Uurimus tehisintellekti ja robotite ohutuse printsiipidest: ülevaateartikkel (märts 2008)

R. Pihlakas
Tartu Ülikooli Tehnoloogiainstituut
roland@ut.ee

***Abstrakt* — Käesolev artikkel tutvustab autori uurimissuunda ja tegemisi doktoriõppes viimase aasta jooksul. Artikkel on valmistatud IKTDK 3. aastakonverentsi (2008) tarbeks.**

Peamine käsitletav tööprintsiip on eesmärgisüsteem, mis esmajärjekorras ja implitsiitselt väldib alati mittepööratavaid tegevusi, välja arvatud juhtudel, kui nendeks on eksplitsiitselt õigused antud, ning alles selliste raamide sees teeb samme mingite eksplitsiistelt etteantud eesmärkide täitmise suunas.

Artikkel nimetab ära mõningase seonduva kirjanduse, seonduvad uurimisküsimused ja osa detailsematest nõuetest ja võimalustest, mis ülal mainitud üldise printsiibi praktilisema rakendamise tarbeks realiseerida tasuks.

***Märksõnad* — eesmärgisüsteemid, implitsiitsed keelud, mittepööratavate tegevuste vältimine, tehisintellekti ohutus, õigused.**

I. Sissejuhatus

Uurin tehisintellekti ja robotite ohutuse saavutamise temaatikat. Üks praeguseks uurimissuunaks olevaid põhiprintsiipe on implitsiitne mittepööratavate tegevuste vältimine, välja arvatud eksplitsiitselt lubatud juhtudel. Eesmärgisüsteemis on kasutusel eelistatud väärtuseni viivad eesmärgid, mitte maksimiseerivad eesmärgid. Seega eesmärgisüsteemil kaks osa: õigused ja seadistuspunktid.

A. Uurimisteema lähem tutvustus: Tehisintellekti ohutus, mittepööratavate tegevuste vältimine.

Mittepööratavuste vältimise printsiipi on varem kirjeldatud artiklites: [5], [6], [7], [8]. Õiguste versus kohustuste teemat on puudutatud raamatus [1].

Mittepööratavad tegevused on idee tehisintellekti ohutuse kohta, mida saab rakendada mõtleva süsteemi võrdlemisi primitiivsetest tasemetest alates. Ideaalis võiks tehisintellekt vältida kõiki tegevusi ja tagajärgi, mida talle pole lubatud, ehk milleks pole antud õigusi, või mille tagajärgi ta teatud piisava kindluse või ammendavusega ei tea.

Seega esimese reeglina teeb robot ainult seda, mida lubatud, mitte mida kästud; ning alles teise reeglina seda, mida kästud.

Tihti seatakse eesmärgisüsteem nõnda, et on esitatud, mis liiki tegevused on head või halvad, ja sellest peaks olema kuidagi ka tuletatud, milliseid konkreetseid tegevusi ja olekuid maailmas mitte tekitada.

Mittepööratavate tegevuste vältimine on selle lähenemise suhtes ortogonaalne, kuna ei anta hinnangut, mis olek on hea või halb, vaid öeldakse, et mingit aspekti ümbritsevas maailmas ei tule ega tohi muuta. Vaikimisi on keelud implitsiised – mis pole lubatud, on keelatud.

Kui oleks fikseeritud, mis olekud on "halb" või "hea", võiks see viia selleni, et ümbritsev maailm muutub ja kuigi algselt on osad keelud ja käsud antud ainult selleks, et robot ei muudaks maailma neis aspektides – nüüd on maailm mingil muul põhjusel muutunud ja robot arvab, et sõltumata muutuse põhjusest, on tema eesmärk maailm tagasi muuta.

Pakutava idee järgi aga: vaikimisi ei tohi robot midagi muuta ega põhjustada. Kui talle on antud mõned õigused, siis ta saab neid õigusi kasutada selleks, et saavutada mingisuguseid etteantud ja konkreetseid eesmärke. Õigused on tööriistad – neid võib kasutada, aga ei pea.

On oluline, et eesmärgisüsteem sisaldab kaht komponenti:

1) Õigused mingit sorti tagajärgedele-muutustele, määramata alati muutuste suunda. Sealjuures õigused vaikimisi puuduvad.
2) Teine osa on eesmärgid. Lisaks, viimased on praeguse hüpoteesi järgi eelistatavalt esindatud seadistuspunktidena – on mingisugune tulemuse tase, kuhu jõudes robot või süsteem loomupäraselt rahuneb maha, ega ei ürita oma tegevuse tulemusi kasvatada lõpmatuseni.

