СОФИЙСКИ УНИВЕРСИТЕТ “СВ. КЛИМЕНТ ОХРИДСКИ”
ФИЛОСОФСКИ ФАКУЛТЕТ
КАТЕДРА “ФИЛОСОФИЯ”

АВТОРЕФЕРАТ

на

ДИСЕРТАЦИЯ

за присъждане на образователна и научна степен “доктор”
Научна специалност:
2.3. Философия (Философия с преподаване на английски език)

ЕТИЧЕСКИ И СОЦИАЛНИ ПРОБЛЕМИ НА ИЗКУСТВЕНИЯ ИНТЕЛЕКТ

Докторант: Невена Руменова Георгиева

Научен ръководител: проф. д.ф.н. Анета Карагеоргиева

Дисертационният труд е обсъден и одобрен за защита на заседание на катедра "Философия" към Философски Факултет на СУ „Св. Климент Охридски", проведено на 25.04.2016 г.

Защитата на дисертационния труд ще се състои през 2015 г. в СУ „Св. Климент Охридски" пред жури в състав:

доц. д-р Димитър Иванов, СУ „Св. Климент Охридски", председател

проф. дфн Анета Карагеоргиева, СУ „Св. Климент Охридски"

проф. дфн Димитър Димитров, ВТУ „Св. Св. Кирил и Методий"

доц. д-р Асен Димитров, ИИОЗ при БАН

доц. д-р Росен Люцканов, ИИОЗ при БАН

Дисертационният труд е в обем 219 страници и се състои от увод, две части, заключение и библиография. Библиографията съдържа 77 заглавия (10 на български, 66 на английски и 1 на руски).

СЪДЪРЖАНИЕ НА АВТОРЕФЕРАТА

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

УВОД

Проблемите на изкуствения интелект са актуална тема на дебата в съвременната западна философска мисъл. Развитието му като научна разработка отдавна е прекрачило прага на лабораториите и инженерните работилници. Научната област, занимаваща се с изкуствения интелект, има различни аспекти – 1. трансхуманизма, който се занимава с увеличаването и подобряването на естествените физически и интелектуални способности на човека, 2. прехвърляне на човешкото съзнание в изкуствена среда с оглед постигане на безсмъртие и неограничен достъп до всяка точка на виртуалния свят и 3. разработване на изкуствен интелект с всички възможности, способности и характеристики на човешкия интелект. Последният тип изследвания и разработки са неразривно свързани с инженерните търсения и структурирането на хуманоидни роботи като носители на интелект. Един от основните фокуси на философския дебат е не толкова дали е възможно постигането на изкуствения интелект, колкото върху възможните отношения между хората и изкуствения интелект, когато последният постигне ниво на близко или равно на човешкия. Част от философите и останалия научен свят смятат, че изкуствен интелект, равен на човешкия, е потенциална опасност за развитието на апокалиптични сценарии за бъдещето на човечеството. Има и по-умерени гледни точки, като например предположението, че когато изкуственият интелект достигне точката на сингуларност, той/тя ще си построи кораб, с който ще напусне планетата. С други думи, сценариите се разделят на три основни типа: 1. апокалиптични (сценарии, според които изкуственият интелект ще заеме доминираща позиция спрямо хората); 2. умерени (сценарии, според които хората и изкуственият интелект ще имат отношения на индиферентност) и 3. позитивни (сценарии, според които хората и изкуственият интелект ще съжителстват в сътрудничество).

Понастоящем изкуственият интелект (като изкуствен тесен интелект – artificial narrow intelligence) съществува и е в масово производство, което го прави достъпен за всички. Разбира се, умните машини не са умни, колкото нас, но фактът, че могат да изчистят дома ни сами, вече поставя на дневен ред проблемите, които ще възникнат от все по-нарастващото им присъствие в

ежедневието ни в социален и етически аспект. В този смисъл обектът на настоящото изследване е изкуственият интелект, а предметът са етическите и социалните проблеми, пред които е изправено нашето общество. Целта на настоящата дисертация е да изследва социалните проблеми на изкуствения интелект, настоящия статус на разработките и стратегическата рамка за развитие в областта на науката, инженерството и информационните технологии (по-специално програмните разработки), за да се използват за база, върху която да се потърсят възможни етически доктрини, които биха послужили за постигането на оптимално позитивни отношения между хората и машините, носители на изкуствен интелект и основаното на сътрудничество взаимодействие помежду им.

За постигането на целта са изпълнени три задачи. Изследвани са различни философски направления, които биха послужили за основа и обосновка на изкуствения интелект с оглед да се прецени дали изкуствен интелект, подобен на човешкия, е възможен. На второ място са изследвани стратегически и планиращи документи, които задават целите и посочват конкретни стъпки за научните разработки на територията на Европейския съюз, както и дават насоки за поведение от страна на разработчиците. След това са изследвани възможностите да се създаде изкуствен морален агент и, разбира се, някои етически теории, които биха могли да се приложат по отношение на поведението на изкуствения интелект към хората.

