Joni-Roy Piispanen

Yleinen tekoäly

Tietotekniikan kandidaatintutkielma

21. elokuuta 2019

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Joni-Roy Piispanen

Yhteystiedot: joni.r.e.piispanen@student.jyu.fi

Työn nimi: Yleinen tekoäly

Title in English: Artificial General Intelligence

Työ: Kandidaatintutkielma

Sivumäärä: 23+0

Tiivistelmä: Tekoälytutkimus on muodostunut merkittäväksi yhteiskunnalliseksi ja tieteelliseksi aihealueeksi, varsinkin sen viimeaikaisten läpimurtojen takia. Se on saavuttanut uskomattomia asioita esimerkiksi lääketieteessä. Tätä toiminnallisuutta on pitkään pyritty laajentamaan alakohtaisista ongelmista, yleispäteviin ongelmiin ja tilanteisiin. Tavoitteena tutkijoilla on ollut luoda algoritmejä, jotka saavuttaisivat ihmisen tasoisen yleisen älykkyyden ja ongelmanratkaisu kyvykkyyden. Tekoälytutkimuksen alkuajoilla oletettiin, että tällaisia erityistehtävissä toimivia systeemejä pystyttäisiin yhdistämään koherentiksi järjestelmäksi, joka saavuttaisi yleisen älykkyyden, mutta pian huomattiin tämän olevan haasteellista, ellei jopa mahdotonta. Tästä johtuen tutkijat ottivat mallia ihmisen tiedonkäsittelystä ja mielestä, muodostamalla kognitiivisia arkkitehtuureja, joiden pohjalta yleinen älykkyys voisi muodostua. Tulen käsittelemään näitä tutkimuksen aikana ja pyrin selvittämään, onko tämä lähestymistapa yleisen älykkyyden ratkaisemiseksi toimiva.

Avainsanat: tekoäly, tietoisuus, älykkyys, kognitiiviset arkkitehtuurit

Abstract: Artificial intelligence research has become a significant societal and scientific area of interest, especially in light of it's recent breakthroughs. It has for instance achieved incredible things in medicine. Long-time efforts have been made to extend this functionality from sector-specific problems to general problems and situations. The goal of the researchers has been to create algorithms that achieve human-level general intelligence and problem solving abilities. In the early days of artificial intelligence research, it was assumed that such special function systems could be com-

bined into a coherent system that would achieve general intelligence, but this was soon discovered to be challenging, if not impossible. As a result, researchers have modeled human information processing and the human mind, by creating cognitive architectures that could serve as the basis for general intelligence. I will address these during my research and will try to determine whether they can serve as a viable approach in working towards general intelligence.

Keywords: artificial intelligence, consciousness, intelligence, cognitive architectures

Kuviot

Kuvio 1. *LIDA*-mallin kognitiivinen sykli (Friedlander, David Samuel & Franklin Stan 2008) 10
Kuvio 2. Soar-arkkitehtuuri (Laird, John 2012) 12

Sisältö

1 JOHDANTO 1

2 TEKOÄLY 3

3 KONETIETOISUUS 6

4 KOGNITIIVISET ARKKITEHTUURIT 8
4.1 *LIDA* (*Learning Intelligent Distribution Agent*) 9
4.2 *Soar* 11

5 YHTEENVETO 14

KIRJALLISUUTTA 16

1 Johdanto

Yleinen tekoäly kykenee laaja-alaiseen, useita erilaisia taitoja vaativaan ongelmanratkaisemiseen verrattuna arkipäiväisempään heikkoon tekoälyyn, joka on ollut viime aikoina mediassa vahvasti esillä ja joka pääasiassa toimii erilaisissa rajoitetuissa tehtävissä. Yleinen tekoäly suorittaa yleisesti ihmismäistä ajattelua ja kykenee mihinkä tahansa älykkääseen toimintoon, kuten luonnollisen kielen tuottamiseen ja tunnistamiseen, auton ajamiseen ja reaalimaailman ongelmatilanteiden ratkaisemiseen. Tässä tutkimuksessa älykkyys nähdään kykynä järkeillä, strategisoida, ratkaista pulmia, tehdä päätöksiä epävarman tiedon nojalla, esittää tietoa, johon sisältyy maalaisjärki, suunnitella, oppia, kommunikoida luonnollisella kielellä ja yhdistää näitä taitoja yhteiseen tavoitteeseen. Nykyisellään yleisen tekoälyn tutkimus on rajoittunutta, eikä se ole monien asiantuntijoiden mielestä edistynyt merkityksellisesti viimeisen 50 vuoden aikana (Wang, Pei & Goertzel, Ben 2012, s. 68). Sen määrittely ja kyvykkyydet vaihtelevat hieman eri teoreettisissa lähestymistavoissa, mutta alan tutkijoilla on yhtenäinen intuitio yleisen tekoälyn ydinongelmista.

Sellaisen keinotekoisen älykkyyden luominen, joka kykenee suorittamaan monipuolisia tavallisesti ihmismieltä vaativia tehtäviä, aiheuttaisi valtavia yhteiskunnallisia ja filosofisia seurauksia. Sen avulla pystyttäisiin tarkastelemaan tietoisuuden roolia älykkäässä toiminnassa. Lisäksi se on suoranaisesti relevantti mieli-ruumis-ongelmalle, joka on vaivannut filosofeja pitkään. Sillä olisi myös radikaaleja yhteiskunnallisia seurauksia liittyen työllisyyteen, hyvinvointiin ja yhteiskunnan rakenteisiin. On teorisoitu sen luomisen johtavan niin kutsuttuun teknologiseen singulariteettiin, jossa tekoälyn aiheuttama teknologinen kehitys ja muutos on niin nopeaa, ettei sitä pystytä enää hallitsemaan (Kurzweil, Ray 2005). Tekoälytutkimukseen liittyen on esitetty myös muita dystopisia asetelmia ja sen potentiaalisia hyötyjä ja haittoja on vertailtu laajalti (ks. esim. (Irving, Geoffrey & Askell, Amanda 2019)).

