatsomuksen ominaisuudet tekevät hyödyllisemmän, eli paremman/arvokkaamman resurssin? Useimmiten nämä:

- Yksinkertaisuus/lyhyys säästää laskennallisia resursseja, kuten prosessointitehoa, ja muistitilaa.
- Sisäinen, ja ulkoinen koherenssi mahdollistavat päättelyn, suunnittelun ym.
- Kehityskelpoisuus/muokattavuus mahdollistaa oppimisen, eli maailmankatsomuksen osien muokkaamisen uusien kokemusten perusteella.

Lisäksi voi olla lukematon määrä erilaisia tapauskohtaisia kriteerejä, jotka määräytyvät tavoitteen, käytössä olevien resurssien sekä ympäristön ominaisuuksien mukaan. Maailmankatsomuksen hyvyys tai arvo on usein tapauskohtaista ja siten sitä voidaan luonnehtia vain hyvin rajoitetussa määrin. Tämä johtuu siitä, että voi olla olemassa hyvin suuri joukko hyvin erilaisia tavoitteita, resursseja ja ympäristöjä. Ehkä voitaisiin jakaa arvo yleiseen ja tapauskohtaiseen, jolloin yleinen arvo viittaisi niihin ominaisuuksiin, jotka ovat kaikissa tapauksissa (tai ainakin suuressa osassa tapauksista) hyödyllisiä, ja tapauskohtainen vain tietyissä. Useimmiten on ilmeisesti käytännössä kokeiltava, millainen toimii milloinkin.

Tämä herättää myös ajatuksen, että maailmankatsomuksia voidaan optimoida. Maailmankatsomuksen osat ovat kognitiivisia osaprosesseja. Ne kuluttavat laskennallisia resursseja: aikaa, muistitilaa, huomiota, ja voivat olla virheellisiä tai sopimattomia.

Jotkut älykkäät systeemit, ihmiset ainakin jossain määrin, voivat suunnitelmallisesti muuttaa, kehitellä ja rakentaa maailmankatsomuksia. Kommunikaation avulla niitä voidaan ainakin osittain ja jossakin määrin virheettömästi levittää myös toisille. Parempien maailmankatsomusten tuottaminen on järkevä tavoite. Huonompien haittoja tulee jossakin määrin valvoa ja yrittää poistaa.

4.4 Mahdollisimman hyvä maailmankatsomus

Älykkäät systeemit -maailmankatsomus perustuu Walter Fritzin kirjassaan Intelligent systems and their societies (2007) esittämiin ajatuksiin. Sen käsitteet tulevat eri tieteenaloilta, pääasiassa tekoälystä ja systeemiteoriasta. Lisäksi olen kerännyt vakuuttavia ja yhteensopivia ajatuksia muista lähteistä. Sen hyvyys perustuu se omaamiin maailmankatsomuksille toivottaviin ominaisuuksii. Se on tarkoitettu nykyisille ihmisille. Toisin kuin monet muut maailmankatsomukset, se vastaa kaikkiin maailmankatsomuksen osia koskeviin kysymyksiin, ei kuitenkaan täydellisesti tai lopullisesti.

Älykkäät systeemit -maailmankatsomus on kokonainen (vastaa kaikkiin maailmankatsomuksen osia koskeviin kysymyksiin), koherentti (tieteellinen, systeemiteoreettinen kieli), ja kehityskelpoinen (avoin ehkä kieltä, tutkimuksen menetelmää sekä joitakin peruskäsitteitä lukuun ottamatta ja uuden tiedon hankintaa hyödyllisenä pitävä).

Älykkäät systeemit -maailmankatsomuksen hyvyys perustuu sen omaamiin maailmankatsomuksille toivottaviin ominaisuuksiin. Toisin kuin monet muut maailmankatsomukset, se vastaa kaikkiin maailmankatsomuksen osia koskeviin kysymyksiin, ei kuitenkaan täydellisesti tai lopullisesti. Se on kokonainen (vastaa kaikkiin maailmankatsomuksen osia koskeviin kysymyksiin), koherentti (tieteellinen, systeemiteoreettinen kieli), ja kehityskelpoinen (avoin ehkä kieltä, tutkimuksen menetelmää sekä joitakin peruskäsitteitä lukuun ottamatta ja uuden tiedon hankintaa hyödyllisenä pitävä). Se on hyvä peruslähtökohta, jota voidaan täydentää ja tarkentaa tutkimuksen edetessä. Maailmankatsomusten osia koskeviin kysymyksiin se vastaa karkeasti kuvailtuna seuraavalla tavalla:

1) Malli maailmasta (ympäristöstä). (Ontologia), 2) malli menneisyydestä. (Historia), ja 3) malli tulevaisuudesta. (Futurologia): Tieteellinen tutkimus ja sen menetelmä tarjoavat luotettavimman tiedon ympäristöstä, menneisyydestä ja tulevaisuudesta. Universumi on kaikki mitä on olemassa, kokonaisuutena tarkasteltuna (Fritz 2007, Glossary). Universumissa on tai voidaan havaita säännönmukaisuuksia, korrelaatioita, joiden perusteella sitä voidaan jakaa

osiin. Agentti tai älykäs systeemi on osa universumia. Enemmän tai vahvempia korrelaatioita on systeemin eri osien välillä, kuin systeemin osien ja systeemin ulkopuolisten universumien osien välillä. (Fritz 2007, Glossary.) Systeemin ympäristöä on se osa universumista, joka on kommunikaatiossa systeemin kanssa, mutta ei ole osa sitä (Fritz 2007, Glossary). Kommunikaatio on materian tai energian liikettä kahden universumin osan välillä (Fritz 2007, Glossary). Agentti havainnoi ympäristöään, eli vastaanottaa ympäristöstä tulevia kommunikaatioita. Sen toimintaperiaatteena on toistaa syöte-vaste -sykliä: Se saa ympäristöstään syötteen, valitsee siihen (tavoitteensa kannalta) sopivan vasteen, ja suorittaa sen. Vaste yleensä muuttaa ympäristön tilaa, ja systeemi saa uuden syötteen, ja sykli alkaa alusta. Kaikkien älykkäiden systeemien toimintaperiaate on sama, vaikkakin se voidaan käytännössä toteuttaa monin eri tavoin.

