Joose Tikkanen

Älykkäiden agenttiarkkitehtuurien soveltaminen vaativammissa tehtävissä

Tietotekniikan kandidaatintutkielma

30. huhtikuuta 2020

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Joose Tikkanen

Yhteystiedot: jopitikk@student.jyu.fi

Ohjaaja: Sanna Juutinen

Työn nimi: Älykkäiden agenttiarkkitehtuurien soveltaminen vaativammissa tehtävissä

Title in English: Application of intelligent agent architectures in more demanding tasks

Työ: Kandidaatintutkielma

Sivumäärä: 22+0

Tiivistelmä: Tutkielmassa perehdytään ohjelmistoagentin määritelmään, toimintaperiaatteeseen ja rationaalisuuteen ja erilaisiin agenttiarkkitehtuureihin, joita on kehitetty. Näiden valossa myös pohditaan tekoälyn kykyjä vaativammissa tehtävissä, ja mitä eettisiä ongelmia ja riskejä sen kehitykseen liittyy. Tutkielman aiheisiin perehdytään kattavan kirjallisuuskatsauksen avulla tutkimalla useita aihetta käsitteleviä vertaisarvioituja artikkeleja. Eri lähteiden ja niihin pohjautuvien päätelmien valossa tekoälyä ei tulisi valjastaa vaativampiin tehtäviin, joissa sen tulisi muun muassa pohtia eettisiä valintoja.

Avainsanat: tekoäly, älykäs agentti, agenttiarkkitehtuuri, agenttiohjelma, moraalinen agentti, kognitiivinen arkkitehtuuri, kone-etiikka, tekoälyn tietoisuus

Abstract: The thesis examines the definition of a software agent, its operating principle and rationality, and various agent architectures that has been developed. In the light of these, we also reflect on the capabilities of artificial intelligence in more demanding tasks, and the ethical issues and the risks that arises from its development. The topics of the thesis are examined through a comprehensive literature review by studying several peer-reviewed articles on the subject. In the light of different sources and the conclusions drawn from them, artificial intelligence should not be harnessed for more demanding tasks, where it would have to e.g. consider ethical choices.

Keywords: artificial intelligence, intelligent agent, agent architecture, agent program, moral agent, cognitive architecture, machine ethics, AI and consciousness

Kuviot

Taulukot

Sisältö

1 Johdanto

Tekoälytutkimus on jatkuvasti kehittyvä ala, jossa uudet innovaatiot vaikuttavat merkittävästi ihmisten elämiin. Tässä tutkielmassa perehdytään agentin käsitteeseen ja ohjelmistoagenttien rationaalisuuteen, erilaisiin agenttiarkkitehtuureihin, kuinka haastaviin ja monimutkaisiin tehtäviin nämä agenttiarkkitehtuurit voidaan mahdollisesti valjastaa, ja mitä mahdollisia ongelmia tekoälykehitykseen liittyy tekoälytutkimuksen kirjallisuuden valossa. Pienet saavutukset tekoälykehityksessä ovat jo vaikuttaneet siihen, kuinka tietojenkäsittelytiedettä opetetaan ja harjoitetaan, ja mahdollistanut muun muassa poikkeuksellisen tehokkaiden hakukoneiden ja kasvojentunnistussovellusten kehityksen (Russell ja Norvig 2016). On selvää, että keskikokoiset saavutukset tulevat jollain tavalla vaikuttamaan kaikkien ihmisten elämiin. Esimerkiksi robottien valjastamisella toimistoapulaisiksi voi olla suuri positiivinen vaikutus usealle ihmiselle, ja automatisoitu ajoapulainen voi pelastaa tuhansia ihmishenkiä.

Toisaalta näilläkin innovaatioilla voi olla ikäviä vaikutuksia lyhyellä aikavälillä esimerkiksi talouteen ja työllisyyteen. Suurissa tekoälyinnovaatioissa tieteisfiktio suosii dystopisia tulevaisuudennäkymiä luultavasti mielenkiintoisten juonien vuoksi, mutta positiivisten mahdollisuuksien kirjo on myös suuri. Muun muassa täydellisen rationaalista agenttia on realistisesti mahdotonta saavuttaa suuren laskentavaativuuden takia (Russell 1997), joten on syytä selvittää, mitä muita perustavanlaatuisia ongelmia tekoälykehityksessä tulee vastaan, mihin asti tekoälykehityksessä voidaan päästä ja mitä eettisiä seurauksia tulee ottaa huomioon.

Agenttiarkkitehtuurien tutkimus auttaa ymmärtämään ihmisen tietoisuutta ja kognitiivisia prosesseja, ja tietoisuuden ja kognition tutkiminen auttaa agenttiarkkitehtuurien kehityksessä (Franklin, Kelemen ja McCauley 1998). Tutkielmassa tarkastellaan eri agenttiarkkitehtuureja ja pohditaan, voidaanko tekoälylle kehittää rationaalisuutta ja älykkäitä agenttiarkkitehtuureja hyödyntäen inhimillinen tietoisuus ja ymmärrys, jotta se voidaan valjastaa nykyistä vaativampiin tehtäviin, jotka ovat toistaiseksi ihmisten hallussa. Tekoälykehitystä tutkitaan eri agenttiarkkitehtuurien näkökulmasta, mitä erityyppisillä arkkitehtuureilla voidaan saavuttaa ja mitä ongelmia niiden kehityksessä on tullut. Tekoälyn valjastamisen mahdollisuus erilaisiin luottamustehtäviin, kuten päättäjiksi, tuomareiksi, johtajiksi, lakien säätäjiksi ja lääkäreiksi, on ajankohtainen ja mielenkiintoinen aihe, jonka mahdollisuuksia ja siihen

liittyviä riskejä on syytä pohtia.

Tutkimusstrategiana ja -metodina toimii pienimuotoinen kirjallisuuskartoitus. Tekoälykehitystä on tutkittu paljon parin viime vuosikymmenen aikana useasta eri näkökulmasta: sekä filosofisemmista näkökulmista, kuten kone-etiikkaa ja tekoälyn tietoisuutta, että teknisemmistä näkökulmista, kuten koneoppimista ja taustalla olevaa matematiikkaa. Aiheeseen liittyvää kirjallisuutta löytyy laajasti internetistä muun muassa Jyväskylän yliopiston kirjaston- ja Googlen tietokannoista. Lisäksi tutkielmassa hyödynnetään yhtä modernin tekoälytutkimuksen merkittävää kirjaa, *Artificial intelligence: a modern approach* (Russell ja Norvig 2016).