Tekoälytutkimuksen alkuaikoina tutkijat pyrkivät rakentamaan ajattelevia koneita, jotka kykenisivät ihmismielen tasoisiin toimintoihin (Feigenbaum, Edward Albert & Feldman, Julian 1963). Merkittävimmät yritykset edellä mainittuun tavoitteeseen

olivat Herbert A. Simonin, J. C. Shawin ja Allen Newellin kehittämä *General Problem Solver* (Newell, Allen & Shaw, John Clifford & Simon, Herbert Alexander 1959) ja Japanin talous-, kauppa- ja teollisuusministeriön *Fifth Generation Computer Systems* -aloite (Shapiro, Ehud 1983). Nämä kuitenkin epäonnistuivat tavoitteessaan. Tästä johtuen valtaosa tekoäly tutkimuksesta alkoi kohdistua yksittäisiin kognitiivisiin toimintoihin ja alakohtaisiin ongelmiin (Wang, Pei & Goertzel, Ben 2012, s. 2). Vuosien varrella on käyty kiivasta keskustelua siitä, onko mahdollista rakentaa yleisesti älykäs kone yhdistämällä eri osa-alueisiin erikoistuneita järjestelmiä. Valtaosa yleisen tekoälyn tutkijoista pitää tätä mahdottomana (Brachman, Ronald Jay 2006).

Nykyisellään on muodostunut joitain hankkeita ja projekteja, jotka käsittelevät yleistä tekoälyä ja sen sisältämää aihepiiriä. Eräs merkittävimmistä on *DeepMind*, jonka tavoitteena on ratkaista älykkyys luomalla yleispäteviä oppimisalgoritmejä (Wang, Jane & Kurth-Nelson, Zeb & Tirumala, Dhruva & Soyer, Hubert & Leibo, Joel & Munos, Rémi & Blundell, Charles & Kumaran, Dharshan & Botvinick, Matthew 2017). Toinen tärkeä aihepiiriin liittyvä on *The Human Brain* -projekti, jonka tavoitteena on tutkia ja analysoida ihmismieltä ja sen rakennetta, sekä luoda ICT-pohjainen tutkimusinfrastruktuuri aivotutkimukselle ja neurotieteelle, (Amunts, Katrin & Ebell, Christoph & Muller, Jeff & Telefont, Martin & Knoll, Alois & Lippert, Thomas 2016). *OpenAI* on hanke, joka pyrkii tukemaan ja kehittämään ihmisyydelle ystävällistä tekoälyä (OpenAI 2018).

Alan historian aikana on esitetty monia teorioita yleisen tekoälyn rakenteesta, mutta asiantuntijat eivät ole päässeet yhteisymmärrykseen sen luonteesta. Luvussa 2 tarkastelen alan historiaa tarkemmin, sen lähtökohdista nykyaikaan saakka. Luvussa 3 käsittelen älykkyyden ja tietoisuuden suhdetta, sekä sitä miten tietoisuus koneissa ilmenee. Suosituimpia ehdotuksia yleisen tekoälyn pohjaksi ovat kognitiivisiin arkkitehtuureihin, kuten *LIDA* ja *Soar*, pohjautuvat systeemit. Tulen perehtymään näihin syvällisemmin luvussa 4. Viimeisessä luvussa käsittelen tutkimuksen ongelmia, merkitystä ja alan ammattilaisten keskeisimpiä mielipiteitä. Tulen käsittelemään viimeisessä luvussa myös tutkimuksen keskeisimpiä johtopäätöksiä.

2 Tekoäly

Tekoälytutkimuksen katsotaan muodostuneen alana, vuonna 1956 pidetyssä Dartmouthin konferenssissa, (McCarthy, John & Minsky, Marvin Lee & Rochester, Nathaniel & Shannon Claude Elwood 1955) vaikka esimerkiksi Alan Turingin tekemää aikaisempaa työtä voidaan karakterisoida tekoälyyn liittyväksi. Hän esitteli 1950-luvulla julkaistussa tutkimuspaperissaan monia vielä tänäkin päivänä relevantteja konsepteja, kuten koneoppimisen, geneettiset algoritmit ja vahvistusoppimisen (Turing, Alan Mathison 1950). Lisäksi hän muotoili Turingin testin, jonka tarkoitus on tutkia koneen ihmismäisyyttä keskustelutilanteessa. Mikäli kuulustelija ei kykene erottamaan koneen vastauksia ihmisen vastauksista, läpäisee kone testin ja täten ainakin näyttää ajattelevan (Warwick, Kevin & Shah Huma 2016). Samankaltaisia ajatuksia ajattelevista koneista esitettiin myös Dartmouthin konferenssissa, jossa Allen Newell, Herbert A. Simon ja Cliff Shaw esittelivät *Logic Theorist* -tietokoneohjelmansa (Gugerty, Leo 2017). Sitä pidetään yleisesti ensimmäisenä tekoälyohjelmana ja se ohjelmoitiin imitoimaan ihmismäistä ongelmanratkaisemista. Se kykeni todistamaan matemaattisia lausekkeita matemaatikon tasolla.

John McCarthy, joka oli konferenssin järjestäjä, julkaisi vuonna 1959 artikkelin, jossa hän esitteli hypoteettisen algoritmin, joka kykenisi muutamalla yksinkertaisella aksioomalla muodostamaan suunnitelman lentokentälle ajamisesta (McCarthy, John 1959). Algoritmi kykenisi myös oppimaan uusia aksioomia toiminnan aikana, joka mahdollistaisi sen autonomisen toiminnan, ilman uudelleen ohjelmointia. Samoihin aikoihin Allen Newell, Herbert A. Simon ja Cliff Shaw rakensivat *General Problem Solver* -tietokoneohjelman, jonka ongelmanratkaisu perustui ihmismäiseen osaongelmien löytämiseen ja ratkaisemiseen (Newell, Allen & Shaw, John Clifford & Simon, Herbert Alexander 1959). Se kykeni ratkaisemaan yksinkertaisia pulmia, kuten Hanoin tornit ongelman. Huomattiin kuitenkin, ettei ohjelma pysty ratkaisemaan reaalimaailman ongelmia, johtuen haun kombinatorisesta räjähdyksestä, jossa vaikuttavia tekijöitä on liian suuri määrä.

Onnistumisista syntynyt optimismi johti moniin uskaliaisiin lausuntoihin. Muun

muassa Herbert A. Simon totesi vuonna 1957 koneiden ongelmanratkaisukyvyn kehittyvän lähitulevaisuudessa samankaltaiseksi ihmisen ongelmanratkaisukyvyn kanssa (Russell, Stuart Jonathan & Norvig, Peter 2010). Tätä asennetta tukivat myös Dartmouthin konferenssiin osallistuneen Marvin Minskyn oppilaiden 1960-luvulla rakentamat ohjelmat, jotka ratkaisivat eri älykkyyden osa-alueita vaativia ongelmia. Näitä olivat esimerkiksi Daniel Bobrowin *STUDENT*-tietokoneohjelma, joka ratkaisi algebrallisia sanaongelmia (Bobrow, Daniel Gureasko 1964).