4) Teoria tiedosta. (Epistemologia): Älykkäillä systeemeillä on erilaiset kyvyt ja keinot hankkia tietoa. Ne saavat ympäristöstä kommunikaatioita eri energiamuotoja tunnistavien aistinelintensä eli sensoriensa kautta. Niillä on usein hyvin rajallinen määrä sensoreita, ja nekin ovat yleensä rajoittuneita vain tietyn tyyppisten kommunikaatioiden vastaanottamiseen, koska rajalliset resurssit kannattaa käyttää vain varmasti relevantin tiedon hankkimiseen ja käsittelyyn. Nämä tosiasiat rajoittavat älykkään systeemin vastaanottamien kommunikaatioiden tyyppejä ja määrää ja siten myös informaatiota ja tietoa, joka sillä voi ympäristöstään olla. Tietoa voidaan saada havainnoista, tai sitten päättelemällä havaintojen perusteella. Tieto perustuu muistiin ja ennustamiseen: aikaisempien kokemuksien perusteella muodostetaan ennusteita tulevaisuuden kehityksistä. Keinotekoisiin älykkäisiin systeemeihin tietoa voidaan myös ohjelmoida etukäteen. Myös perimää voidaan pitää tietona. Luonnollisten älykkäiden systeemien DNA toimii muistina, joka sisältää niille elintärkeää tietoa, joka johtaa niiden rakenteen, kykyjen ja taipumusten kehittymiseen (Hawkins 2005, 186).

5) Teoria arvoista. (Aksiologia): Kaikilla älykkäillä systeemeillä on tavoite (engl. main objective). Monet voivat myös oppia luomaan ja käyttämään alitavoitteita (engl. sub objective). Alitavoitteet ovat alemman tason ja/tai väliaikaisia tavoitteita. Alitavoitteen saavuttamalla älykäs systeemi lähenee tavoitteensa saavuttamista. (Fritz 2007, Objectives.) Tarkoitushakuisuus tai tavoitteellisuus syntyi elämän mukana noin neljä miljardia vuotta sitten.

Evoluutio kasvatti tästä pienestä alusta yhä selvempiä merkityksiä ja yhä selvempiä arvoja preferenssijärjestelmän, jonka mukaan eliöt panevat asioita ja elämyksiä arvojärjestykseen. (Salmi 2002, 54- 55.) Luonnollisten älykkäiden systeemien ympäristöjen vaarat ovat aiheuttaneet sen, että niiden tavoitteeksi on tullut elossa pysyminen ja lisääntyminen (Fritz 2007, Main objective). Yksilöt, joilla on tavoitteita, ominaisuuksia tai tapoja, jotka eivät edesauta tai haittaavat selviytymistä, menehtyvät jonakin vaikeana hetkenä. Paremmin selviytymiseen kykenevillä yksilöillä on keskiarvoltaan enemmän jälkeläisiä, ja siten suurempi osa seuraavasta sukupolvesta on niitä. (Fritz 2007, The objectives of humans.) Vaihtoehtojen lisäys eliöiden aivojen kapasiteetin kasvaessa koskee ennen muuta keinojen valintaa, ei niinkään suuressa määrin perimmäisiä tavoitteita. Keinotekoisten älykkäiden systeemien tavoitteet voivat olla mitä tahansa mitä ohjelmoija ohjelmoi niihin. Ihmistä palvelevan robotin tavoitteena voi olla esimerkiksi ihmisen hyväksymisen maksimointi ja moitteen minimointi. (Fritz 2007, Main objectives.)

6) Teoria toiminnasta. (Prakseologia): Älykäs systeemi oppii käyttäytymään niin, että se saavuttaa tavoitteensa: Se aistii ympäristöään, ja kokemustaan käyttäen valitsee toiminnon ja suorittaa sen. (Fritz 2007, Create a scientific ethics.) Seuraava ohjaava periaate pätee kaikkiin älykkäisiin systeemeihin, ihmiset mukaan lukien:

- Älykäs systeemi toimii aina saavuttaakseen tavoitteensa. (Fritz 2007, Interaction between intelligent systems.)

Tämä koskee myös tilanteita, joissa se on vuorovaikutuksessa muiden älykkäiden systeemien kanssa. Ainakaan nykyään ei ole tietoa, onko olemassa esim. universaalia, transsendentaalia, absoluuttista, ikuista, jne. hyvää ja miten siitä voidaan tietää. Älykäs systeemi oppii "Hyvän" käsitteen, kuten muutkin käsitteet kokemustensa kautta. Ympäristössään toimiessaan se saa kokemuksia ja vähitellen oppii, mikä auttaa sitä saavuttamaan tavoitteensa.
Hyvän määritelmä on seuraava:

- Älykkäälle systeemille se, mikä auttaa sitä saavuttamaan tavoitteensa, on hyvää, ja se, mikä estää, on pahaa.

Moniagenttisissa ympäristöissä tilanne on usein monimutkaisempi kuin yksiagenttisissa ympäristöissä. Moniagenttisissa ympäristöissä toiset älykkäät systeemit saattavat reagoida sen toimintoihin, ja älykkään systeemin täytyy ottaa tämä huomioon arvioidessaan oman

toimintansa hyödyllisyyttä. Etiikkaa voidaan tarkastella älykkäiden systeemien näkökulmasta seuraavasti: Koska ihminen on älykäs systeemi, ja voidaan yrittää selvittää, mitä mahdollisimman älykäs systeemi tekisi tietyssä tilanteessa, tämä on se miten ihmisenkin tulisi toimia. (Fritz 2007, Extensive summary.) Kun älykäs systeemi toimii hyödyllisesti (tavalla mikä auttaa sitä saavuttamaan tavoitteensa) moniagenttisissa ympäristöissä, Fritz sanoo sen toimintaa eettiseksi. Toisaalta löysemmin myös parasta tapaa pyrkiä edellä mainittuun tulokseen voidaan sanoa eettiseksi toiminnaksi.

Moniagenttisissa ympäristöissä toiset älykkäät systeemit saattavat reagoida sen toimintoihin, ja älykkään systeemin täytyy ottaa tämä huomioon arvioidessaan oman toimintansa hyödyllisyyttä. Hyödyllisten reaktioiden todennäköisyyden lisäämiseksi ja haitallisten vähentämiseksi moniagenttisissa ympäristöissä älykkäiden systeemien kannattaa käyttäytyä seuraavan periaatteen ohjaamana:
Älykkään systeemin kannattaa käyttää toimintoja, jotka
1) auttavat sitä saavuttamaan tavoitteensa, ja
2) ovat samalla enemmän hyödyksi kuin haitaksi kaikille, joihin toiminta vaikuttaa. (Fritz 2007, Ethics as a science.)