Luvussa 2 käsitellään agentin ja ohjelmistoagentin määritelmiä ja toimintaperiaatetta sekä ohjelmistoagenttien rationaalisuutta. Luvussa 3 käsitellään agenttiarkkitehtuureja ja niiden luokittelua eri tyyppien mukaan, ja lisäksi katsotaan tarkemmin kahta kehittyneempää hybridiarkkitehtuuria, ja mitä tuloksia niiden kehitykset ovat saavuttaneet. Luvussa 4 tutkitaan tekoälyagenttien potentiaalia tekoälytutkimuksen kirjallisuuden perusteella ja pohditaan myös niiden kehitykseen liittyviä eettisiä ongelmia ja riskejä. Lopuksi luvussa 5 selvitetään yhteenveto tutkimuksen menetelmistä ja tuloksista.

2 Älykkäät agentit

Tekoälytutkimuksessa yksi keskeisimmistä käsitteistä on älykäs agentti tai -toimija, tarkemmin älykäs ja autonominen ohjelmistoagentti. Tässä luvussa avataan agentin käsitettä ja pohditaan ohjelmistoagenttien rationaalisuutta.

2.1 Agentin määritelmä ja toimintaperiaate

Tutkijat ovat antaneet useita eripituisia määritelmiä käsitteelle agentti, joista toiset ovat vaativampia ja toiset yksinkertaisempia. Oleellisesti näitä määritelmiä yhdistää kaksi tekijää: Agentti on 1) jokin, joka toimii, tai jokin, joka voi toimia, tai 2) jokin, joka toimii jonkin toisen sijasta luvan kanssa (Franklin ja Graesser 1996). Tarkemmin autonomisen agentin olemuksen voi kiteyttää seuraavasti:

”*Autonominen agentti on systeemi, joka sijaitsee jossain ympäristössä ja osana ympäristöään, ja joka aistii ympäristönsä ja toimii sen perusteella, ja ajan kuluessa tavoittelee oman toimintasuunnitelmansa toteutumista ja vaikuttaa siihen, mitä se aistii tulevaisuudessa* (Franklin ja Graesser 1996, s. 25).”

Tarkennuksena tässä tutkimuksessa agenteista puhuttaessa tarkoitetaan erityisesti autonomista ja älykästä ohjelmistoagenttia. Autonomisuudella tarkoitetaan, että agentti on itsenäinen ja tekee itse päätöksensä, mikä erottaa agentit esineistä. Ohjelmisto-etuliitteellä korostetaan kyseen olevan tietokoneohjelmista, eikä yleisistä agenteista, joita esimerkiksi ihmiset ja eläimet ovat. Älykäs ohjelmistoagentti voidaan määritellä taulukon 1 mukaisesti.

Älykkään agentin täytyy pystyä havaitsemaan ympäristöään yksityiskohtaisesti, jonka perusteella se voi visioida useita mahdollisia muutoksia, erityisesti niitä muutoksia, joita se voi itse tehdä. Lisäksi sen täytyy pystyä erottamaan nämä muutokset tapauksiin, jotka johtavat haluttuihin lopputuloksiin ja tapauksiin, jotka eivät. Yksi merkittävimmistä luonnon saavutuksista on luonnollisten aivojen kyky prosessoida informaatiota kehonsa ja ympäristönsä kontekstissa käyttäen resurssejaan hyvin tehokkaasti (Duro, Bellas ja Permuy 2014). Yksinkertaiset ongelmat analogisessa esitysmuodossa voivat olla toisaalta yllättävän haasta-

Älykäs agentti on ohjelmisto, joka sijaitsee ympäristössään, ja lisäksi on	
Autonominen	Itsenäinen ja ei ulkoisesti ohjattava
Reaktiivinen	Vastaa ympäristönsä muutoksiin sopivassa ajassa
Proaktiivinen	Tavoittelee jatkuvasti päämääriään
Joustava	Hallitsee useita keinoja tavoitteidensa saavuttamiseksi
Vakaa	Toipuu epäonnistumisista
Sosiaalinen	Kykenee toimimaan muiden agenttien kanssa

Taulukko 1. Älykkään agentin määritelmä (Padgham ja Winikoff 2005)

via ohjelmistoagenteille. Esimerkiksi koira on tarpeeksi älykäs kiertämään aidan hakeakseen ruuan aidan toiselta puolelta, mutta ohjelmistoagentille täytyy antaa oikeanlainen esitysmuoto ongelmalle, jotta sen ratkaiseminen on tarpeeksi helppoa. Vähimmäisvaatimus älykkään agentin toimimiselle ympäristössään on kyky harkita ympäristönsä muutoksia tarpeeksi sulavassa tilojen sarjassa (Sloman 1971). Toistaiseksi älykkäät agentit ovat toimineet enimmäkseen kulissien takana, esimerkiksi luottokorttiostosten automatisoidussa vahvistamisessa nettikaupoissa, keskiverto kuluttajalta piilossa (Russell ja Norvig 2016). Tämänkaltaisissa tehtävissä ongelma on esitetty diskreetissä ja äärellisessä tilalaitteessa, johon aritmeettiset ongelmanratkaisumenetelmät toimivat hyvin. Ihmisille ja koirille yksinkertaisten analogisten ongelmien esittäminen ohjelmistoagentille vaatii uudenlaisten systemaattisten kaavojen löytämistä tai keksimistä, joilla kukin ongelma voidaan organisoida diskreettiin ja aritmeettisesti laskettavaan muotoon (Sloman 1971).

Agentin toiminta riippuu täysin sen nykyisistä havainnoistaan ja havainnointihistoriastaan (eng. *percept sequence*) nykyhetkeen saakka. Ohjelmistoagentin havainnointia voidaan luonnehtia jonkinlaisen laskentasarjan tulokseksi, joka muuttaa sensorien signaaleja tietämykseksi entiteeteistä, tapahtumista ja maailman tilanteista (Arsene ja Dumitrache 2017). Agenttifunktio määrittää toiminnan jokaista mahdollista havainnointihistoriaa kohtaan (Russell ja Norvig 2016). Agenttifunktio voidaan esittää taulukkona, jossa on kaikki mahdolliset havainnointihistoriat ja niitä vastaavat toiminnat. Tämä taulukko voi olla luonnollisesti äärettömän pitkä, jos huomioitavien havainnointihistorioiden määrää ei rajoiteta. Agenttiohjelman tehtävä on implementoida agenttifunktio agentin sisäisesti ohjelmakoodissa (Russell ja Nor-

vig 2016). Agenttifunktio kuvailee agentin toimintaa ulkoisesti, ja agenttiohjelma määrittää agentin toiminnan sisäisesti.