Nämä ensimmäiset tekoälyohjelmat käyttivät samaa perusalgoritmiä ja toimivat yleisesti samalla tavalla. Ratkaistakseen jonkin ongelman ne etenivät askel kerrallaan yrittäen eri kombinaatioita, kunnes ratkaisu annettuun ongelmaan löytyi. Tämä ongelmanratkaisutapa tuotti tuloksia esimerkiksi James Staglen, joka oli myös Minskyn oppilas, luomassa *SAINT*-tietokoneohjelmassa, joka ratkaisi symbolisia integrointiongelmia käyttäen apunaan heuristisia sääntöjä (Stagle, James Robert 1961). Tekoälytutkimuksen alkuaikoina oletettiin tämän ongelmanratkaisu paradigman olevan skaalautuva myös suuremmille ongelmille, pian kuitenkin huomattiin tämän olevan virheellinen olettamus. Ohjelmat eivät kyenneet ratkaisemaan vaikeampia tai yksinkertaisesti laajempia ongelmia, johtuen muun muassa kombinatorisesta räjähdyksestä.

Tämän huomaaminen johti asiantuntijajärjestelmien syntymiseen 1960-luvun loppu puolella ja ne dominoivat 1970-luvun ja 1980-luvun tekoälytutkimusta. Nämä järjestelmät luotiin käyttäen jonkin alan asiantuntijoiden erikoistietämystä ja ne sisälsivät eräänlaisen tietopankin, joka oli erillinen ohjelman päättelyalgoritmistä. Ne erosivat aikaisemmista ohjelmista siinä, ettei niitä luotu yleispäteviksi ongelmanratkaisijoiksi, vaan pikemminkin rajattuihin alakohtaisiin käyttötarkoituksiin. 1980-luku näki myös kunnianhimoisen *Fifth Generation Computer Systems* -hankkeen, joka oli Japanin talous-, kauppa- ja teollisuusministeriön aloittama projekti, jonka kymmenen vuoden tavoite oli luoda älykkäitä koneita (Shapiro, Ehud 1983). Hanke ei lopulta onnistunut saavuttamaan sille asetettuja tavoitteita, mikä edelleen ohjasi tutkimusta kohti alakohtaisten ongelmien ratkaisemiseen. Tätä siirtymää pois yleisistä algoritmeistä on kritisoitu myöhempinä vuosina ja sitä pidetään osasyynä alan hitaal-

le kehitykselle ja esteenä alkuperäisen tavoitteen saavuttamiseen. Esimerkiksi John McCarthy ja Marvin Minsky ovat puoltaneet ja kannattaneet paluuta ihmistasoisen tekoälyn luomiseen (Minsky, Marvin & Singh, Push & Sloman, Aaron 2004).

Kokonaisuudessaan ala on keskittynyt tähän päivään saakka konkreettisiin erillisiä älykkyyden osa-alueita vaativiin tehtäviin, mutta myös joitain yksittäisiä, yleistä tekoälyä tavoittelevia projekteja on aloitettu. Ajatus yleisen tekoälyn taustalla on luoda systeemejä, jotka kykenevät laaja-alaiseen yleiseen älykkääseen toimintaan. Tämä ajatus pystytään jäljittämään tekoälytutkimuksen alkuajoille saakka, josta monen epäonnistumisen jälkeen päädyttiin tutkimukseen, joka laajalti keskittyy heikkoon tekoälyyn ja systeemeihin, jotka ylittävät ihmisen kyvykkyydet jossakin rajatussa tehtävässä, mutta eivät kykene yleistämään tätä kykyä muihin alueisiin tai tehtäviin. Yhteisymmärrys yleisen tekoälyn tutkijoilla tänä päivänä on, ettei näitä irrallisia systeemejä pystytä yhdistämään koherentiksi, yleistä älykkyyttä omaavaksi järjestelmäksi. Tästä johtuen alettiin muodostamaan ja toteuttamaan kognitiivisia arkkitehtuureja, joiden pohjalta yleinen tekoäly voisi rakentua. Arkkitehtuurien pohjalta voidaan luoda agentteja, joilla tarkoitetaan autonomisia entiteettejä, jotka toimivat jossakin ympäristössä saavuttaakseen tavoitteita, jotka kykenevät laaja-alaiseen ongelmanratkaisemiseen.

3 Konetietoisuus

Yleistä tekoälyä tutkiessa päädytään välttämättä tilanteeseen, jossa tietoisuuden rooli ja yhteys älykkääseen toimintaan nousee esille. Niiden suhde näyttäisi olevan hienovarainen, sillä on itsestäänselvää tietoisuuden olevan tarpeellinen älykkäällä toimijalla, mutta toisaalta esimerkiksi ihmistä tutkiessa, huomataan monen kognitiivisen prosessin, jotka mahdollistavat älykkään toiminnan, olevan tiedostamattomia. Tutkiakseen tätä yhteyttä syvemmin, täytyy älykkyydelle muodostaa käyttökelpoinen määritelmä, jonka avulla tietoisuutta pystytään käsittelemään. Lopulliseen yhteisymmärrykseen alan tutkijat eivät ole päässeet, mutta älykkyyden katsotaan yleisesti jakautuvan kahteen aspektiin (Wang, Pei & Goertzel, Ben 2012, s. 263). Ensimmäisenä on syntaktinen puoli, joka mielletään kyvykkyytenä manipuloida symboleita heuristiikkojen avulla. Toinen puoli liittyy älykkäämpiin ominaisuuksiin, joita tarvitaan merkityksen tuottamiseen ja symboleiden maadoittamiseen todellisuudessa.