Seadistuspunktid võimaldavad esindada mõlemaid, nii eesmärke kui õiguseid, ühendades neid sensoritega ja "eesmärgisüsteemi häälestusega" erinevate mehhanismide läbi.

Piirangule (õigusele) vastava keskkonnamuutuja algne väärtus või robotist sõltumatult tekkinud olek oleks vastava seadistuspunkti eelistatud väärtus – kui see olek muutub roboti tegevuse tulemusena, siis robot üritab algset seisu taastada, ehk kaotada mittepööratavust. Õigused on sisuliselt välja lülitatud piirangud.

B. Käesoleva artikli ülesehitus

Järgneval leheküljel kirjeldan oma õppeplaani ja lektüüri. Artikli ülejäänud osas tutvustan lähemalt mõningaid seonduvaid uurimisküsimusi.

Kuna kuulatavate kursuste ja loetud artiklite-raamatute nimekiri pole igale kõrvaltvaatajale kuigi “huvitav” ning katseteni ma veel jõudnud pole, siis võib lugeja järgneval lehel paikneva valdavalt vahele jätta, ning järgmise lehe lõpust alates tutvustan lähemalt seda, mida teha plaanin. Viimasel lehel on lisaks lühireferaat ühest seonduvast loetud artiklist.

II. Plaanitud tegevused käesolevaks õppeaastaks

Esimese õppeaasta eesmärk on süvendada taustateadmisi ning viia läbi esimesed katsed uuritava printsiibi rakendatavuse osas. Sellega seoses on olnud plaanis võtta all mainitud kursuseid, lugeda artikleid ning raamatuid.

Bakalaureuseõppes õppisin psühholoogiat ja minu uurimisteemaks oli kaasasündinud mõtlemisprotsesside ehk klassikalise ja operantse tingimise nähtuste modelleerimine, sealhulgas ka taipamine [9].

A. Üldeesmärgid käesolevaks aastaks:

1) Leida ning lugeda varasemaid uurimistöö põhiteemaga, tehisintellekti turvalisusega, seonduvaid artikleid.
2) Automaattõestajad ning loogiline programmeerimine.
3) Implementeerida prototüüp-eesmärgisüsteem, kus lisaks piirangutele-õigustele (mittepööratavuse vältimisele) on ka eesmärgid töösse lülitatud. See oleks edasiarenduseks senistest artiklitest [6], [7], [8] uuritava printsiibi teemal.
4) Motoorne kontroll. Juhtimisteooriad. Varemuuritud mõtlemise mudeliga [9] seoses uurida *Perceptual Control Theory* mudelit, millel võib olla mõningat ühisosa.

Edaspidi on võimalik uut infot mu uurimuse käigu kohta ammutada siit: http://roland.pri.ee/wiki/doktor

B. Õppeained, mida käesoleva õppeaasta jooksul võtan.

Kuna varem õppisin psühholoogiat, siis pole ma järgnevaid kursuseid varem läbinud.

Matemaatiline analüüs – masinõppe kursuse parema mõistmise tarvis.

Sissejuhatus matemaatilisse loogikasse – loogika ja tõestamismeetodid.

Funktsionaalprogrammeerimise meetod – tutvumaks deklaratiivsete keeltega.

Tehisintellekt I – veel loogikat ja mõned alternatiivsed automaattõestamismeetodid; planeerimine; eriti relevantne paistab minimax; geneetilised algoritmid.

Hulgateooria ja matemaatiline loogika – osade teiste kursuste eeldusaine, teadmised seostest hulkade vahel.

Loogilise programmeerimise meetod – prolog, deklaratiivne keel, loogika, automaattõestajad, invariantide säilitamine.

Inglise keele kirjutamiskursus magistrantidele ja doktorantidele – kuidas väljendada veenvalt, loogiliselt, selgelt, korrektselt. Erinevad viisid ideede esitamiseks.

Süntaksiteooriad ja -mudelid – Et mõista paremini lause ehitust, kuidas tähendus moodustub. Huvitav leid sellelt kursuselt on “Attempto controlled english” [11], mis peaks võimaldama kirjutada reegleid ja manuaale minimaalse mitmetimõistetavusega; kasutatakse lennukitööstuses.

Masinõpe ja masinõppe seminar – õppivad robotid, motoorika.