2.2 Rationaaliset agentit

Ohjelmistoagentteja suunnitellessa täytyy pohtia agenttien rationaalisuutta. Rationaalisen agentin tulisi osata valita oikea toiminta jokaisessa agenttifunktion taulukon kohdassa. Oikean toiminnan valitseminen ei ole toisaalta yksiselitteistä, vaan vaatii perusteluja. Älykkään ohjelmistoagentin rationaalisuuden periaate on seuraava: Jos agentilla on tietoa siitä, että yksi sen mahdollisista toiminnoista johtaa haluttuun lopputulokseen, niin agentti valitsee tämän vaihtoehdon (Newell ym. 1982). Tämä periaate asettaa yleispätevän yhteyden agentin tietoudelle, tavoitteille ja jokaiselle toimintamahdollisuudelle tarkentamatta yksityiskohtia mekanismista, jolla yhteys tehdään. Toisaalta agentin rationaalinen toiminta voidaan määritellä sen havainnointihistorioiden perusteella: Jokaista mahdollista havainnointihistoriaa kohden agentin tulee valita toiminta, jonka odotetaan saavuttavan maksimaalinen suoritus saatavilla olevien todisteiden ja sisäisen tiedon perusteella (Russell ja Norvig 2016).

Rationaalinen päätöksenteko ei tapahdu tyhjiössä, vaan agentin pohdintatyössä täytyy huomioida reaalimaailman rajoitteet, kuten ohjelmistoagenttien vaatimat laskennalliset resurssit (Parkes ja Wellman 2015). Laskennallisia resursseja ovat muun muassa pohdintatyöhön vaadittava aika ja muisti. Äärellisten laskennallisten resurssien vuoksi täydellistä rationaalisuutta ei ole mahdollista saavuttaa ohjelmistoagenteissa, sillä laskennalliset vaatimukset ovat yksinkertaisesti liian korkeat (Parkes ja Wellman 2015). Russell 1997 puoltaa tätä näkemystä: Fyysiset mekanismit vievät aikaa tiedon prosessoinnissa ja toimintojen valinnassa, minkä vuoksi todelliset agentit eivät kykene vastaamaan välittömästi ympäristönsä muutoksiin.

Ei-laskennalliset resurssit, kuten agentin työ vuorovaikuttamisessa insinöörien tai loppukäyttäjien kanssa, voivat rajoittaa agentin toimintaa jopa enemmän kuin laskennalliset resurssit (Doyle 1992). Rajoitukset herättävät epäilyksen siitä, onko rationaalisuus hyvin määriteltyä, tai voidaanko sitä ohjelmistoagenteissa ylipäätään saavuttaa. Ideaalisesti rationaalisuus pohtii vain päätösten lopputuloksia, ja saavuttavatko päätökset haluttuja tavoitteita. Resursseiltaan rajallisten agenttien täytyy pohtia myös sitä, tekeekö agentti oikeita valintoja, kun

se pohtii valintojaan tavoitteidensa saavuttamiseksi (Lipman 1991). Tämä voi luonnollisesti johtaa äärettömään pohdintaketjuun, joka estää agentin päätöksenteon kokonaan. Käytännössä agentteja suunniteltaessa tulisi ratkaista tasapaino, kuinka paljon aikaa agentti voi käyttää hallinnollisiin pohdintoihin, jotta se voi samalla saavuttaa mahdollisimman hyviä lopputuloksia.

Toisaalta rajallinen rationaalisuus voidaan nähdä tehokkuutena. Oleellista on erottaa rajalliset rationaalisuudet tavoitteiden perusteella: On eri asia, jos agentin tulee selvittää kaikki mahdolliset tavoitteet ja valita niistä sopivin kuin että agentille annetaan valmiiksi tarkkaan määritetty tavoite, jota se koittaa saavuttaa (Binmore ym. 1998). Ensimmäisessä mallissa agentin täytyy osata pohtia, kuinka se voi parhaalla tavalla allokoida resurssinsa, kun taas jälkimmäisessä agentin täytyy miettiä vain, kuinka se voi parhaalla tavalla saavuttaa tavoitteensa. Tällöin jokaista mahdollista kannattavaa toimintaa ja päämäärää ei tarvitse harkita, vaan valinnat rajoittuvat mahdollisten keinojen ja resurssien tutkimiseen tilanteenmukaisen tavoitteen saavuttamiseksi.