Antonio Chellan ja Riccardo Manzottin mukaan ensimmäinen puoli pystytään toteuttamaan ilman tietoisuutta, sen sijaan tietoisuutta vaaditaan saavuttaakseen merkityksellinen älykäs toiminta (Chella, Antonio & Manzotti, Ricardo 2009), johon myös pyritään, kun yritetään luoda yleistä älykkyyttä osoittavia keinotekoisia systeemejä. Samankaltaisen kahtiajaon tekee myös David Chalmer, jonka näkökannan mukaan tietoisuuteen liittyy helppo ja vaikea ongelma (Chalmers, David John 1995, s. 4). Helppo ongelma liittyy ärsykkeiden kategorisointiin, ärsykkeisiin reagoimiseen, informaation integrointiin ja sisäisten tilojen esittämiseen. Toisaalta vaikeampi ongelma on selittää kokemuksen subjektiivinen ominaisuus. Tämän erottelun avulla pystytään kategorisoimaan yleistä älykkyyttä osoittavat systeemit niihin, jotka käyttäytyvät ikään kuin ne olisivat tietoisia ja sellaisiin, jotka kykenevät aitoon tietoiseen kokemukseen ja toimintaan. Tekoälytutkimuksen historian aikana jälkimmäinen kategoria on pitkälti sivuutettu, koska sitä on pidetty epäoleellisena. Muun muassa Alan Turing, joka muotoili kuuluisan turingin testin, kyseenalaisti aidon tietoisuuden merkityksen systeemissä, joka kykenisi ihmisen kaltaiseen älykkääseen

toimintaan ja käytökseen (Turing, Alan Mathison 1950).

Turing lähti alun perin selvittämään, pystyvätkö koneet ajattelemaan. Hän kuitenkin hylkäsi kysymyksen huomatessaan ajattelemisen olevan vaikeasti määriteltävissä ja sen sijaan ehdotti toisenlaista kysymystä, joka kuitenkin liittyy vahvasti edeltävään. Hän muotoili testin, jota on myöhempinä aikoina kutsuttu standardiksi Turingin testiksi, jossa kuulustelija pyrkii erottamaan koneen ja ihmisen toisistaan, niiden antamien vastausten perusteella. Testin tavoitteena on selvittää koneen kyky ihmisen kaltaiseen älykkääseen toimintaan, tekstipohjaisessa keskustelutilanteessa. On tärkeä huomata, ettei koneelta vaadita oikeita vastauksia, sen sijaan koneen kyky vastata ihmismäisesti on selvitettävänä. Kone läpäisee testin, jos kuulustelija ei kykene erottamaan sitä ihmisestä. On väitetty koneen saavuttavan ihmisen tasoisen älykkyyden, mikäli se läpäisee testin (Warwick, Kevin & Shah Huma 2016). Testi on saanut myös kritiikkiä, koska on väitetty vain symboleiden manipuloinnin, ilman laajempaa ymmärrystä käsiteltävästä asiasta, olevan tarpeellista, jotta testin voi läpäistä (Searle, John 1980).

Testiin on myöhempinä vuosina ehdotettu muutoksia, muun muassa Giulio Tononi muotoili uudenlaisen testin, joka pohjautuu informaation integrointiin (Tononi, Giulio 2004). Hänen mukaansa tietoisen kokemuksen aste määräytyy integroidun informaation määrän perusteella. Täten keinotekoisten systeemien kyvykkyyttä merkityksen luontiin pystytään testaamaan, selvittämällä niiden kyvykkyys sisäistää eri lähteistä poimittua tietoa, esimerkiksi laatimalla selityksiä kuvissa tapahtuville asioille. Keinotekoiset systeemit ovat edistyneet merkityksellisesti tässä tehtävässä, niin kuin myös luonnollisen kielen ymmärtämisessä ja tuottamisessa varsinkin viime aikoina ja molempia taitoja pidetään olennaisina vaatimuksina yleiselle älykkyydelle. Avoin kysymys nykyisellään vielä on, mikä yhteys näillä on tietoisuuteen ja kuinka merkityksellinen tietoisuus tekoälyssä on. Se on pyritty huomioimaan rajallisessa kapasiteetissa keinotekoisten systeemien luonnissa. Esimerkiksi Allen Newellin sisällytti tietoisuuden kriteerin hänen muodostamalle listalle, jota käytetään eräänlaisena mittapuuna agenttien, jotka perustuvat johonkin kognitiiviseen arkkitehtuuriin, arvioimisessa (Newell, Allen 1990).

4 Kognitiiviset arkkitehtuurit

Kognitiiviset arkkitehtuurit perustuvat teoriaan ja malliin ihmisen tiedonkäsittelystä ja mielestä (Lieto, Antonio & Bhattb, Mehul & Oltramaric, Alessandro & Vernond, David 2017). Näiden teorioiden pohjalta pystytään luomaan keinotekoisia agentteja, jotka kykenevät järkeilemään, oppimaan, toimimaan, suunnittelemaan toimintaansa, havainnoimaan ja sopeutumaan ympäristöönsä. Yleisen tekoälyn luominen kognitiivisten arkkitehtuurien avulla on ollut suosittu lähestymistapa, vaikka se aiheuttaakin tiettyjä ongelmia, sillä ihmisen osoittaman yleisen älykkyyden heterogeenisuus, tekee laaja-alaisen mittausjärjestelmän muodostamisen yleiselle tekoälylle lähes mahdottomaksi (Kotseruba, Iuliia & Tsotsos, John 2018). Vaikka älykkyyden mittaaminen tuottaakin ongelmia, voidaan siitä huolimatta selvittää vaatimuksia, jotka tällaisen kognitiiviseen arkkitehtuurin perustuvan agentin tulisi toteuttaa ja täten vertailla eri teorioita ja niiden instantiaatioita näitä kriteerejä vastaan.