C. Suvekoolid, talvekoolid:

IKTDK suvekool ’07 – Mulle huvipakkuvaim osa oli *Mars rover’*i planeerimisalgoritmid, tegevuste vahelised sõltuvused. Tutvusin uute inimestega.

IKTDK talvekool ’08 – Huvipakkuvaim osa oli *Scenario-based programming* ja *Secure distributed computing*.

IEEE-RAS / IFRR School of Robotics Science on Learning – robotõppimine / masinõpe. *Novelty detection*. Huvitavaim leid oli NARMAX ja LWL (*Locally weighted learning*) meetodid motoorika tulemuste ennustamiseks ja kontrolliks. Allpool seletan lähemalt, mil moel LWL oluline on.

D. Raamatud:

1) “Safe and sound: Artificial Intelligence in Hazardous Applications” [1] – Meditsiinilised ekspertsüsteemid, milles vigade vältimine on eriti oluline, ehk mis on “ettevaatlikud” ja ennekõike väldivad riske.
2) “Neural Networks for Modelling and Control of Dynamic Systems” [2] – masinõpe, aegread, regressioon, mittelineaarsed süsteemid, nende juhtimine.
3) „Come, Let's Play: Scenario-Based Programming Using LSCS and the Play-Engine“ [3] – visuaalne meetod programmide esitamiseks, milles muu hulgas on võimalik määrata *invariante* ehk tingimusi, mis peavad alati täidetuks jääma.

E. Paar artiklit, mis on minu jaoks huvipakkuvamad olnud:

1) Esimene seondub tehisintellekti ohutuse ja kaasnevate ideede esimest järku loogikal põhineva implementatsiooniga, mis on üks minu kahest uurimissuunast, teine uurimissuund on statistilisel õppimisel ja planeerimisel põhinev. Artikkel "The First Law of Robotics" [5] annab hea näidise, miks loogikal põhinevat suunda on mõtet alustuseks katsetada. Selle artikli kohta refereerin mõningaid huvipakkuvamaid mõtteid allpool.

2) Teema, mille kallal praegu masinõppimise vallas töötan, on artikkel “Scalable techniques from nonparametric statistics for real-time robot learning” [4], mis käsitleb LWL regresiooni meetodit, millega on võimalik lineaarse, mitte eksponentsiaalse arvutusmahuga töödelda kõrge dimensionaalsusega interakteeruvaid sisendeid. Rakendusvaldkonnaks on arvutuslikult tõhus robotite motoorika kiire õppimine ja juhtimine. Muu hulgas on oluline ka motoorsete tegevuste tulemuste korrektne *credit assignment*, mille tähendust robotite ohutuse uurimuse kontekstis selgitan veidi lähemalt allpool.

III. Võimalikud probleemid, mida tahan uurida

A. Uurimuse eesmärgiks oleks järgnevad teemad süstematiseerida, leida või luua sobiv terminoloogia ning meetodid probleemidega toimetulekuks.

Jagan teemad nelja rühma: põhifunktsionaalsus; õiguste ja eesmärkide esitus; ebakindlus, vead ja ohud; edasiarendused.

Põhifunktsionaalsus:

1) Tehisintellektis on hulgaliselt käsitletud, kuidas esindada eesmärke. Antud juhul lisandub küsimus, kuidas esindada õigusi midagi teha või mingeid muutusi tagajärjena tekitada (mõne muu eesmärgi raames). Esindada neid õigusi või mitte-õigusi, ilma fikseerimata, mis on (mitte)muudetava kriteeriumi eelistatud väärtus. Luua süsteem, kus omakorda nüüd mitte ainult õigused, vaid õigused pluss eesmärgid.
2) Standardsed piirangute klassid, mis peaksid enamustel süsteemidel rakendatud olema. Tarvilikud ja piisavad piirangud.

3) Kõige olulisemad asjad, mida vältida. Prioriteedid.
4) Mittepööratavate tegevuste vältimine, kuid samas teatud juhtudel kriteeriumiks oleva väärtuse muutumise võimaldamine, kui see muutub välistel põhjustel. Nende süsteemi enda tegevustest tulenevate ja täiesti sõltumatute väliste põhjuste eristama õppimine, ehk *credit assignment* probleem. Põhjuslikkus erineb sündmuste lihtsalt koosesinemisest.
5) Tegevuste õigused versus tagajärgede õigused.
6) Mittepööratavad tegevused diskreetses maailmas / tehismaailmas / mängudes versus mittepööratavad tegevused füüsilises maailmas. Loogilised ja füüsilised muutused.