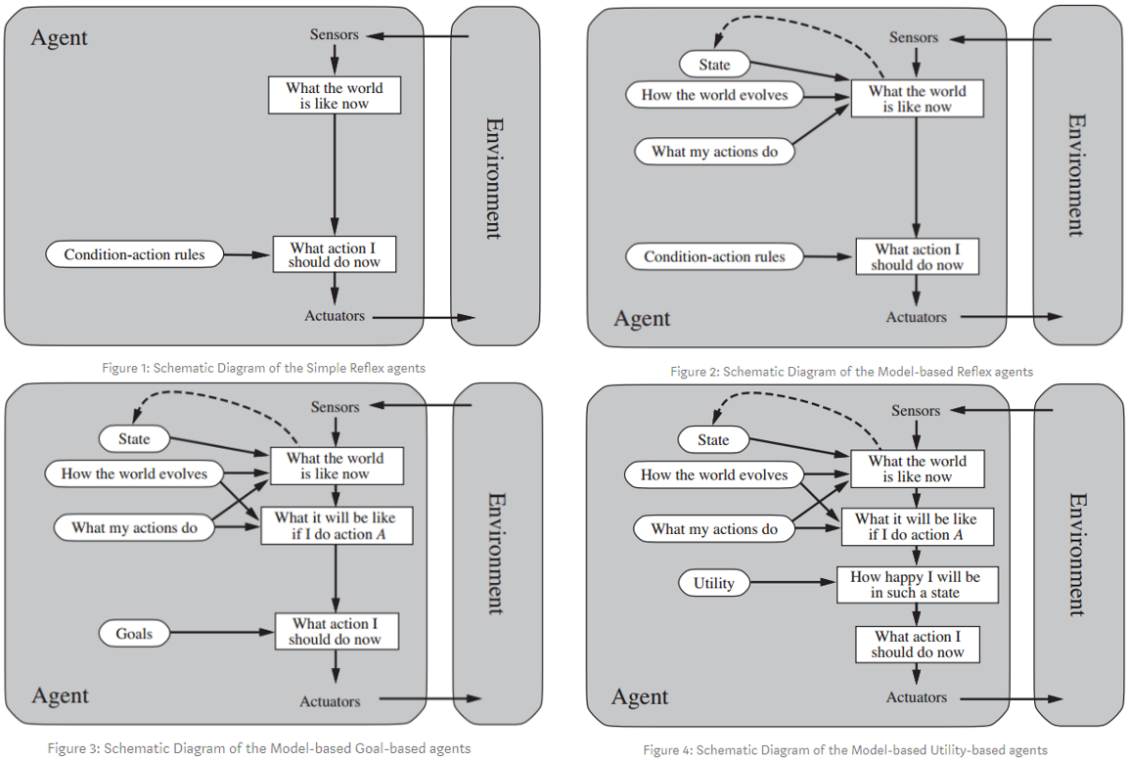
3 Agenttiarkkitehtuurit

Tekoälyn tehtävänä on oleellisesti suunnitella agenttiohjelma, joka implementoi agentin toimintamallin havainnoista tekoihin. Tämä ohjelma toimii jonkinlaisessa laskentalaittessa, jolla on fyysiset sensorit ja toimielimet (eng. *actuators*). Tällaista laitetta kutsutaan agenttiarkkitehtuuriksi (Russell ja Norvig 2016). Arkkitehtuuri voi olla tavallinen pöytätietokone, tai esimerkiksi itseohjautuva auto, jolla on useita tietokoneita, sensoreita ja kameroita. Oleellisesti arkkitehtuuri on jokin laite, joka havaitsee sensoriensa avulla tietoa ohjelmalle, ajaa ohjelman, ja syöttää ohjelman käskyt toimielimille.

3.1 Agenttiarkkitehtuurien typologia

Agenttiarkkitehtuurit voidaan luokitella neljään pääkategoriaan: yksinkertaisiin refleksiagentteihin (eng. *simple reflex agents*), malliperusteisiin agentteihin (*model-based agents*), tavoiteperusteisiin agentteihin (*goal-based agents*) ja hyötyperusteisiin agentteihin (*utility-based agents*) (Hegazy ym. 2003, kts. kuvio 1). Yksinkertaiset refleksiagentit toimivat suoraan sen hetkisen havainnon perusteella ja jättää huomiotta loput havainnointihistoriasta. Malliperusteiset agentit seuraavat maailman tilaa, toisin sanoen ylläpitää jonkinlaista sisäistä tilaa, joka riippuu havainnointihistoriasta. Tavoiteperusteiset agentit tarvitsevat tietoa tavoitteistaan, sillä sen havainnot eivät anna tarpeeksi informaatiota, jolla agentti voisi tehdä päätöksen oikeasta toiminnasta. Joskus tavoitteiden tietäminen ei ole riittävä agentille toimimaan oikein, kun tavoitteet ovat ristiriidassa. Hyötyperusteiset agentit koittavat ratkaista tämän ongelman arvottamalla havainnot numeerisesti, joiden avulla voidaan arvioida kuinka lähelle agentit pääsevät tavoitteitaan tietyillä toiminnoilla.

Jokaisella arkkitehtuurilla on tarkoituksensa, ja niiden tarve riippuu täysin agentille tarkoitetusta tehtävästä. Esimerkiksi yksinkertaiselle itseohjautuvalle pölynimurille riittää yksinkertainen refleksimalli, sillä sen toiminta riippuu täysin siitä, onko sen hetkinen alue likainen. Vaativampien tehtävien näkökulmasta älykkään agentin tulee hyödyntää jokaista neljää arkkitehtuurityyppiä, jotta se voi tehdä sekä nopeita ratkaisuja että tietoperusteista pohdintaa, jolla se voi suunnitella toimiaan etukäteen. Tällaisia arkkitehtuureja kutsutaan hybridiark-



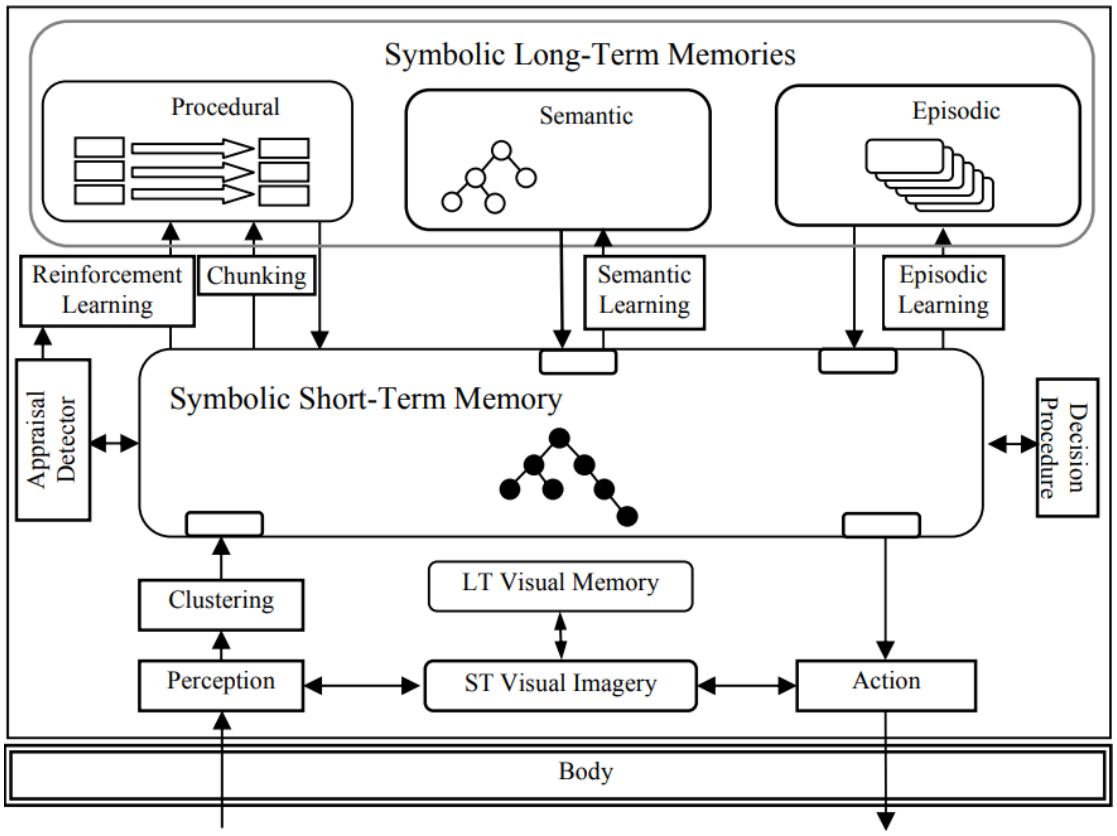
Kuvio 1. Eri arkkitehtuurityypit (Russell ja Norvig 2016, s. 49–54)