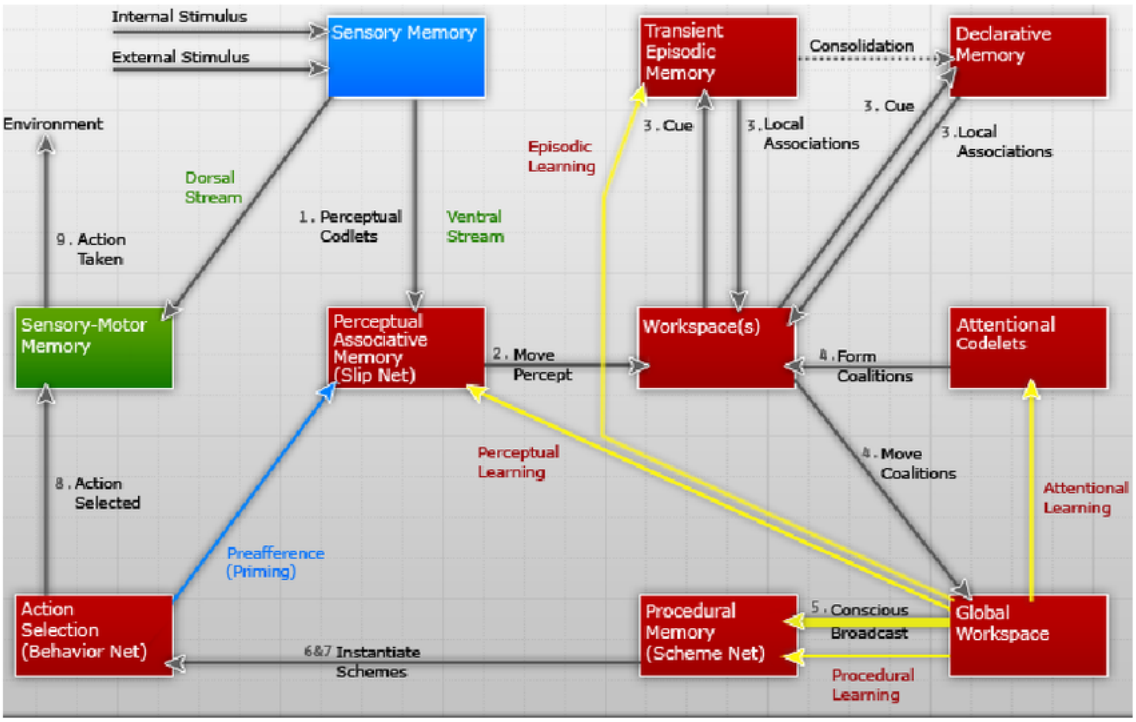
Eräs käyttökelpoiseksi todettu listaus kriteerejä on John R. Andersonin ja Christian Lebieren muodostama kokoelma (Anderson, John Robert & Lebiere, Christian 2003), joka perustuu alun perin Allen Newellin ehdottamiin kriteereihin (Newell, Allen 1990). Heidän mukaan johonkin kognitiivisen arkkitehtuurin perustuvan agentin täytyy osoittaa joustavaa käytöstä, jolla tarkoitetaan kykyä olla joustava toimenpiteiden valinnassa, reaaliaikaista toimintaa, sopeutuvaa käytöstä, jolla tarkoitetaan funktionaalista käytöstä reaalimaailmassa, laajaa tietokantaa ja kykyä hakea tietoa tehokkaasti ja toimia sen perusteella, dynaamista käytöstä, jolla tarkoitetaan kykyä reagoida vaihtuviin olosuhteisiin, uuden tiedon integrointia ja yhdistämistä aikaisempaan, luonnollisen kielen ymmärtämistä, tietoisuutta, moninaista oppimista ja kehitystä, jolla tarkoitetaan kykyä hankkia toimivuutta ajan myötä, ollakseen funktionaalinen. Tietoisuuden kriteeri on kiistelty lisäys tällä listalla, jonka Anderson ja Lebiere ovat itsekin kyseenalaistaneet (Anderson, John Robert & Lebiere, Christian 2003, s. 590). Muitakin listauksia on muodostettu, esimerkiksi John Laird julkaisi yhdessä Robert Wrayn kanssa pohjan yleiselle tekoälylle ja sen vaatimuksille (Laird, John & Robert, Wray 2010). Kokoelmat kriteerejä sisältävät yleisesti monia

samankaltaisuuksia, mistä johtuen tulen käsittelemään kognitiivisia arkkitehtuureja pääasiassa Andersonin ja Lebieren näkökulmasta.

4.1 *LIDA* (*Learning Intelligent Distribution Agent*)

Kognitiivisia arkkitehtuureja yleiselle tekoälylle on luotu useita alan historian aikana (Kotseruba, Iuliia & Tsotsos, John 2018). Eräs tunnetuimmista on Stan Franklinin *LIDA*-arkkitehtuuri (Franklin, Stan & Patterson, Jr. 2006). Kognitiivinen malli, johon *LIDA* pohjautuu, on täysin integroitu keinotekoinen kognitiivinen systeemi, joka kykenee alhaisen tason havainnoinnista/toiminnasta korkean tason päättelyyn (Wang, Pei & Goertzel, Ben 2012, s. 105). Sen toiminta perustuu kognitiiviseen sykliin, jossa se havainnoi sisäisen ja ulkoisen ympäristönsä, luo merkityksen ympäristössä havaituille asioille ja päättää mitä tehdä seuraavaksi (Ramamurthy, Uma & Baars, Bernard & D'Mello Sidney & Franklin, Stan 2006). Lisäksi se oppii kokemuksistaan selvittämällä uusien havaittujen objektien suhteita, sille entuudestaan tunnettuihin objekteihin. Se pyrkii myös selvittämään toiminnan ja objektien välisiä riippuvuuksia. Kuviossa 1 näkyy *LIDA*-arkkitehtuuri ja sen koostumus.

LIDA-arkkitehtuuri voidaan jakaa viiteen osaan. Ensimmäisenä on aistipohjainen assosiatiivinen muisti (*Perceptual Associative Memory*), joka mallintuu neurologisesti ihmisen aivokuoren osiin. Tämä muisti mahdollistaa sisäisen ja ulkoisen informaation erottelun, luokittelun ja tunnistamisen. Sen avulla pystytään myös oppimaan uusia objekteja, luokkia ja suhteita lisäämällä ja vahvistamalla yhteyksiä niiden ja aikaisemman tiedon välillä muistin sisällä. Toisena on työtila (*Workspace*), joka suunnilleen vastaa ihmisen työmuistin esitietoista puskuria. Se sisältää havaitut struktuurit ja lokaalit assosiaatiot episodisista muisteista, joiden avulla se luo mallin ympäristön tämän hetkisestä tilasta. Kolmantena on episodiset muistit (*Episodic Memories*), jotka tallentavat tapahtumia ja niiden yksityiskohtia. Tapahtumat tallennetaan ensin lyhytaikaiseen episodiseen muistiin (*Transient Episodic Memory*), josta ne saatetaan vahvistaa deklaratiiviseen muistiin (*Declarative Memory*), joka on pitkäaikainen, mikäli ne koetaan merkityksellisiksi. Neljäntenä on tarkkaavaisuus muisti (*Attentional Memory*), jonka tarkoitus on etsiä työtilan luomasta ympäristön



Kuvio 1. *LIDA*-mallin kognitiivinen sykli (Friedlander, David Samuel & Franklin Stan 2008)

mallista erityisiä ärsykkeitä. Se etsii esimerkiksi uhkia ja niitä löydettyään lisää ne globaaliin työtilaan (*Global Workspace*), jossa niiden vakavuuden mukaan muodostetaan toimenpiteitä. Viidentenä on proseduraalinen muisti (*Procedural Memory*), toimenpiteen valinta ja sensomotorinen muisti (*Sensory-motor Memory*). Proseduraalinen muisti käsittelee tulevia toimenpiteitä, muodostamalla mallin toimista ja niiden seurauksista, jonka se välittää toimenpiteen valinta mekanismiin, joka vuorostaan valitsee toimenpiteen, jonka sensomotorinen muisti toteuttaa.