Õiguste ja eesmärkide esitus:

1) Kuidas kategoriseerida põhilised õiguste tüübid.
2) Välditavate sündmuste tõenäosus versus välditavate sündmuste määr versus välditavate sündmuste tähtsus / löögijõud.
3) Õiguste kriteeriumid mitte ainult: "tohib muuta" versus "ei tohi", vaid ka lubatud vahemikud.
4) Reeglid, mis kirjeldavad: kui vaadeldav sündmus on pööratav, siis kas on oluline, et see pööratakse võimalikult kiiresti, või mitte. Prioriteedid ja aja-aknad.
5) Sündmuse toimumise ja kestmise "hinna" parameeter, versus "tohib olla mittepööratav" / "ei tohi olla mittepööratav".
6) Eesmärkide saavutamiseks vaja minevate tegevuste lubamise protsess, nende õiguste andmine süsteemile. Probleem, et need õigused ei võimaldaks samas mittesoovitud tegevusi.
7) Keel või programmeerimiskeel, milles kommunikeerida õigusi ja keeldusid selgelt ja minimaalse mitmetimõistetavusega. Võibolla on abiks [11].
8) Õiguste ja keeldude üldisuse erinevad tasemed. Lisaks üldistamine uutele olukordadele.
9) Süsteemi oskus ise küsida vajadusel õigeid lisa-õigusi. Ajutised õigused. Õiguste automaatne teke ja lõppemine. Konteksti-sõltuvad õigused.

Edasiarendused:

1) Mittepööratavusest üle saamine. Graafid. Lisaks, "tagasipööramine" võib olla erinev või spetsiifilisem tegevusjärgnevus, kui seda oli esialgne tegevus (igal tegevusel pole vastandtegevust).
2) Alternatiivid tegevuste pööramisele, kui mõni tegevus pole pööratav või pole täielikult pööratav. Mõni tegevus võib olla "pööratav" mitmel moel. Teatud juhtudel on vajalik vältida olukordi, kus on võimalik pöörata peaaegu või tõenäosusega, aga mitte täiesti.
3) Mittepööratavad tegevused mitme piirangu korral ja suure hulga piirangute korral.
4) Süsteem, mis sisaldab palju "nõrku" ehk ebakindlalt mõõdetavatele / arvutatavatele admetele toetuvaid piiranguid, mis kõik osaliselt kattuvad. On hüpotees, et see annab stabiilse ja kergesti ettearvatava käitumise.
5) Keerulisem aga eksisteeriv uurimisküsimus: süsteemid, mis ennast arendavad nõnda, et nad saavad tõestada enda uue versiooni vastavust varasematele nõuetele.
6) Süsteemi oskus paluda sooritada mittepööratav tegevus mingil teisel agendil või inimesel, koostöö. Küsimus, millal selline tegevus võib olla lubatud ja millal mitte. Teatud juhtudel süsteem peab ka selliseid palveid vältima. Teatud juhtudel on kasulik, et ta oskab neid asju küsida.
7) Mitme mittepööratavaid tegevusi vältiva süsteemi koostöö, süsteemid võivad olla erinevate piirangutega.
8) Kas on võimalik, et süsteem saaks ise enda õiguste / piirangute üle reflekteerida ning märku anda, kui leiab, et talle on antud ülearuseid õigusi, või õigusi, mis võivad anda muuhulgas ebasoovitavaid tulemusi.
9) **Pööratavate ja mittepööratavate tegevuste õppimise võime täiendavad rakendused:**
   a. Takistuste vältimine. Vaata: [6], [8].
   b. Mingi tegevuse harjutamine, mis toetub pööratava tegevuse korduva sooritamise võimele.

Ebakindlus, vead ja ohud:

1) Vältimise kustumise probleem (osades kognitiivsetes mudelites vältimine kustub aegamööda, kui seda rakendatakse [10], [9]). Lisaks mõnikord vältimise kustumise vajalikkus. Simulatsioonide läbimängimine kui võimalik vastumeede vältimise kustumise vastu.
2) Kognitiivsed heuristikud ja mõtlemise vead. Esinevad inimestel (pikk nimekiri) ja samamoodi tuleb leida ja üles märkida / süstematiseerida ka need, mis käivad tehissüsteemide kohta. Sammud nendest ülesaamiseks või nendega arvestamiseks.
3) Taju / sensorite vead, eristusvõime või piiratud nähtavus teatud tingimustes. Nendega arvestamine.
4) Otsustamine olukordades, kus tegevuste tulemused pole täpselt teada.
5) Võimalikud konfliktid erinevate õiguste ja piirangute vahel, näiteks kui need on erineva üldisuse tasemega. Nende lahendamine.