Kohtaan (2) on kaksi pääsyytä: Sitä noudattamalla älykäs systeemi voi (ehkä) välttää muiden haitalliset reaktiot ja tehdä yhteistyötä. Jos älykkään systeemi suorittamasta toiminnosta on jollekin toiselle älykkäälle systeemille enemmän haittaa kuin hyötyä, se saattaa reagoida ei-toivotulla tavalla. Paras tapa varmistaa, ettei haitallisia reaktioita tule, on toimia niin, että jokaiselle älykkäälle systeemille, johon toiminto vaikuttaa, on siitä enemmän hyötyä kuin haittaa. Jos on joitakin, joihin älykäs systeemi ei tiedä toiminnon vaikuttavan, se ei voi tehdä laskelmia näiden osalta. Toiminto on melkein varmasti eettinen (auttaa sitä saavuttamaan tavoitteensa), jos siitä on hyötyä sen suorittavalle älykkäälle systeemille, ja enemmän hyötyä kuin haittaa kaikille muille älykkäille systeemeille, joihin se tietää toiminnon vaikuttavan. (Fritz 2007, Ethics of the artificial intelligent system.)

Kaikki älykkäät systeemit eivät välttämättä osaa tai halua käyttäytyä eettisesti. Tällöin toiset älykkäät systeemit tai yhteisö voi yrittää muuttaa sen käytöstä. Jos älykäs systeemi ei halua

käyttäytyä eettisesti, eli se uskoo saavuttavansa tavoitteensa paremmin toiminnalla, josta on enemmän haittaa kuin hyötyä toisille, tämä voidaan yrittää estää osoittamalla, että sen usko ei pidä paikkaansa. Tämä voidaan tehdä joko sanoin tai teoin, uhkaamalla tai rankaisemalla. Mahdollinen rangaistus vähentää todennäköisyyttä, että älykäs systeemi valitsee tietyn vasteen tulevaisuudessa tekemällä siitä todennäköisesti vähemmän hyödyllisen.

Yhteisö syntyy seuraavasti. Joku älykäs systeemi uskoo, että se voi saavuttaa tavoitteensa paremmin toisten avustuksella kuin yksin ja kommunikoi ajatuksensa toisille älykkäille systeemeille. Tämä kommunikointi voidaan tehdä sanoin tai teoin. Se saa ne toimimaan yhdessä tavoitteen saavuttamiseksi, joten ryhmällä on yhteinen tavoite. Oletetaan vielä, että ryhmän tavoite on pitkäaikainen. Tällöin, koska se on koostunut monista osista jotka ovat älykkäitä systeemejä, ja joilla on yhteisiä tavoitteita ja systeemin olemassaolon aika ylittää sen yksittäisen jäsenen elinajan, sitä voidaan sanoa yhteisöksi edellä annetun määritelmän mukaan. (Fritz 2007, Creation of a society.)

Ihmiselle yhteisö on tärkeä ainakin seuraavista syistä:
1) *Mahdollisuus käyttää yhteisön yhteisiä resursseja, energiaa, tavaroita ja tietoa.* (Firtz 2007, The importance of societies.)
2) *Mahdollisuus tehdä yhteistyötä, työn jakaminen*. (Firtz 2007, The importance of societies.)

Suunnittelun ja päätöksenteon teoriat ja näiden tehtävien ainakin osittainen automatisointi on nyt ja tulevaisuudessa hyödyllistä mm. taloudessa, politiikassa, hallinnoinnissa ja projektien suunnittelussa.

Lähteet

Achinstein, P. 1968. Concepts of science, a philosophical analysis. London: The Johns Hopkins Press 1968.

Aerts, Diederick, Apostel, Leo, De Moor, Bart, Hellemans, Staf, Maex, Edel, Van Belle, Hubert, Van der Veken, Jan. 1994. *World views. From Fragmentation to Integration.* VUB Press. Translation of (Apostel and Van der Veken 1991) with some additions. http://www.vub.ac.be/CLEA/pub/books/worldviews.pdf (accessed July 4, 2007).

Banathy, B.H. 1997. "A Taste of Systemics", The Primer Project. Luettu 11.3.2008. URL = http://www.newciv.org/ISSS_Primer/asem04bb.html

Bennet, S., (1979) A history of control engineering 1800-1930. IEE Control Engineering Series 8, Peter Peregrinus Ltd., London, 1979.

Berkeley, G., 1710, Principles of Human Knowledge.

Berrien, K. F. (1968). General and Social Systems, Rutgers University Press, New Brunswick, NJ.

Bertalanffy, L. von (1949). Das Biologische Weltbild, Franke A. G. Verlag, Bern (in English: Problems of Life, Watts, London, 1952).

Bertalanffy, L. von (1950). An outline of general system theory. British Journal for the Philosophy of Science l(2), 134-165.

Bertalanffy, Ludwig von 1968. General system theory: foundations, development, applications. New York, 1968: George Braziller.

Bertsekas, D. P. ja Tsitsiklis, J. N. *Neuro-Dynamic Programming*. Athena Scienti¯c, Belmont, MA, 1996.

Binet. A., & Simon, T. (1916). The development of intelligence in children. Baltimore, Williams & Wilkins. (Reprinted 1973, New York: Arno Press; 1983, Salem, NH: Ayer Company).

Black, Donald 2000a. The Purification of Sociology. Contemporary Sociology, Vol. 29, No. 5. (Sep., 2000), pp. 704-709. URL: http://links.jstor.org/sici?sici=0094-3061%28200009%2929%3A5%3C704%3ATPOS%3E2.0.CO%3B2-4

Black, Donald 2000b. Dreams of Pure Sociology. Sociological Theory, Vol. 18, No. 3. (Nov., 2000), pp. 343-367. URL: http://links.jstor.org/sici?sici=0735-2751%28200011%2918%3A3%3C343%3ADOPS%3E2.0.CO%3B2-P

Boulding, K. (1953). Toward a general theory of growth. Canadian Journal of Economics and Political Science 19 (reprinted in General Systems Yearbook, vol. 1, 1956).

Brandom, R. (1994). Making It Explicit: Reasoning, Representing, and Discursive Commitment. Cambridge, MA: Harvard University Press.

Broad, Charlie Dunbar. 1924. Critical and Speculative Philosophy, in *Contemporary British Philosophy: Personal Statements* (First Series), ed. J. H. Muirhead (London: G. Allen and Unwin, 1924): 77-100.

Broad, Charlie Dunbar 1947. "Some methods of speculative philosophy". *Aristotelian Society Supplement* 21, p1- 32. http://www.ditext.com/broad/smsp.html (accessed July 4, 2007).