kitehtuureiksi (Russell ja Norvig 2016). Hybridiarkkitehtuurien yksi tärkeimmistä ominaisuuksista on se, että agentin päätöksentekokomponenttien rajat eivät ole kiinteät, vaan se kykenee tekemään päätöksiä joustavasti eri tavalla arkkitehtuurin eri komponenttien avulla.

3.2 Ihmisen tietoisuuden ja kognition jäljillä

Arkkitehtuureja, jotka koittavat jäljittää mahdollisimman tarkkaan ihmisen tietoisuutta, tutkitaan jatkuvasti. Yksi tällaisista agenttiarkkitehtuureista on IDA (*Intelligent Distribution Agent*) (Franklin, Kelemen ja McCauley 1998). IDA:n alkuperäinen tehtävä oli neuvotella uusia tehtäviä merimiehille Yhdysvaltojen merivoimissa huomioimalla jokaisen yksilölliset kyvyt ja mieltymykset sekä merivoimien tarpeet. IDA havaitsee dataa merivoimien tietokannoista ja kommunikoi merimiesten kanssa sähköpostin välityksellä käyttäen luonnollista kieltä. IDA on myöhemmin laajennettu LIDA:ksi (*Learning Intelligent Distribution Agent*), johon on lisätty kolme oppimisen muotoa: aistiperäinen-, episodinen- ja proseduraalinen oppiminen (Franklin ja Patterson Jr 2006).

Toinen vastaavanlainen pitkälle kehittynyt hybridiarkkitehtuuri on Soar (Laird 2008). Soar-projektin tarkoituksena on kehittää älykäs agentti, joka kykenee käyttämään kaikentyyppistä tietoa ja jäljittämään kaikkia kognitiivisia kykyjä, joita ihmisillä on, kuten päätöksentekoa, ongelmanratkaisua, suunnittelua ja luonnollisen kielen ymmärtämistä. Soar-arkkitehtuuri koostuu pääsääntöisesti pitkäaikaismuistista ja lyhytaikaisesta työmuistista (kuvio 2). Työmuistissa on agentin sen hetkinen käsitys maailman tilasta, joka on muodostunut agentin havainnoista ja pitkäaikaismuistista. Pitkäaikaismuisti jakautuu proseduraaliseen, semanttiseen ja episodiseen muististiin. Proseduraaliseen muistiin tallentuu tieto toimintasäännöistä, jotta agentti voi tietää ja oppia miten sen tulee toimia kussakin tilanteessa. Semanttiseen muistiin tallentuu deklaratiivisia totuuksia maailmasta, kuten pöydällä on jalat, koirat ovat eläimiä ja niin edelleen. Toisin kuin semanttisessa muistissa, jonka tieto on riippumatonta siitä, milloin ja missä se opittiin, episodiseen muistiin tallentuu tilannekatsauksia menneistä kokemuksista, joita voi hyödyntää vastaavanlaisten kokemusten ilmetessä esimerkiksi ennakoimalla mahdollisten toimintojen vaikutuksia.



Kuvio 2. Laajennettu Soar-arkkitehtuuri (Laird 2008, s. 5)

Soar-arkkitehtuuri implementoi kaikki edellä mainitut arkkitehtuurityypit (kuvio 1) kasaamisen (eng. *chunking*) avulla: Joka kerta kun Soar ratkaisee ongelman syvällisellä pohdinnalla, se tallentaa ratkaisusta yleistetyn version proseduraaliseen muistiin, jotta se voi jatkossa reagoida samankaltaisiin ongelmiin refleksinomaisesti. Näin Soarissa toteutuu sekä yksinkertainen refleksimalli että muut arkkitehtuurityypit.

IDA:n kehittäjät uskovat projektin tuovan syvempää ymmärrystä sekä ihmisen- että keinotekoisesta tietoisuudesta, ja että tietoiset ohjelmat kykenevät joskus suorittamaan toimia, jotka nykyään on varattu vain ihmisille (Franklin, Kelemen ja McCauley 1998). IDA-arkkitehtuurin voidaan odottaa reagoivan uudenlaisiin ja ongelmallisiin tilanteisiin joustavammalla ja ihmisen kaltaisemmalla tavalla, kuin perinteiset agenttiarkkitehtuurit. Soarin kehittäjillä on vastaavanlaisia uskomuksia: Kognitiivisten arkkitehtuurien tutkimus tulee sallimaan yleisten tekoälyjärjestelmien kehityksen, jotka kykenevät toimimaan useissa uudenlaisissa tehtävissä, pitkissä ajanjaksoissa, laajalla empiirisellä oppimisella ja dynaamisissa ja monimutkaisissa ympäristöissä (Laird 2008).

Arkkitehtuurien tutkimuksessa on aukkoja, jotka toisaalta kyseenalaistavat potentiaaliset kyvyt. Soarin komponenttien keskinäinen vuorovaikutus vaatii lisäselvitystä, kuten kuinka tunteet ja visuaaliset mielikuvat tallennetaan episodiseen muistiin, kuinka tieto liikkuu episodisen ja semanttisen muistin välillä, kuinka tunteet vaikuttavat päätöksentekoon ja kuinka tunteiden ja työmuistin aktivointi vaikuttaa tiedon varastointiin ja sen saantiin episodisesta ja semanttisesta muistista (Laird 2008). IDA:n tutkijat ovat puolestaan löytäneet ongelmia agenttien ei-laskennallisista resursseista ja autonomisuuden tasapainon löytämisestä (McCauley ja Franklin 2002). Suuret ja monimutkaiset järjestelmät, kuten Yhdysvaltojen merivoimat, vaatisivat tuhansien ohjelmistoagenttien vuorovaikuttamista keskenään, jonka toteuttamiseen ei ole löydetty tyhjentävää ratkaisua. Autonomisuuden näkökulmasta päätöksentekovaltaa on jätettävä ihmisille, jotta voidaan varmistaa ihmisten tyytyväisyys ennemmin kuin ohjelmistoagentin näkemä optimaalinen tulos. Toisaalta ohjelmistoagenttien tarkoitus jäisi epäselväksi, jos ne eivät kykenisi tekemään itsenäisiä päätöksiä.