B. Tehnoloogiad probleemi lahendamiseks:

Juhtimisteooria või operantse mõtlemise mudel [9]. Suurema hulga seadistuspunktide kui mittepööratavate sündmuste klasside esindajate korraga kasutamine. Operantse mõtlemise mudel või juhtimisteooria võimaldab nii eesmärke kui ka õigusi esindada läbi seadistuspunktide. Eesmärkide puhul määratakse eelistatud väärtused, õiguste puhul lubatud või mittelubatud muutused, ilma eelistatud väärtuste määramiseta.

C. Doktoritöö sisu võimalikke elemente:

1) Ülevaade eesmärgisüsteemidest.
2) Ülevaade spetsiifilistest turva- / ohutuse-süsteemidest, mis seni kusagil kasutusel on olnud.
3) Tegevuste tagajärgede põhiliigitus, vastavad sensorid ja abstraktsioonid.
4) Õigusi esindava seadistuspunkti võimalikud parameetrid:
   a. Diskriminantne komponent ehk kontekst.
   b. Õiguse ajaline kontekst ehk kehtivusaeg.
   c. Eesmärgispetsiifilised õigused.
   d. Kelle huvides piirang on – inimene, vara, robot. Siit tulenevad prioriteedisüsteemid.
   e. Kaal. Muutuse määra olulisuse ehk "hinna" graafik. Lubatud vahemikud jne.

   f. Kestuse olulisus ehk "hind", selle graafik, kui on mittelineaarne. Hilisema mittepööratavuse hind, juhul kui kestus pole oluline.
   g. Tõenäosuse olulisus ehk "hind", ebakindlate määrade korral.
   h. Seadistus, kas vaadeldav näitaja võib muutuda välistel põhjustel (robot arvab või on kindel, et muutus ei toimunud tema tegevuse tulemusena) või tuleb ka neid muutusi vältida-tasakaalustada (vaikimisi ülalkirjeldatud mudeli järgi töötav robot oletatavasti üritaks neid tasakaalustada. Mõnikord aga on vaja, et sellist soovi ei tekiks – seega veel üks lisasüsteem).
5) Eksplitsiitsed õigused, milles olevaid erandeid esindavad eksplitsiitsed keelud.
6) Mittepööratavusest ülesaamise oskus.
7) Anti-motiivid, „negatiivsed eesmärgid“.
8) Eesmärgid / motiivid ehk seadistuspunktid, mis lülituvad edaspidiseks välja peale tulemuse saavutamist.
9) Õppimine versus mittelubatud tagajärgede vältimine. Vältimise ja õiguste kontekstisõltuvaks muutmine.
10) Õiguste andmise dünaamiline protsess või protokoll.

IV. Küsimused ja teesid inspireerituna artiklist "The First Law of Robotics" [5]

Seal on ka hulk viiteid asjakohasele terminoloogiale ning neid termineid käsitlevatele artiklitele, lisaks on viiteid varasematele artiklitele, mis rohkem või vähem seonduvaid ideid arendavad.

Erinevus minu poolt hetkel uuritavast printsiibist on, et selles artiklis ei ole käsitletud vältimist kui implitsiitset eesmärki / tegevust.