Broad, Charlie Dunbar 1958. "Philosophy", *Inquiry* I, p99-29. http://www.ditext.com/broad/philo.html (accessed July 4, 2007).

Buckley, W. (1967). Sociology and Modern Systems Theory, Prentice-Hall, Englewood Cliffs, NJ.

Buckley, W. (ed.) (1968). Modern Systems Research for the Behavioral Scientist, Aldine, Chicago.

Bartholomew, David J. 2004. Measuring Intelligence: Facts and Fallacies. Cambridge University Press.

Buss, D., (ed.), 2005, The Handbook of Evolutionary Psychology, Hoboken, NJ: Wiley, Hoboken, NJ.

Buss, D., 2007, Evolutionary Psychology: The New Science of the Mind, Boston: Allyn and Bacon.

Calvin, William H. "Pumping Up Intelligence: Abrupt Climate Jumps and the Evolution of Higher Intellectual Functions during the Ice Ages," in *The Evolution of Intelligence*, edited by R. J. Sternberg (Erlbaum, 2001), pp. 97-115. See also *http://WilliamCalvin.com/1990s/1999intelligence-chapter.htm*.
Carey, S. (1985). Conceptual Change in Childhood. Cambridge, MA: MIT Press.

Cat, Jordi, "The Unity of Science", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/scientific-unity/>.

Chomsky, N. (1980). Rules and Represenations. New York: Columbia University Press

Churchland, P. (1981). Eliminative Materialism and Propositional Attitudes. Journal of Philosophy, 78.2, 67-90.

Churchland, P. 1989. A neurocomputational perspective: the nature of mind and the structure of science. Cambridge (Mass.): MIT Press 1989.

Churchland, P. (1995). The Engine of Reason, The Seat of the Soul: A Philosophical Journey into the Brain, MIT Press, 1995.

Churchland, Paul M. 1996. The engine of reason, the seat of the soul: a philosophical journey into the brain. The MIT Press: Cambridge, Massachusetts.

Collins & Quillian 1969. Retrieval time from semantic memory. Journal of erbal learning & verbal behaviour, 8, 240-248.

Craig, Edward 2005. Ontology. Entry in: The shorter routledge encyclopedia of philosophy. New York: Taylor & Fracis group.

Csanyi, V. (1989). The replicative model of selforganization. In Dalenoort, G. j. (ed.), The Paradigm of Self-organization, Gordon & Breach, New York.

Daly, H. (1973). Towards a Steady-state Economy, Freeman, San Francisco.

Danielson, P. 1998. Modeling Rationality, Morality, and Evolution. New York: Oxford Press.

Darwin, Charles 1874. The descent of man and selection in relation to sex. 2. painos. 2 osaa. American home library. New York.

Dewitt, R. 2004. Worldview: an introduction to the history and philosophy of science. Malden, MA : Blackwell , 2004

Dummett, M. (1993). Seas of Language. Oxford: Oxford University Press.

Duntley ja Buss 2004. The evolution of evil. Teoksessa: Miller, Arthur G. (Editor) 2004. Social Psychology of Good and Evil. Guilford Publications, Incorporated.

Elchardus, Marc. (ed.) 1998. *Wantrouwen en Onbehagen*, VUB Press, Brussels.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). Rethinking Innateness: A Connectionist Perspective on Development. Cambridge, MA: MIT Press.

Enqvist, Kari. Olemisen porteilla. WSOY: Porvoo 1998.

Eysenck, Michael W. ja Mark T. Keane 2005. Cognitive psychology: a student's handbook. 5th ed. Hove: Psychology Press , 2005

Feibleman, James K. 1961. The Scientific Philosophy. Philosophy of Science, Vol. 28, No. 3. (Jul.1961), pp. 238-259. URL: http://links.jstor.org/sici?sici=0031-8248%28196107%2928%3A3%3C238%3ATSP%3E2.0.CO%3B2-A

Fieser, James 2006. Ethics. The internet encyclopedia of philosophy. URL = http://www.iep.utm.edu/e/ethics.htm

Flinn M. V. et al. 2005. Ecological dominance, social competition, and coalitionary arms races: Why humans evolved extraordinary intelligence. Evolution and Human Behavior 26 (2005) 10-46. Teksti luettu ja saatavilla 24.1.2008 osoitteessa: URL: http://web.missouri.edu/~gearyd/Flinnetal2005.pdf

Floridi, Luciano 2004. Information. Teoksessa: The Blackwell guide to the philosophy of computing and information. edited by Luciano Floridi. Malden, MA: Blackwell Pub, 2004

Floridi, Luciano 2007. Semantic Conceptions of Information. The Stanford Encyclopedia of Philosophy (Spring 2007 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2007/entries/information-semantic/>.

Fodor, J. (1975). The Language of Thought. Cambridge, MA: Harvard University Press.

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

Fodor, J. & Lepore, E. (1992). Holism: A Shopper's Guide. Oxford: Blackwell.

Fodor, J.A. ja Pylyshyn, Z. (1988), "Connectionism and Cognitive Architecture: A Critical Analysis," Cognition 28: 3-71.

Fogel D. B. 1995. Review of computational intelligence: Imitating life. Proc. of the IEEE, 83(11), 1995

Foster, M.H.; Martin, M.L. (eds.), *Probability, Confirmation,and Simplicity*. New York: The Odyssey Press, 1966.

Franklin ja Graesser 1996. Is it an agent, or just a program? : a taxonomy for autonomous agents. Institute for intelligent systems. University of Memphis. Proceedings of the third international workshop

on agent theories, architectures, and languages, Springer-Verlag, 1996. Haettu 30.10. 2008. URL = http://www.msci.memphis.edu/~franklin/AgentProg.html

François, Charles 1999. Systemics and cybernetics in a historical perspective. Systems research and behavioral science
Syst Res. 16, 203-219 (1999).

François, Charles 2000. History and philosophy of the systems sciences. URL = http://wwwu.uniklu.ac.at/gossimit/ifsr/francois/papers.htm#4

Frigg, Roman, Hartmann, Stephan, "Models in Science", The Stanford Encyclopedia of Philosophy (Spring 2006 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2006/entries/models-science/>

Fritz, Walter 2007. Intelligent Systems and their Societies. http://www.intelligent-systems.com.ar/intsyst/index.htm#IS. First version: Jan 27, 1997. Last Edited 23.2. 2007 / Walter Fritz. Copyright © 1996 New Horizons Press.