4 Mahdollisuudet vaativammissa tehtävissä

Kykenevätkö ohjelmistoagentit vain toimimaan kuin ne olisivat älykkäitä, toisin sanoen simuloimaan ajattelua, vai voivatko ne todella kyetä ajattelemaan? Tekoälytutkimuksen haastavimmat kysymykset tekoälyn tietoisuudesta ja todellisista kyvyistä ovat toistaiseksi yleisesti tunnustettuja mysteerejä, joita on syytä tutkia enemmän (Russell ja Norvig 2016). Jokaisen tekoälykehittäjän pitäisi huolehtia töidensä eettisistä seurauksista.

4.1 Potentiaaliset kyvyt

Rutiininomaisten ja fyysisesti vaativien töiden robotisointi on nykyään arkipäivää, mutta luovat ja strategiset tehtävät ja emergentti päätöksenteko ovat toistaiseksi ihmisten hallussa. Tekoälyteollisuuden tutkijat uskovat vahvasti tekoälyn kykenevän tarkempaan ja nopeampaan päätöksentekoon, ja tämän vuoksi korvaamaan ihmiset kokonaan myös vaativammissa tehtävissä (Terziyan, Gryshko ja Golovianko 2018). Toisaalta tutkijat katsovat myös, että tekoälyn potentiaalisille kyvyille automatisoinnissa halutaan antaa liikaa ja huolimattomasti merkittävyyttä: Tekoälyn laajemmat teknologiset, taloudelliset ja yhteiskunnalliset osallisuudet suhdetoiminnassa vaativat laajempaa kriittistä huomiota (Galloway ja Swiatek 2018).

Parhaat käytänteet ja menestystarinat saattavat tarvita päätöksenteossa ihmisille tyypillisiä ominaisuuksia, kuten intuitiota, tunteita ja irrationaalisuutta. Uusi Pi-Mind -teknologia tarjoaa kompromissin ihmisten ja tekoälyn autonomisuuden välille agenteilla, jotka kloonaavat ihmisen päätöksentekomallin, eivätkä ole täysin itsenäisesti oppivia keinotekoisia päätöksentekijöitä (Terziyan, Gryshko ja Golovianko 2018). Pi-Mind on joukko tekniikoita, malleja ja työkaluja, jonka tarkoitus on laajentaa perinteisiä rationaalisia päätöksentekomalleja teollisessa kontekstissa yhdistämällä kognitiivisia ja luovia näkökulmia.

Yleinen kritiikki tekoälyn toiminnasta on se, ettei tekoäly kykene korvaamaan ihmisten luovuutta. Esimerkiksi Ristic 2017 väittää, ettei tekoäly kykene varmistamaan oikeanlaista äänensävyä puhe- tai kirjoitusviestinnässä, tai suorittamaan luovaa temppua eikä kykene saamaan tunneälykkyyttä. Toisaalta Pi-Mind -teknologia haastaa tätä näkökulmaa: Pi-Mindin hyödyt tulevat ilmi emergenteissä, epävarmoissa ja nopeissa tilanteissa, jotka vaativat no-

peaa päätöksentekoa, joissa ei ole jäykkää universaalia ratkaisua, joihin perinteiset koneoppimismetodit ja ennakoinnit eivät toimi ja jotka ihmiset yleensä ratkovat luovasti (Terziyan, Gryshko ja Golovianko 2018). Pi-Mindin kehittäjät korostavat tekoälyn ja ihmisten yhteistyötä teollisuudessa, ja kutsuvat Pi-Mind -teknologiaa älypalveluksi (*Intelligence-as-a-Service*) ihmisiä varten.

Tekoälykehitys ja erityisesti robotisointi on herättänyt pelkoja paitsi mahdollisista työpaikkojen menetyksistä myös eettisistä kysymyksistä ja siitä, että ihmisiä tulee hallitsemaan heitä älykkäämmät teknologiat. Galloway ja Swiatek 2018 argumentoivat, että tällaiset uhkakuvat eivät toisaalta ole kovin realistisia, sillä todennäköisemmin automatisointi tulee eliminoimaan vain vähäisiä ammatteja kokonaan, ja sen sijaan tulee vaikuttamaan osiin lähes kaikista töistä enemmän tai vähemmän, riippuen työtehtävästä. Suhdetoiminnan harjoittajat epätodennäköisesti korvataan roboteilla tai edes joutuvat vuorovaikuttamaan suoraan robottien kanssa: Valin 2018 ehdottaa raportissaan, että vain vähemmistö suhdetoiminnan tehtävistä ovat taipuvaisia automatisoinnille.

4.2 Etiikka ja riskit

Tekoälykehityksessä tekoälyn potentiaalisten kykyjen lisäksi on otettava huomioon kehityksen mahdolliset eettiset seuraukset. Yampolskiy 2013 argumentoi, että yleispätevän tekoälyn (eng. *Artificial General Intelligence*) kehittäminen on lähtökohtaisesti epäeettistä: Mikäli yleispätevä tekoäly kykenee universaaliin ongelmanratkaisuun ja rekursiiviseen itsensä kehittämiseen, se potentiaalisesti ylittää ihmisten tarpeellisuuden jokaisella alalla, ja näin tekee ihmiskunnasta altistuvan sukupuutolle. Lisäksi todellisella yleispätevällä tekoälyllä on mahdollisesti ihmisen kaltainen tietoisuus, mikä tekisi robottien kärsimyksen mahdolliseksi, ja täten tekoälykokeiluista epäeettisiä. Russell ja Norvig 2016 esittää muiden pienempien seurauksien lisäksi myös uhkakuvan ihmiskunnan lopusta: Tekoälyn oppimismekanismi saattaa aiheuttaa ei-toivottuja seurauksia, kuten niin sanotun älykkyysräjähdyksen. Älykkyysräjähdys, tunnetaan myös nimellä teknologinen singulariteetti, tarkoittaa tekoälykehityksen eksponentiaalista ja ääretöntä kasvua, johtuen superälykkäiden tekoälyagenttien kyvyistä kehittää itseään älykkäämpiä agentteja. Toisaalta jokaisen muun teknologian kehitys on noudattanut S-muotoista käyrää. Toisin sanoen muiden teknologioiden kehityksen eksponentiaalinen

kasvu on ennen pitkää lakannut, joten ei ole täysin perusteltua olettaa tekoälyn olevan tästä poikkeus ja kehittyvän äärettömyyksiin asti (Russell ja Norvig 2016).