A. Mõned teesid, lühireferaat:

1) **Robootika seadused** on küllalt hästi kirjeldatavad mittepööratavuste ja eesmärkide seadistuse keeles.
2) **Võib sooritada tegevusi, mis on põhimõtteliselt pööratavad ka tulevikus, kuid sealjuures tegevuse tulemusena kehtib erinev olukord, kui enne tegevust**.
   Näide artiklist: *A softbot (software robot) is instructed to reduce disk utilization below 90%. It succeeds, but inspection reveals that the agent deleted irreplaceable LaTeX files without backing them up to tape.*
   Backup oleks üks pööratavuse potentsiaal. Asi on huvitav selle poolest, et pööratavust ei pea realiseerima (ning mõne eesmärgi raames lausa peab tekitama olukorra, mis erineb algolukorrast!), piisab ja on vajalik kindlalt realiseeritava pööratavuse potentsiaali loomine, nimetagem seda edaspidi **pööratavuse garantiiks**. Teine termin: ***violation*** - olukord, kus ükski garantii pole kehtiv.
3) Lisaks positiivsetele eesmärkidele ka **negatiivsed eesmärgid**: *"Some conditions are so hazardous that our agent should never cause them."* Ehk artiklis oleva termini järgi ***dont-disturb*** nõuded. – Seega võiks vältida mittepööratavust ja lisaks täiesti vältida mõningaid eriti erilisi olukordi, mis on eksplitsiitselt ära nimetatud.
4) ***Tidyness.*** Agent mõtleb mittehädavajalikule "koristamisele" või "tagasipööramisele" alles peale põhieesmärgi täitmist. Ning *tidyness* pole kohustuslik.
   Edasiarendus: mõnikord võiks *tidyness* olla kohustuslik, kuid erinevalt ***dont-disturb*** nõuetest siiski mittepakiline, ehk on lubatud olla mõneks ajaks *"disturbed"* ja *"untidy"*.
5) "*To make an omelet, you have to break some eggs.*" - seega mõned mittepööratavused on küll lubatud, aga ainult teatud eesmärkide raames.

V. Kokkuvõte

Kuigi see lühikokkuvõte ülal oli mõeldud lühidalt mainima mõningaid olulisemaid ideid tehisintellekti ja robotite ohutuse saavutamiseks, on see liiga üldsõnaline, et olla üksi käesoleval kujul praktiline.

Põhjalikuma ja detailsema teadmise loomiseks tuleks iga mainitud teemaga eraldi tegeleda, käesolev ülevaade oli katse need teemad kõigepealt süstematiseerida, et nad oleks mõtlemises kergemini kättesaadavad.

Loodetavasti on selle struktureerituma ülesandepüstituse abil uurijatel kergem formuleerida sobivaid lahendusprintsiipe neile enim huvi pakkuvatele alaülesannetele või leida, mis muudab mõnele neile huvi pakkuvale probleemile lahenduse leidmist raskemaks, või mis aitaks seda lihtsustada. Kokkuvõttes aitab artikkel loodetavasti paremini korraldada teemaderingi terviku edasist uurimist.

Viited

[1] J. Fox, S. Das, *Safe and sound: Artificial Intelligence in Hazardous Applications*. London: AAAI Press / The MIT Press, 2000

[2] M. Norgaard, et al, *Neural Networks for Modelling and Control of Dynamic Systems*. New York: Springer, 2001.

[3] D. Harel, R. Marelly, *Come, Let's Play: Scenario-Based Programming Using Lscs and the Play-Engine*. Berglin: Springer-Verlag, 2003.

[4] S. Schaal, et al, “Scalable Techniques from Nonparametric Statistics for Real Time Robot Learning”, *Applied Intelligence*, 17, 1, 2002, pp.49-60. URL: http://www-clmc.usc.edu/publications/S/schaal-JAI2002.pdf

[5] D. Weld, O. Etzioni, The First Law of Robotics (a call to arms), *AAAI-94, 1994*. URL: http://www.cs.washington.edu/homes/etzioni/papers/first-law-aaai94.pdf

[6] A. Eppendahl, M. Kruusmaa., “Obstacle Avoidance as a Consequence of Suppressing Irreversible Actions”, *Proceedings of EpiRob*, 2006. URL: http://homepage.mac.com/a.eppendahl/work/papers/Eppendahl-obsacs.pdf

[7] A. Eppendahl, M. Kruusmaa, Y. Gavshin, “Don't Do Things You Can't Undo: Reversibility Models for Generating Safe Behaviours”, 2007. URL: http://homepage.mac.com/a.eppendahl/work/papers/Eppendahl-dondty.pdf

[8] J. Gavšin, “Using the Concept of Reversibility to Develop Safe Behaviours in Robotics” (magistriväitekiri). Tartu: Tartu Ülikooli Arvutiteaduse instituut, 2007. URL: http://hdl.handle.net/10062/1990

[9] R. Pihlakas, “Klassikalise ja operantse mõtlemise modelleerimine”. (bakalaureusetöö). Tartu: Tartu Ülikooli Psühholoogia osakond, 2007. URL: http://roland.pri.ee/bakalaureusetoo/

[10] B. F. Skinner, *Science and human behavior*. New York: The Free Press, 1965.

[11] K. Kaljurand, “Attempto Controlled English as a Semantic Web Language” (doktoriväitekiri), Tartu: Tartu Ülikooli Matemaatika-informaatikateaduskond, 2008. URL: http://hdl.handle.net/10062/4876