Fulcher, John (Editor) 2006. *Advances in Applied Artificial Intelligence*. Hershey, PA, USA: Idea Group Publishing.
http://site.ebrary.com/lib/jyvaskyla/Doc?id=10116545&ppg=132

Gardner, H. 1993. Frames of Mind: Theory of multiple intelligences. Fontana Press, 1993.

Geary, D. C. (2005). The origin of mind: Evolution of brain, cognition, and general intelligence. Washington, DC: American Psychological Association.

Gershenson, Carlos. 2007. *Design and Control of Self- organizing Systems*, PhD Thesis, VUB,

Gettier, E. (1963). Is Justified True Belief Knowledge? Analysis, 23, 121-123.

Georgescu-Roegen, N. (1971). The Entropy Law and the Economic Progress, Harvard University Press, Cambridge, M.A.

Gintis, H. 2007. A framework for the unification of the behavioral sciences. URL = http://people.umass.edu/gintis/ Julkaistu: Behavioral and brain sciences (2007) 30,1-61.

Gopnik, A., & Meltzoff, A. (1997). Words, Thoughts, and Theories. Cambridge, MA: MIT Press.

Gottfredson, L. S. 1997. Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. Intelligence, 24(1):13-23, 1997.

Gottfredson, L. S. 2002. g: Highly general and highly practical. In R. J. Sternberg and E. L. Grigorenko, editors, The general factor of intelligence: How general is it?, pages 331-380. Erlbaum, 2002.

Gupta, Anil, "Definitions", *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/definitions/>.

Haaparanta ja Niiniluoto 1998. Johdatus tieteelliseen ajatteluun. Helsinki 1998: Hakapaino Oy.

Hammond, Debora 2003. *Science of Synthesis: Exploring the Social Implications of General Systems Theory*. Boulder, CO, USA: University Press of Colorado.
http://site.ebrary.com/lib/jyvaskyla/Doc?id=10069588&ppg=292

Harris, Errol E. 1952. Scientific Philosophy. The Philosophical Quarterly, Vol. 2, No. 7. (Apr., 1952), pp. 153-165. URL: http://links.jstor.org/sici?sici=0031-8094%28195204%292%3A7%3C153%3ASP%3E2.0.CO%3B2-D

Harva, Urpo 1980. Maailmankatsomuksen ongelmia. Otava: Keuruu 1980.

Hawkins, Jeff 2005. Älykkyys: uusi tieto aivoista ja älykkäät koneet. Helsinki: Edita, 2005.

Henrich, J. , R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, K. Hill, F. Gil-White, M. Gurven, F. Marlowe, J. Q. Patton, N. Smith, and D. Tracer, 'Economic Man' in Cross-cultural Perspective: Behavioral Experiments in 15 Small-scale Societies. URL = http://www.sscnet.ucla.edu/anthro/faculty/boyd/Publications.htm Juolkaistu: Behavioral and Brain Sciences, 28: 795-855, 2005.

Herman L. M. ja Pack, A. A. 1994. Animal intelligence: Historical perspectives and contemporary approaches. In R. Sternberg, editor, Encyclopedia of Human Intelligence, pages 86-96. Macmillan, New York, 1994

Herrick, Paul 1999. Reason and Worldview: An Introduction to Western Philosophy. Toronto: Thompson Publishing, 1999.

Heylighen, Francis. 1997. Objective, subjective and intersubjective selectors of knowledge, *Evolution and Cognition* 3:1, p. 63-67. http://pespmc1.vub.ac.be/Papers/KnowledgeSelectors.pdf (accessed July 4, 2007).

Heylighen, F. 1998. Basic concepts of the systems approach. Principia cybernetica web. 13.3.2008. URL = http://pespmc1.vub.ac.be/SYSAPPR.html

Heylighen, F. 1998: "Technological acceleration", in: F. Heylighen, C. Joslyn and V. Turchin (editors): Principia Cybernetica Web (Principia Cybernetica, Brussels), URL: http://pespmc1.vub.ac.be/TECACCEL.html

Heylighen, F.1999: "Happiness", in: F. Heylighen, C. Joslyn and V. Turchin (editors): *Principia Cybernetica Web* (Principia Cybernetica, Brussels), URL: http://pespmc1.vub.ac.be/HAPPINES.html

Heylighen, Francis and Bernheim, Jan. 2000. Global Progress I: empirical evidence for increasing quality of life, *Journal of Happiness Studies* 1 (3): 323-349. http://pespmc1.vub.ac.be/Papers/ProgressI&II.pdf (accessed July 4, 2007).

Heylighen, F. 2000. "World View", in: F. Heylighen, C. Joslyn and V. Turchin (editors): *Principia Cybernetica Web* (Principia Cybernetica, Brussels), http://cleamc11.vub.ac.be/WORLVIEW.html

Heylighen, F. 2000: "What is a world view?", in: F. Heylighen, C. Joslyn and V. Turchin (editors): Principia Cybernetica Web (Principia Cybernetica, Brussels), URL: http://pespmc1.vub.ac.be/WORLVIEW.html

Heylighen, F. , Joslyn C. 1992. What is systems theory? Principia cybernetica web. 11.3.2008. URL = http://pespmc1.vub.ac.be/SYSTHEOR.html

Heylighen, Francis 2008. Web Dictionary of Cybernetics and Systems. Principia cybernetica web. 6.3.2008. URL= http://cleamc11.vub.ac.be/ASC/IndexASC.html

Hofweber, Thomas, "Logic and Ontology", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/logic-ontology/>.

Hume, D. (1739/1978). A Treatise of Human Nature. Oxford: Oxford University Press.

Hutter, M. 2000. Algorithmic "Kolmogorov" Complexity (AC). Käyty 19.7.2008. URL = http://www.hutter1.net/ait.htm#intro

Hyvönen, Eero 1993. Tekoäly ja tietämystekniikka. Teoksessa: Tekoälyn ensyklopedia. Karisto Oy: Hämeenlinna 1993.

Jaffe, K. 2002. An Economic Analysis of Altruism: Who Benefits from Altruistic Acts? Journal of Artificial Societies and Social Simulation 5 (3).

Kant I. Critique of judgement: including the first introduction. Indianapolis: Hackett 1987.

Karanta, Ilkka 1993. Tehtäväsuunnittelu. Artikkeli teoksessa: Tekoälyn ensyklopedia. Karisto Oy: Hämeenlinna 1993.

Kaufman, A. S. 2000. Tests of intelligence. In R. J. Sternberg, editor, Handbook of Intelligence. Cambridge University Press, 2000.