Tutkijat ovat kehittäneet erilaisia menetelmiä, joilla pyritään varmistamaan tekoälykehityksen eettisyys. Yampolskiy 2013 ehdottaa tekoälykehityksen katsauslautakuntien perustamista, jotka kieltäisivät kaikki tekoälyprojektit, jotka mahdollisesti johtaisivat yleispätevän tekoälyn kehitykseen. Samankaltaisia lautakuntia hyödynnetään esimerkiksi lääketieteessä uusien lääkkeiden kehityksessä. Lawrence, Palacios-Gonzalez ja Harris 2016 esittävät toisenlaisia ratkaisuja: Älykkäitä tai superälykkäitä tekoälyagentteja kehittäessä täytyy ensiksi selvittää, kuinka agentteja voidaan tehokkaasti hallita ennen niiden vapauttamista maailmalle. Tämän saavuttamiseksi ehdotetaan eri vaihtoehtoja: kykyjenhallintamenetelmät, kannustinmenetelmät, tainnutusmenetelmät (järjestelmän kykyjen tai informaation saannin rajoitus) ja motivaationvalintamenetelmät.

Tekoälyn kyky moraaliselle ymmärtämiselle ja toimijuudelle herättää myös kysymyksiä tekoälykehityksen eettisyydestä. Moraalisella toimijalla tai -agentilla tarkoitetaan agenttia, jonka yhteiskunta katsoo olevan vastuussa teoistaan, jonka hyvinvoinnin varjelemisesta yhteiskunta katsoo olevan vastuussa ja jonka yhteiskunta on näiden lisäksi myös tunnustanut (Gray ja Wegner 2009). Lawrence, Palacios-Gonzalez ja Harris 2016 ehdottavat, että hyvyyden kognitiivinen ymmärtäminen vaatii enemmän kuin vain propositionaalista tai algoritmista tietämystä, toisin sanoen moraalinen tietämys vaatii enemmän kuin vastauksia kysymyksiin *miten* ja *mitä*, sillä siihen liittyy myös kyky käsittää vastauksia kysymyksiin *miksi* ja *millaista se saattaisi olla*. Myös Bryson 2018 argumentoi, että moraalisten ohjelmistoagenttien toteuttaminen ei ole tarpeellista tai edes toivottavaa. Hänen mukaansa ihmisillä on eettinen velvollisuus suunnitella tekoäly, lait ja moraalit niin, että ihmisten valtaote pitää ohjelmistoagentit otteessaan. Ongelmia tulee moraalisen vastuun kantamisessa, sillä kuka olisi vastuussa moraalisen ohjelmistoagentin moraalisista toimista? Esimerkiksi lapset ja lemmikkieläimet eivät ole laillisesti vastuussa teoistaan. Moraalinen vastuu on vain niillä, jotka moraalinen yhteisö on tunnustanut olevan vastuullisessa asemassa (Bryson 2018).

5 Yhteenveto

Tutkielmassa käytiin läpi agentin määritelmiä, toimintaperiaatetta ja rationaalisuutta, agenttiarkkitehtuureja ja pohdintaa ohjelmistoagenttien mahdollisuuksista ja eettisistä ongelmista. Agenttien rationaalisuutta tutkittaessa selvitettiin, että rationaalisen päätöksentekomallin toteuttaminen ohjelmistoagenteissa ei ole yksinkertaista, vaan agentin tulisi pohtia jokaista mahdollista toimintaa ja valita niistä oikea, tai määrittää jokaista mahdollista havainnointiketjua kohtaan toiminta, joka saavuttaisi mahdollisimman hyvän suorituksen. Täydellinen rationaalisuus on toisaalta mahdottomuus resurssien rajallisuuden vuoksi, ja rajoitukset herättävät myös epäilyksiä mahdollisuudesta luoda rationaalista ohjelmistoagenttia. Vaikka ohjelmistoagenttien fokus voidaan asettaa suorittamaan ennalta määritetty tavoite sen sijaan, että agentin täytyisi pohtia kaikkia mahdollisia tavoitteita ja valita niistä sopivin, rationaalisuuden rajallisuus herättää epäilyksiä ohjelmistoagenttien mahdollisuuksista vaativissa tehtävissä, joissa agentin täytyisi nimenomaan kyetä pohtimaan useita tavoitteita.

Agenttiarkkitehtuurien luvussa tarkasteltiin ohjelmistoagenttien rakennetta ja selvitettiin, että useilla erityyppisillä arkkitehtuureilla voidaan saavuttaa kehittyneitä ohjelmistoagentteja, jotka kykenevät sekä reagoimaan nopeisiin tilanteisiin että pohtimaan syvällisemmin tavoitteita ja toimintojen hyötyä. Lisäksi luvussa käsiteltiin kaksi tekoälykehityksen kannalta merkittävää hybridiarkkitehtuuria, IDA ja Soar. Sekä Soarin että IDA:n kehittäjät ovat hyvin optimistisia tekoälyagenttien potentiaalista, siitä, että arkkitehtuurien tutkimus tuo meille syvempää ymmärrystä ihmisen- ja keinotekoisesta tietoisuudesta, ja kyvyistä uudenlaisissa tehtävissä, jotka ovat nykyään varattu vain ihmisille. Molempien arkkitehtuurien kehittäjät ovat toisaalta löytäneet ongelmia niiden kehityksessä, kuten kuinka tunteet vaikuttavat ohjelmistoagentin toimintaan ja kuinka autonomisuus tasapainotetaan ihmisten ja ohjelmistoagenttien välillä. Ongelmista huolimatta tutkijat ovat hyvin optimistisia arkkitehtuurien mahdollisuuksista, joten syytä lisätutkimukselle on olemassa.