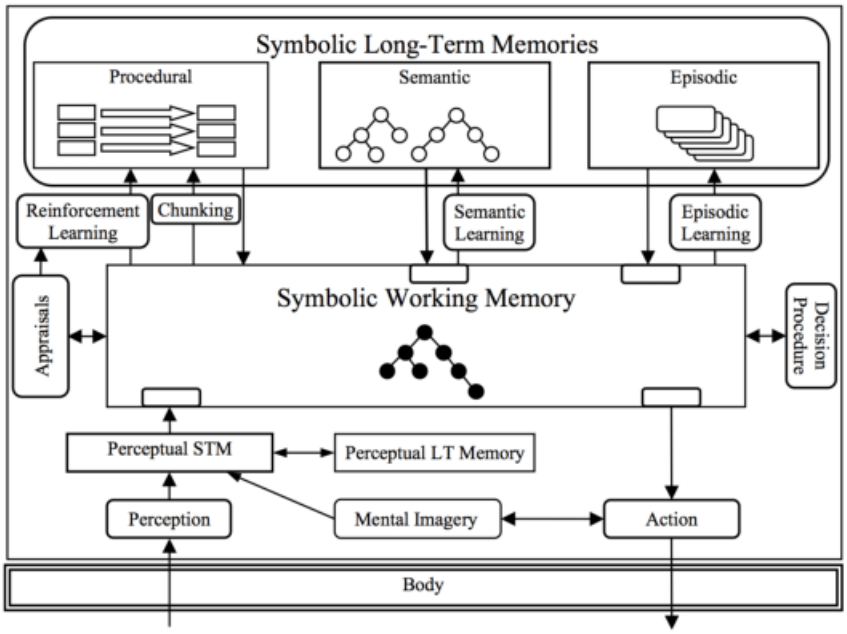
LIDA-arkkitehtuuria voidaan myös tarkastella Andersonin ja Lebieren kriteerien nojalla (Anderson, John Robert & Lebiere, Christian 2003), selvittääkseen onko se käyttökelpoinen lähtökohta yleiselle tekoälylle. Heidän mukaan kognitiivisen arkkitehtuurin tulee osoittaa joustavuutta toimenpiteiden valinnassa. Tämä toteutuu, kun huomioidaan *LIDA*-arkkitehtuurin mahdollistama kyvykkyys oppimiseen ja toiminnan mukautuminen ärsykkeiden johdosta (Ramamurthy, Uma & Baars, Bernard & D'Mello Sidney & Franklin, Stan 2006). Lisäksi sen täytyy toimia reaaliajassa

reaalimaailmassa, mikä toteutuu Usef Faghihin ja Stan Franklinin mukaan (Wang, Pei & Goertzel, Ben 2012, s. 112). Arkkitehtuuri osoittaa myös kyvykkyyttä suurten tietomäärien hallintaan, tiedonhakuun ja oppimiseen, kun tarkastellaan kuinka se säilyttää informaatiota ja valitsee toimenpiteitä. Luonnollisella kielellä toimiminen on osa-alue, joka *LIDA*-arkkitehtuurin perustuvalle agentille täytyy erikseen opettaa. Kuten huomataan, *LIDA*-arkkitehtuuri täyttää Andersonin ja Lebieren asettamat kriteerit funktionaaliselle kognitiiviselle arkkitehtuurille ja on täten sopiva lähtökohta yleisen tekoälyn muodostamiselle. Tulen vertailemaan sitä Soar-arkkitehtuurin kanssa, jota pidetään toisena varteenotettavana mallina yleiselle tekoälylle.

4.2 *Soar*

Soar on John Lairdin, Allen Newellin ja Paul Rosenbloomin 1980-luvulla luoma kognitiivinen arkkitehtuuri (Laird, John & Newell, Allen 1983). Tutkijat pyrkivät luomaan agentin, joka pystyy suorittamaan laajan kirjon tehtäviä ja toteuttamaan kaikki kognitiiviset kyvykkyydet, joita ihmiseltä löytyy. *Soar* on samalla teoria siitä, mitä kognitio on ja laskennallinen implementaatio tämän pohjalta. Sen alkuperäinen toimintaperiaate pohjautui muutamaan yksinkertaiseen hypoteesiin. Tutkijat ajattelivat sääntöjen olevan riittäviä esittämään kaiken pitkäaikaisen tiedon, yhden oppimismekanismin olevan riittävä kaikelle oppimiselle ja symbolisten esitysten olevan riittäviä kaikelle lyhytaikaiselle ja pitkäaikaiselle tiedolle (Laird, John 2012). Arkkitehtuuria on myöhemmin laajennettu sisältämään monia ominaisuuksia, jotka ovat ominaisia moderneille kognitiivisille arkkitehtuureille, kuten useat pitkäaikaismuistit ja erilaiset oppimismekanismit. Tavoitteena projektissa on ollut saavuttaa laajamittainen teoria ja arkkitehtuuri yleispäteville agenteille. *Soar*-arkkitehtuurin rakenne näkyy kuviossa 2.