Keil, F. C. 1989. Concepts, kinds and cognitive development. Cambridge, MA: MIT Press.

Ketonen, Oiva 1981. Eurooppalaisen ihmisen maailmankatsomus.WSOY: Juva 1981.

Kitcher, P., 1989, 'Explanatory Unification and the Causal Structure of the World', in *Scientific Explanation,* P. Kitcher and W. Salmon, 410-505. Minneapolis: University of Minnesota Press.

Kitcher, P., The Advancement of Science: Science without Legend, Objectivity without Illusions. Oxford: Oxford University Press, 1993.

Koikkalainen ja Orponen 2002. Tietotekniikan perusteet. Moniste Jyväskylän yliopistossa luennoitavan tietotekniikan approbatur-kurssin ``Tietotekniikan perusteet'' tarpeisiin. Haettu 17.3.2007 osoitteesta http://erin.mit.jyu.fi/pako/kurssit/perusteet/kirja/kirja.html

Kuipers, T., From Instrumentalism to Constructive Realism. Dordrecht: D. Reidel, 2000.

Kuratowski, K. ja A. Mostowski 1976. Set Theory: With an Introduction to Descriptive Set Theory. North-Holland, Amsterdam.

Kurzweil R. 2000. The age of spiritual machines: When computers exceed human intelligence. Penguin, 2000.

Körner, Stephan. 1969. *Fundamental Questions of Philosophy. One philosopher's answers.* 4th edition, Redwood Burn Limited.

Lakatos, I. and Musgrave, A. (eds.), Criticism and the Growth of Knowledge. Cambridge: Cambridge University Press, 1970.

Langton, C. (ed.) (1989). Artificial Life. Santa Fe Institute for Studies in the Sciences of Complexity, Addison-Wesley, Reading, MA.

LeDoux, Joseph E. 2003. Synaptinen itse: miten aivot tekevät minusta minut. Helsinki : Terra Cognita , 2003

Le, S. ja Boyd, R. 2007. Evolutionary Dynamics of the Continuous Iterated Prisoner's Dilemma. URL = http://www.sscnet.ucla.edu/anthro/faculty/boyd/Pub. Julkaistu: Journal of Theoretical Biology, Volume 245, 258-267.

Legg ja Hutter 2005. A universal measure of intelligence for artificial agents. Haettu 21.3.2007 osoitteesta http://www.hutter1.net/index.htm

Legg ja Hutter 2006. A formal measure of machine intelligence. In Proc.15th Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'06), pages 73-80, Ghent, 2006.

Legg ja Hutter 2007a. A collection of definitions of intelligence. Julkaisussa: Frontiers in Artificial Intelligence and Applications, Vol.157 (2007) 17-24. Käyty: 23.2.2008. URL: http://arxiv.org/abs/0706.3639v1
Legg ja Hutter 2007b. Universal intelligence: A definition of machine intelligence. Minds and Machines, 17:4 (2007) 391-444. URL = http://www.hutter1.net/

Lin, Y. 1989. A multi-relation approach of general systems and tests of applications. Synthese: Int. J. Epistem. Methodol. Phil. Sci., 79, 473? 488.

Lin, Y. 1999. *General Systems Theory: A Mathematical Approach.* New York, NY, USA: Kluwer Academic Publishers. http://site.ebrary.com/lib/jyvaskyla/Doc?id=10052618&ppg=24

Lin, Y. 2003. Systems and cybernetics: new theories and applications - part II. Emerald Group Publishing Limited.

Locke, J. (1690/1975). An Essay Concerning Human Understanding. New York: Oxford University Press.

López, B. (Editor) 2005. Artificial Intelligence Research and Development, Volume 131. Amsterdam, NLD: IOS Press.
http://site.ebrary.com/lib/jyvaskyla/Doc?id=10155203&ppg=2

Luisi, Pier Luigi 2002. Emergence of Life : From Chemical Orgins to Synthetic Biology. Cambridge University Press.

Macy, M. W. ja Willer, W. 2002. From Factors to Actors: Computational Sociology and Agent-Based Modeling, Annual Review of Sociology 28, 143-166.

Mainzer, Klauz 2004. System: an introduction to systems science. Teoksessa: The Blackwell guide to the philosophy of computing and information.

Margolis ja Stephen 2007. "Concepts", The Stanford Encyclopedia of Philosophy (Winter 2007 Edition), Edward N. Zalta (ed.), Luettu: 21.2.2008. URL = <http://plato.stanford.edu/archives/win2007/entries/concepts/>.

Maturana, H. ja Varela, F. (1980). Autopoiesis and Cognition, Reidel, Boston, MA.

Merriam-Webster online dictionary 2007. Universe. 5.12. 2007. URL: http://m-w.com/dictionary/universe.

Merriam-Webster online student dictionary 2007. Universe. 5.12. 2007. URL: http://www.wordcentral.com/cgi-bin/student?universe.

Miller, J. G. (1978). Living Systems, McCraw Hill, New York.

Millikan, R. (2000). On Clear and Confused Ideas. Cambridge: Cambridge University Press.

Mulkay, M.J. 1975. Three Models of scientific development. The sociological review 23 (1975).

Murphy, G. (2002). The Big Book of Concepts. Cambridge, MA: MIT Press.

Myers, David G. 1993. *The Pursuit of Happiness.* Avon Books.

Naugle, D. K. 2002. Worldview: the history of a concept. Grand Rapids, Mich. : W. B. Eerdmans Pub. Co , cop. 2002

Neisser U. , G. Boodoo, T. J. Bouchard, Jr., A. W. Boykin, N. Brody, S. J. Ceci, D. F. Halpern, J. C. Loehlin, R. Perloff, R. J. Sternberg, and S. Urbina. 1996. Intelligence: Knowns and unknowns. American Psychologist, 51(2):77-101, 96.

Newall, Paul 2005. Ethics. The galilean library. Manuscripts. URL = http://www.galilean-library.org/int11.html

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

Newell, Allen 1990. Unified theories of cognition. Cambridge, MA : Harvard University Press, 1990.

Niemelä, Ilkka 1993. Logiikka tietämyskielenä. Artikkeli teoksessa: Tekoälyn ensyklopedia. Karisto Oy: Hämeenlinna 1993.

Niiniluoto, Ilkka 1980. Johdatus tieteenfilosofiaan: käsitteen- ja teorianmuodostus. Helsinki: Otava, 1980.

Niiniluoto, Ilkka 1984. Tiede, filosofia ja maailmankatsomus. Helsinki: Otava, 1984.