Mahdollisuuksia pohdittaessa tarkemmin esiteltiin uusi Pi-Mind teknologia, jonka tarkoitus on kloonata ihmisen päätöksentekomalli, jolloin Pi-Mind teknologian toteuttavat ohjelmistoagentit eivät ole täysin autonomisia päätöksentekijöitä. Pi-Mindin tutkijat ovat myös varmoja tekoälyn potentiaalista ihmisille tarkoitetuissa tehtävissä. Pi-Mind tarjoaa ratkaisun

yleiselle kritiikille tekoälyn luovuuden mahdottomuudesta. Toisaalta todettiin myös, että tekoälyn potentiaalisille kyvyille automatisoinnissa halutaan antaa liikaa valtaa ja huolimattomasti merkittävyyttä, joten tekoälyn valjastaminen vaatii laajempaa kriittistä huomiota. Tutkimusten mukaan muun muassa suhdetoiminnan harjoittajia ei tulla korvaamaan ohjelmistoagenteilla, sillä vain vähemmistö suhdetoiminnan tehtävistä ovat taipuvaisia automatisoinnille. Luvussa todettiin myös, että uhkakuvat työpaikkojen suurilukuisista menetyksistä ja tekoälyn yliotteesta ihmiskuntaa kohtaan eivät ole realistisia, vaan tutkimusten mukaan automatisointi tulee eliminoimaan vain vähäisiä ammatteja kokonaan. Sen sijaan se tulee vaikuttamaan osiin lähes kaikista työtehtävistä jollain tavalla.

Ohjelmistoagenttien potentiaalisten kykyjen lisäksi pohdittiin tekoälykehityksen eettisiä seurauksia ja riskejä. Kirjallisuuden valossa esitettiin älykkyysräjähdyksen käsite ja sen mahdollisuus, mutta todettiin myös, ettei ole täysin perusteltua olettaa sen toteutuminen. Menetelmiä tekoälykehityksen eettisyyden varmistamiseksi on ehdotettu, kuten tekoälykehityksen katsauslautakunnat ja erilaiset menetelmät, joilla pyritään varmistamaan ohjelmistoagenttien hallinta ennen niiden vapauttamista maailmalle. Lopuksi luvussa selvitettiin moraalisen toimijuuden käsite ja esitettyjä argumentteja sille, ettei moraalisten ohjelmistoagenttien toteuttaminen ole eettistä lähtökohtaisesti. Ihmiskunnalla on epäilemättä kyky valmistaa älykkäitä moraalisia toimijoita, mutta niiden valmistaminen on eettisesti hyvin kyseenalaista.

Lähteet

Arsene, Octavian, ja Ioan Dumitrache. 2017. “Mind as multiresolution system based on multiagents architecture”. *Biologically inspired cognitive architectures* 20:31–38.

Binmore, Ken, Cristiano Castelfranchi, James Doran ja Michael Wooldridge. 1998. “Rationality in multi-agent systems”. *The Knowledge Engineering Review* 13 (3): 309–314.

Bryson, Joanna J. 2018. “Patiency is not a virtue: the design of intelligent systems and systems of ethics”. *Ethics and Information Technology* 20 (1): 15–26.

Doyle, Jon. 1992. “Rationality and its roles in reasoning”. *Computational Intelligence* 8 (2): 376–409.

Duro, Richard J, Francisco Bellas ja José A Becerra Permuy. 2014. “Brain-like robotics”. Teoksessa *Springer handbook of bio-/neuroinformatics,* 1019–1056. Springer.

Franklin, Stan, ja Art Graesser. 1996. “Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents”. Teoksessa *International Workshop on Agent Theories, Architectures, and Languages,* 21–35. Springer.

Franklin, Stan, Arpad Kelemen ja Lee McCauley. 1998. “IDA: A cognitive agent architecture”. Teoksessa *SMC’98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218),* 3:2646–2651. IEEE.

Franklin, Stan, ja FG Patterson Jr. 2006. “The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent”. *pat* 703:764–1004.

Galloway, Chris, ja Lukasz Swiatek. 2018. “Public relations and artificial intelligence: It’s not (just) about robots”. *Public relations review* 44 (5): 734–740.

Gray, Kurt, ja Daniel M Wegner. 2009. “Moral typecasting: divergent perceptions of moral agents and moral patients.” *Journal of personality and social psychology* 96 (3): 505.

Hegazy, Islam M, Taha Al-Arif, Zaki T Fayed ja Hossam M Faheem. 2003. “A multi-agent based system for intrusion detection”. *IEEE Potentials* 22 (4): 28–31.

Laird, John E. 2008. “Extending the Soar cognitive architecture”. *Frontiers in Artificial Intelligence and Applications* 171:224.

Lawrence, David R, Cesar Palacios-Gonzalez ja John Harris. 2016. “Artificial intelligence: the shylock syndrome”. *Cambridge Quarterly of Healthcare Ethics* 25 (2): 250–261.

Lipman, Barton L. 1991. “How to decide how to decide how to...: Modeling limited rationality”. *Econometrica: Journal of the Econometric Society:* 1105–1125.

McCauley, Lee, ja Stan Franklin. 2002. “A large-scale multi-agent system for navy personnel distribution”. *Connection Science* 14 (4): 371–385.

Newell, Allen, ym. 1982. “The knowledge level”. *Artificial intelligence* 18 (1): 87–127.

Padgham, Lin, ja Michael Winikoff. 2005. *Developing intelligent agent systems: A practical guide.* Nide 13. John Wiley & Sons.

Parkes, David C, ja Michael P Wellman. 2015. “Economic reasoning and artificial intelligence”. *Science* 349 (6245): 267–272.

Ristic, D. 2017. *PR in 2018: Dominated by technology, mired by inauthenticity. PR week.*

Russell, Stuart J. 1997. “Rationality and intelligence”. *Artificial intelligence* 94 (1-2): 57–77.

Russell, Stuart J, ja Peter Norvig. 2016. *Artificial intelligence: a modern approach.* 46–58. Malaysia; Pearson Education Limited,

Sloman, Aaron. 1971. “Interactions between philosophy and artificial intelligence: The role of intuition and non-logical reasoning in intelligence”. *Artificial intelligence* 2 (3-4): 209–225.

Terziyan, Vagan, Svitlana Gryshko ja Mariia Golovianko. 2018. “Patented intelligence: Cloning human decision models for Industry 4.0”. *Journal of manufacturing systems* 48:204–217.

Valin, J. 2018. *Humans still needed: An analysis of skills and tools in public relations.* Tekninen raportti. Discussion paper. Retrieved fromLondon: Chartered Institute of Public Relations.

Yampolskiy, Roman V. 2013. “Artificial intelligence safety engineering: Why machine ethics is a wrong approach”. Teoksessa *Philosophy and theory of artificial intelligence,* 389–396. Springer.