Arkkitehtuurin toiminta perustuu samankaltaiseen kognitiiviseen sykliin, kuin *LIDA*-arkkitehtuurin toiminta. *Soar* havainnoi ympäristönsä, josta havaitut objektit siirtyvät symboliseen työmuistiin (*Symbolic Working Memory*). Tämä alue sisältää arkkitehtuurin käsityksen nykyisestä tilanteesta, johon pitkäaikaismuistit tuovat in-



Kuvio 2. Soar-arkkitehtuuri (Laird, John 2012)

formaatiota, sen ollessa tarpeellista. Symbolinen työmuisti sisältää myös puskurin, johon toiminnat, joita halutaan toteuttaa, kiinnitetään siirrettäväksi aktuaattoreille. *Soar*-arkkitehtuuri sisältää useita muisteja, jotka tallentavat erilaista tietoa. Proseduraalinen muisti (*Procedural Memory*) koostuu säännöistä, jotka kertovat mitä seurauksia toiminnat aiheuttavat. Nämä säännöt kertovat agentille, mitä toimintoja täytyy suorittaa, mikäli *Soar* havaitsee tämän hetkisessä tilassaan proseduraalisesta muistista löytyviä ehtoja. Tämän yhteydessä agentti oppii, muokkaamalla ja luomalla uusia sääntöjä proseduraaliseen muistiin, toimintojen seurausten perusteella. Arkkitehtuuri sisältää proseduraalisen muisti lisäksi semanttisen muistin (*Semantic Memory*), joka tallentaa yleistä tietoa maailmasta, kuten tiedon siitä, että koirat ovat eläimiä. Se pitää sisällään myös episodisen muistin (*Episodic Memory*), johon tallennetaan tiettyjä instansseja työmuistissa samaan aikaan esiintyvistä struktuureista, jotka mahdollistavat kokemusten kontekstien ja kokemusten ajallisten suhteiden muistamisen. *Soar*-arkkitehtuuria tutkittaessa on huomattu tämän muistin mahdollistavan monia korkean tason kognitiivisia toiminnallisuuksia, kuten tilanteiden sisäisen simuloinnin ja toimintamallien oppimisen. Symbolisen työmuistin luomasta ympäristön mallista etsitään aikaisemmin havaittuja objekteja, struktuureja ja ehto-

ja, joiden pohjalta päätetään, mitä toimintoja seuraavaksi toteutetaan. Symbolinen työmuisti sisältää erillisen valintamekanismin, joka evaluoi ja lopulta päättää mitä tehdä.

Soar-arkkitehtuuri toteuttaa osittain Andersonin ja Lebieren asettamat kriteerit funktionaaliselle kognitiiviselle arkkitehtuurille (Anderson, John Robert & Lebiere, Christian 2003). Arkkitehtuuri mahdollistaa monenkaltaista oppimista eri tiedonkäsittelyn vaiheissa, kuten eri tapahtumien välisten suhteiden ymmärtämistä episodisen muistin avulla, mikä tarkoittaa, että se kykenee joustavaan, dynaamiseen ja sopeutuvaan käytökseen opitun informaation mukaan. Proseduraalinen ja semanttinen muisti mahdollistavat uuden tiedon integroinnin ja laajojen tietokantojen muodostamisen. Lisäksi eri muistit tukevat pitkäaikaista oppimista ja kehitystä. Arkkitehtuuri ei toteuta luonnollisella kielellä toimimista, eikä se ota kantaa tietoisuuden kriteerin, mutta siitä huolimatta se on potentiaalinen lähtökohta, yleistä älykkyyttä osoittavalle agentille (Laird, John 2012).

Kun arkkitehtuureja vertaillaan, huomataan niiden muistuttavan etäisesti toisiaan. Niiden toiminnassa ja rakenteessa ilmenee monia samankaltaisuuksia, kuten useat pitkäaikaismuistit, jotka tallentavat erilaista tietoa, mutta ne eroavat myös paikoittain toisistaan merkityksellisesti. Molemmat arkkitehtuurit tallentavat ympäristössä havaittua informaatiota symbolisesti. Eroavaisuutena arkkitehtuurien välillä on subsymbolinen aktivaatiomekanismi *LIDA*-arkkitehtuurissa, joka määrää ärsykkeisiin reagoimisen, niiden ollessa tarpeeksi merkityksellisiä tarkkaavaisuus muistin mukaan. *Soar*-arkkitehtuuriin on lisätty sen kehityksen aikana monia muissa arkkitehtuureissa hyväksi havaittuja ominaisuuksia, mistä johtuen arkkitehtuurit ovat nykyisellään samankaltaisia. Arkkitehtuurien luonti ja laskennallisten agenttien toteutus on auttanut tutkijoita luomaan kattavampia malleja ihmisen tiedonkäsittelystä ja mielestä, mutta nykyiset mallit ovat vieläkin keskeneräisiä ja usein ristiriitaisia keskenään (Duch, Wlodzislaw & Oentaryo, Richard & Pasquier Michel 2008).

5 Yhteenveto

Tutkimuksessa tarkasteltiin yleistä tekoälyä ja sen rakennetta, vertailemalla sitä arkipäiväisempään heikkoon tekoälyyn ja selvittämällä mistä lähtökohdista se on lähtenyt muodostumaan. Huomattiin yleistä älykkyyttä osoittavien koneiden olevan vanha tavoite, jota tekoälytutkimus ei monien tutkijoiden mielestä ole merkityksellisesti lähestynyt, vaikka ala kokonaisuudessaan tuntuu kehittyvän vilkasta vauhtia. Tutkimuksessa selvitettiin alan historian kehitys viime vuosituhannen puolivälistä nykypäivään saakka, samalla tarkastelemalla minkälaisia pyrkimyksiä tutkijoilla oli ja mihin aihealueeseen ala oli kokonaisuudessaan milloinkin keskittynyt.

Älykkyyden ja tietoisuuden suhdetta selvitettiin siinä kapasiteetissa, missä ne liittyvät keinotekoisiin systeemeihin ja yleiseen älykkyyteen, käyttäen apunaan eri tutkijoiden näkökulmia ja yhteisymmärryksiä. Huomattiin niiden olevan haasteellisia ongelmia, johtuen muun muassa eri määritelmien ja ominaisuuksien eroavaisuuksista, sekä tutkijoiden erimielisyyksistä. Tutkimuksessa saatiin kuitenkin selvitettyä toiminnallinen kuvaus älykkyydelle, jonka pohjalta kognitiivisia arkkitehtuureja pystyttiin tarkastelemaan ja vertailemaan. Huomattiin yleisen tekoälyn tutkijoilla olevan laajalti erilaiset lähtökohdat systeemisuunnittelussa ja tavoitteissa koneiden toiminnallisuuksille, vaikka joitakin standardeja arkkitehtuurien analysoimiseen onkin muodostunut.