Niiniluoto, Ilkka 1989. Informaatio, tieto ja yhteiskunta: filosofinen käsiteanalyysi. Hki: Valtion painatuskeskus.

Niiniluoto, Ilkka 1994. Järki, arvot ja välineet - kulttuurifilosofisia esseitä. Otava: Keuruu 1994.

Niiniluoto ja Haaparanta 1998. Johdatus tieteelliseen ajatteluun. Helsinki: Hakapaino Oy, 1998.

Niiniluoto, Ilkka 2003. Totuuden rakastaminen. Helsinki: Otava, 2003.

Niiniluoto, Ilkka, "Scientific Progress", *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/scientific-progress/>.

Odum, H. (1971). Environment, Power and Society, Wiley, New York.

Orr, Larry R. 1999. Social experiments: Evaluating public programs with experimenal methods. Sage publications, Inc. : USA 1999.

Passmore, John. 1967. *Philosophy* entry, p. 216-226. *The Encyclopedia of Philosophy*, P. Edwards (ed.), Macmillan, London.

Peacocke, C. (1992). A Study of Concepts. Cambridge, MA: MIT Press.

Pinker, S. (1994). The Language Instinct: The New Science of Language and Mind. London: Penguin.

Pitt, David, "Mental Representation", The Stanford Encyclopedia of Philosophy (Spring 2007 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2007/entries/mental-

representation/>. Port, R. & van Gelder, T. (1995). Mind as Motion: Explorations in the Dynamics of Cognition. Cambridge, MA.: MIT Press

Poincaré, H. (1898). Les Méthodes Nouvelles de la Mécanique Céleste, reprint: Dover, New York, 1957.

Pollock, John L. 2006. Thinking about Acting: Logical Foundations for Rational Decision Making. New York, London: Oxford University Press.

Pollock 2008. OSCAR: a cognitive architecture for intelligent agents. URL = http://www.sambabike.org/ftp/publications.html

Popper, Karl. 1958. On the Status of Science and of Metaphysics *Ratio,* 1, No. 2, pp. 97-115. in: *Conjectures and Refutations. The Growth of Scientific Knowledge* (5th edition, revised; London & New York: Routledge, 1989), 184-200.

Popper, Karl R. 1995. Arvauksia ja kumoamisia. Helsinki: Gaudeamus, 1995.

Project Worldview 2007. Worldviews--An Introduction. URL: http://www.projectworldview.org/worldviews.htm

Psillos, S., Scientific Realism: How Science Tracks Truth. London: Routledge, 1999.

Ree, M. J. & Earles, J. A. (1992). Intelligence is the best predictor of job performance. Current Directions in Psychological Science, 1, 86-89.

Reichenbach, Hans 1956. The rise of scientific philosophy. Los Angeles: University of California press.

Rescher, N., Methodological Pragmatism. Oxford: Blackwell, 1977.

Rescher, Nicholas 2001. *Philosophical Reasoning. A Study in the Methodology of Philosophizing.* Blackwell publishers.

Richerson, Bettinger ja Boyd 2005. Evolution on a restless planet: Were environmental variability and environmental change major drivers of human evolution? Handbook of Evolution Vol. 2 edited by Franz M. Wuketits and Francisco J. Ayala, Pp 223-242. 2005.

Richerdson, Boyd, Gintis and Bowles. 2003. The Evolution of Altruistic Punishment. URL = http://www.des.ucdavis.edu/faculty/Richerson/recent%20cultutral%20new.htm Julkaistu: Proceedings of the National Academy of Sciences (USA) 100: 3531-3535.

Richerson ja Boyd 2006a. Ei ainoastaan geneistä. Miten kulttuuri muunsi ihmisen evoluution. Hakapaino: Helsinki.

Richerson ja Boyd 2006b. Culture and the evolution of the human social instincts. In: Roots of Human Sociality: Culture, Cognition, and Interaction. Edited by N.J. Enfield and Stephen C. Levinson. Berg. Pp. 453-477.

Richerson, Paciotti, Boyd 2006. Cultural Evolutionary Theory: A Synthetic Theory for Fragmented Disciplines. URL = http://www.des.ucdavis.edu/faculty/Richerson/recent%20cultutral%20new.htm In Bridging Social Psychology: The Benefits of Transdisciplinary Approaches, Paul Van Lange, Editor. Pp 365-70. 2006.

Ricoeur, Paul. 1979. (reporter). *Main Trends in Philosophy.* Holmes & Meier.

Robinson, Howard, "Dualism", The Stanford Encyclopedia of Philosophy (Fall 2006 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2006/entries/dualism/>.

Rorty, Richard 1983. Philosophy and the mirror of nature. Oxford: Blackwell, 1983.

Ross, Don, "Game Theory", The Stanford Encyclopedia of Philosophy (Spring 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2008/entries/game-theory/>.

Rumelhart, D.E. (1989), "The Architecture of the Mind: A Connectionist Approach," in M.I. Posner, ed., Foundations of Cognitive Science, Cambridge, Mass.: The MIT Press: 133-159.

Russell, Bertrand 1948. Länsimaisen filosofian historia, osa 1. WSOY: Porvoo 1948.

Russell ja Norvig 1995. Artificial intelligence: a modern approach. Upper saddle river, NJ 07458: Prentice Hall, Inc.

Russell ja Norvig 2003. Artificial intelligence: a modern approach. 2nd ed. Upper Saddle River (NJ) : Prentice Hall , 2003

Russon ja Begun 2004. Evolutionary origins of great ape intelligence: an integrated view. Teoksessa: Russon, Anne E. (Editor)Evolution of Thought: Evolutionary Origins of Great Ape Intelligence. Cambridge University Press.

Saam, N. J. and Harrier, A. 1999. Simulating Norms, Social Inequality, and Functional Change in Artificial Societies. Journal of Artificial Societies and Social Simulation 2, (1).

Sabbatini, Renato M.E. 2001. The Evolution of Intelligence. Brain & Mind Magazine, February/April 2001. Luettu 20.2.2008. URL: http://www.epub.org.br/cm/n12/mente/evolution05_i.html

Salmi, K. 2002. Aivojen ajatus - kiehtova mieli, Ihminen evoluution tuotteena. Hämeenlinna: Karisto Oy 2002.

Sayre, K. 1993. Three more flaws in the computational model. APA (central division) Annual conference, Chicago, Illinois.

Seiler ja Jenkins 2004. Frequently asked questions about Isaac Asimov. Haettu 18.3.2007 osoitteesta http://www.asimovonline.com/asimov_FAQ.html

Shannon, C. and Weaver, W. (1949). The Mathematical Theory of Communication, University of Illinois Press, Urbana.