Kognitiivisia arkkitehtuureja on alan historian aikana suunniteltu ja toteutettu useita. Tutkimuksessa perehdyttiin kahteen sellaiseen arkkitehtuurin, jotka osoittavat potentiaalia toimiakseen pohjana yleiselle älykkyydelle. Arkkitehtuurien välillä huomattiin monia samankaltaisuuksia, vaikka joitakin eroja myös löydettiin. Selvisi näiden arkkitehtuurien perustuvan kognitiiviseen sykliin, jonka aikana ne havainnoivat ympäristönsä, päättelevät havaintojen perusteella mitä tehdä seuraavaksi ja lopulta toimivat tämän päätelmän mukaan. Syklin aikana tapahtuu paljon muutakin, kuten erilaista oppimista ja arviointia, mutta arkkitehtuurin toiminta perustuu yksinkertaisuudessaan tähän kolmiosaiseen prosessiin. Jatkotutkimuksena olisi kiinnostavaa selvittää minkälaisia agentteja näiden arkkitehtuurien perusteella on to-

teutettu ja selvittää kykenevätkö ne käytännössä samoihin toiminnallisuuksiin, kuin mitä ne teoriassa lupaavat.

Kirjallisuutta

Amunts, K. & Ebell, C. & Muller, J. & Telefont, M. & Knoll, A. & Lippert, T. 2016. *The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain*. Neuron vol. 92. no. 3. s. 574–581.

Anderson, J. R. & Lebiere, C. 2003. *The Newell Test for a theory of cognition*. Behavioral and brain sciences vol. 26. s. 587-601.

Bobrow, D. G. 1964. *Natural Language Input for a Computer Problem Solving System*.

Brachman, R. J. 2006. *AI – more than the sum of its parts*. AI Magazine vol. 27. no. 4. s. 19–34.

Chalmers, D. J. 1995. *The Conscious Mind in Search of a Theory of Conscious Experience*. Väitöskirja, University of California, Santa Cruz.

Chella, A. & Manzotti, R. 2009. *Machine Consciousness: A Manifesto for Robotics*. International Journal of Machine Consciousness vol. 1. no. 1. s. 33-51.

Duch, W. & Oentaryo, R. & Pasquier, M. 2008. *Cognitive Architectures: Where Do We Go from Here?*. Frontiers in Artificial Intelligence and Applications vol. 171. s. 122-136.

Feigenbaum, E. A. & Feldman, J. 1963. *Computers and Thought*. McGraw-Hill, New York.

Franklin, S. & Patterson Jr. F. G. 2006. *The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent*. Integrated Design and Process Technology, IDPT-2006. Society for Design and Process Science.

Friedlander, D. S. & Franklin, S. 2008. *LIDA and a theory of mind*. Frontiers in Artificial Intelligence and Applications vol. 171. s. 137-148.

Gugerty, L. 2017. *Newell and Simon's Logic Theorist: Historical Background and Impact on Cognitive Modeling*. Human Factors and Ergonomics Society Annual Meeting Proceedings vol. 50. no. 9. s. 880-884.

Irving, G. & Askell, A. 2019. *AI Safety Needs Social Scientists*. Distill.

Kotseruba, I. & Tsotsos, J. 2018. *A Review of 40 Years in Cognitive Architecture Research Core Cognitive Abilities and Practical Applications*. Artificial Intelligence Review.

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. Viking

Press.

Laird, J. & Newell, A. 1983. *A Universal Weak Method : Summary of results*. Proceedings of the Eighth International Joint Conference on Artificial Intelligence - vol. 2. s. 771-773.

Laird, J. & Wray, R. 2019. *Cognitive Architecture Requirements for Achieving AGI*. Proceedings of the 3d Conference on Artificial General Intelligence. s. 79-84.

Laird, J. 2012. *The Soar Cognitive Architecture*. AISB Quarterly. No. 134.

Lieto, A. & Bhattb, M. & Oltramaric, A. & Vernond, D. 2017. *The Role of Cognitive Architectures in General Artificial Intelligence*. Cognitive Systems Research vol. 48. s. 1-3.

McCarthy, J. & Minsky, M. L. & Rochester, N. & Shannon C. E. 1955. *A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence*. AI Magazine vol. 27. no. 4. s. 12-14.

McCarthy, J. 1961. *Programs With Common Sense*. Computation and intelligence. s. 479-492.

Minsky, M. & Push, S. & Sloman, A. 2004. *The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence*. AI Magazine vol. 25. no. 2. s. 113-124.

Newell, A. & Shaw, J. F. & Simon, H. A. 1959. *Report on a general problem-solving program*. International Conference on Information Processing. s. 256-264.

Newell, A. 1990. *Unified Theories of Cognition*. Harvard University Press.

OpenAI. 2018. *OpenAI Charter*. Haettu 7.7.2019 osoitteesta (`https://openai.com/charter/`)

Ramamurthy, U. & Baars, B. & D'Mello S. & Franklin, S. 2006. *LIDA: A Working Model of Cognition*. Proceedings of the 7th international conference on cognitive modeling. s. 244–249.

Russell, S. J. & Norvig, P. 2010. *Artificial Intelligence: A Modern Approach 3rd Edition*. s. 20-21.

Searle, J. 1980 *Minds, Brains and Programs*. Behavioral and Brain Sciences vol. 3. no. 3. s. 417–457.

Shapiro, E. 1959. *The Fifth Generation Project - Trip Report*. Communications of the

ACM vol. 26. no. 9. s. 637-641.

Stagle, J. S. 1961. *A Heuristic Program That Solves Symbolic Integration Problems In Freshman Calculus, Symbolic Automatic Integrator (SAINT)*. Väitöskirja, Massachusetts Institute of Technology.

Tononi, G. 2004. *An information integration theory of consciousness*. BMC Neuroscience vol. 5. no. 42.

Turing, A. M. 1950. *Computing Machinery And Intelligence*. Mind vol. 49. s. 433-460.

Wang, J. & Kurth-Nelson, Z. & Tirumala, D. & Soyer, H. & Leibo, J. & Munos, R. & Blundell, C. & Kumaran, D. & Botvinick, M. 2017. *Learning to Reinforcement Learn*. CogSci 2017 Proceedings.

Wang, P. & Goertzel, B. 2012. *Theoretical Foundations of Artificial General Intelligence*. Atlantis Press. s. 2-263.

Warwick, K. & Shah, H. 2016. *Passing the Turing Test Does Not Mean the End of Humanity*. Cognitive Computation vol. 8. no. 3. s. 409–419.