Simonton D. K. 2003. An interview with Dr. Simonton. In J. A. Plucker, editor, Human intelligence: Historical influences, current controversies, teaching resources. http://www.indiana.edu/~intell, 2003.

Skyrms, B., Harms, W. Evolution of Moral norms. URL = http://www.lps.uci.edu/home/fac-staff/faculty/skyrms/Skyrmspapers.html Julkaistavaksi teoksessa: Oxford Handbook on the Philosophy of Biology ed. Michael Ruse.

Smith ja Szathmáry 1995. The major transitions in evolution. Oxford: W. H. Freeman, cop. 1995.

Smolensky, P. (1988), "On the Proper Treatment of Connectionism," Behavioral and Brain Sciences, 11: 1-74.

Sober, Elliott (Editor) 2006. Conceptual Issues in Evolutionary Biology (3rd Edition). MIT Press.

Sowa, John F. 2000. Processes and Causality. Haettu 21.3.2007 osoitteesta http://www.jfsowa.com/ontology/causal.htm

Spearman, C. E. 1927. The abilities of man, their nature and measurement. Macmillan, New York, 1927.

StatSoft, Inc. 2007. Correlation. Electronic Statistics Textbook. Tulsa, OK: StatSoft. URL = http://www.statsoft.com/textbook/stathome.html.

Sternberg, R. J. 1985. Beyond IQ: A triacrchi theory of human intelligence. Cambridge University Press, New York.

Sternberg R. J. (editor) 2000. Handbook of Intelligence. Cambridge University Press.

Sternberg R. J. 2003. An interview with Dr. Sternberg. In J. A. Plucker, editor, Human intelligence: Historical influences, current controversies, teaching resources. http://www.indiana.edu/~intell, 2003

Steup, Matthias, "Epistemology", The Stanford Encyclopedia of Philosophy (Fall 2006 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2006/entries/epistemology/>.

Stoljar, Daniel, "Physicalism", The Stanford Encyclopedia of Philosophy (Winter 2005 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2005/entries/physicalism/>.

Sutton, R. ja Barto, A. *Reinforcement learning: An introduction*. Cambridge, MA, MIT Press, 1998.

Totten, R 2004. Worldview test site. Truth-tests for choosing a worldview. Käyty: 10.12.2007. URL: http://www.geocities.com/worldview_3/truthtests.html

Thurstone, L. L. 1938. Primary mental abilities. University of Chicago Press, Chicago.

Turing, A.M. (1950). Computing machinery and intelligence. Mind, 59, 433-460. URL = http://www.loebner.net/Prizef/TuringArticle.html

Turing, Alan 1950. Computing Machinery and Intelligence. Mind LIX (236): 433-460.

Ukkonen, Esko 1993. Algoritmin käsite. Teoksessa: Tekoälyn ensyklopedia. Karisto Oy: Hämeenlinna 1993.

Ukkonen, Esko 1993. Tietojenkäsittelytiede. Artikkeli teoksessa: Tekoälyn ensyklopedia. Karisto Oy: Hämeenlinna 1993.

Vallée, R. 1993. Systems theory, a historical presentation. In Rodriguez Delgado, R., and Banathy, B. (eds), International Systems Science Handbook, Systemic Publications, Madrid.

Wang P. 1995. On the working definition of intelligence. Technical Report 94, Center for Research on Concepts and Cognition, Indiana University, 1995.

Vidal, C. 2007. An enduring philosophical agenda - worldview construction as a philosophical method. (preprint) haettu: 14.12.2007 osoitteesta: URL: http://cogprints.org/5408/

Wiener, N. 1948. Cybernetics or Control and Communication in the Animal an the Machine. Hermann: Paris.

Wilhelm Jerusalem 1926. Filosofian alkeet. WSOY.

Wilson, George 2008. "Action", The Stanford Encyclopedia of Philosophy (Spring 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2008/entries/action/>.

Viitala, Jussi 2004. Inhimillinen eläin, eläimellinen ihminen: sosiaalisen käyttäytymisen avaimet. Jyväskylä: Atena.

Wilson, George, "Action", The Stanford Encyclopedia of Philosophy (Winter 2007 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2007/entries/action/>.

Wolfram, Stephen 2002. A new kind of science. Champaign, IL: Wolfram Media, cop. 2002.

Wolters, Albert M. 1989. “On the Idea of Worldview and Its Relation to Philosophy.”In *Stained Glass: Worldviews and Social Science*. Christian Studies Today, edited by Paul A. Marshall, Sander Griffioen, and Richard J. Mouw, 14-25. Lanham, MD: University Press of America.

Von Neumann J. ja Morgenstern O. (1947). The Theory of Games and Economic Behavior. Princeton: Princeton University Press, 2nd edition.

Woodward, James, "Scientific Explanation", *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/scientific-explanation/>.

Workman, Lance 2004. *Evolutionary Psychology: An Introduction.* West Nyack, NY, USA: Cambridge University Press, 2004. URL = http://site.ebrary.com/lib/jyvaskyla/Doc?id=10131668&ppg=206

Yolton, R., 1983, Thinking Matter, Minneapolis: University of Minnesota Press.

Younger, S. M. 2002 Discrete Agent Modeling as a Tool For the Study of Individual and Social Development: The MICROS Code (Version 8.0), Los Alamos National Laboratory Report LA-UR-02-7009. URL = http://lib-www.lanl.gov/cgi-bin/getfile?00852082.pdf

Younger, S. M. 2003. Discrete Agent Simulations of the Effect of Simple Social Structures on the Benefits of Resource Sharing. Journal of Artificial Societies and Social Simulation vol. 6, no. 3. URL = http://jasss.soc.surrey.ac.uk/6/3/1.html

Yudkowsky, Eliezer S. 2002. Levels of organization in general intelligence. Research Fellow, Singularity Institute for Artificial Intelligence, Inc. Teksti luettu ja saatavilla 23.1.2008 osoitteessa: URL: http://www.singinst.org/upload/LOGI//. Ilmestyy teoksessa: Real AI: New Approaches to Artificial General Intelligence. Ben Goertzel and Cassio Pennachin, eds.

Zalta, E. (2001). Fregean Senses, Modes of Presentation, and Concepts. Philosophical Perspectives, 15, Metaphysics, pp. 335-59.

Zentall, T. R. 2000. Animal intelligence. In R. J. Sternberg, editor, Handbook of Intelligence. Cambridge University Press, 2000.