Хипотезата на дисертацията е следната: изкуственият интелект е възможен и когато той е в хуманоидна форма (т. нар. андроиди), етическите и социалните проблеми се намират в отношението човек-машина, а не в отговорността, която лежи върху работата на учените, конструкторите и програмните разработчици. Поради тази причина етическите теории, които могат да бъдат вложени в софтуерната разработка на изкуствения интелект, са деонтологичните.

Източниците на досертационната разработка включват не само класиците на философията като Аристотел, Декарт и Лок, Кант, Хегел, Ницше, Сартр, но

изследвани ще бъдат предимно съвременните трактовки в разбирането на проблема душа-тяло с оглед търсене на философската основа за изкуствения интелект, която ще даде и база за научните търсения на невронауката и разработките на програмистите и инженерите. Друга група източници са документи, приети на равнище Европейски съюз, които ясно очертават постигнатите резултати и поставените цели, които трябва да се реализират до 2020 година, както и етическия аспект на проблема. Не на последно място, изследвани са разработки, свързани с актуалната етическа мисъл, релевантни на проблема за навлизането на интелигентни машини-роботи в ежедневието.

Приложените методи са сравнителният анализ между различните философски направления - предмет на изследването, както и критичният метод с цел специфицирането на подходяща философска доктрина, обосноваваща възможността за конструиране на изкуствен общ интелект (artificial general intelligence), т. е. подобие на човешкия. Описателният метод е приложен по отношение на стратегическите и планиращи документи, приети от Европейския съюз, за да се изследват в конкретика приложенията на интелигентните машини у дома, на работа, в магазина и изобщо в ежедневния аспект в живота на хората. От друга страна, по необходимост е използван и критичният метод с цел да се установят конкретните начини, по които изкуственият интелект присъства в ежедневието на хората, както и в търсенето на етика, гарантираща благоприятно „поведение“ на машините. В собствено философски смисъл е приложен феноменологичният метод, тъй като спецификата на основния проблем и научните изследвания са в началната си фаза и поставяните въпроси все още са много повече от получените отговори.

КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

В първата глава като отправна точка приемам класическите теориите, касаещи отношението душа-тяло. За първи път Декарт променя средновековната парадигма, като се насочва към философски търсения за съзнанието (mind). Той набляга на разума, като по този начин ограничава разбирането за душата. В своята теория Декарт определя човека като двусубстанциален, т. е. състоящ се от душа/съзнание/духовна субстанция и тяло/материална субстанция. Според Декарт тялото и душата са два различни, но тясно свързани и взаимодействащи си елемента. Дуализмът има влияние върху философията и днес, тъй като редица философи опитват да докажат, че ако тялото и душата не са едно и също нещо, то те си взаимодействат толкова тясно, че в голяма степен се припокриват.

Втората класическа теория, която разглеждам, е тази на Джон Лок. Лок принадлежи към емпиризма, според който човешкото съзнание е Tabula rasa, която чрез опита се изпълва със съдържание. Фокусът тук е върху това, че човекът придобива знание за света чрез сетивата, а после придобитата информация бива приведена в систематизирано знание. Заслугата на Лок се състои в различаването на първичните (форма, покой) и вторичните (вкус, цвят, звук) качества на предметите. Днес философите наричат вторичните качества квалии, които са съдържание на опита. Тук трябва да се постави въпросът дали изкуствено интелигентните машини ще притежават квалии, т. е. ще имат ли субективен опит, който от своя страна е фундаментален за всеки интелект.

Разглеждам и съвременни теории за отношението душа-тяло, които биха могли адекватно да обосноват възможността за съществуването на изкуствен интелект. Днес дуализмът е също толкова актуален, колкото и за епохата на Декарт. Австралийският философ Дейвид Чалмърс смята, че съзнанието и опитът не са епифеноменални, а напротив – изискват вътрешна предразположеност, за да може познавателната способност да се манифестира. Според Чалмърс тялото и душата са две отделни субстанции, тъй като ако нещо е мислимо, то то е възможно. Това е класическа картезианска постановка. Философът предлага следния мисловен експеримент. Съществува човек

(напълно нормален и обикновен) на име Дейв и негово копие, което има тяло, но никакво съзнание. Ако физикалистите са прави, че физическото поражда съзнателното, то тогава зомби-Дейв би трябвало да притежава характеристики на съзнание. Но това не е така, следователно съзнанието е нещо същностно различно от тялото, то не е негов продукт. Съзнанието и тялото обаче тясно си взаимодействат.

Физикализмът твърди, че съзнанието е продукт на физическото. Според Томас Негъл съзнанието се състои в това да знаеш какво е да си организма, който си. Хората знаят какво е да са хора, но не и какво е да са прилепи или марсианци. В такъв смисъл съзнанието е субективно. Затова Негъл е критичен към физикализма, тъй като вярва, че ни липсва адекватен инструментариум да изучаваме обективно съзнанието и сме ограничени от собствената ни биология.

Освен двете представени съвременни теории са изследвани и две методологии, които по-конкретно обосновават изкуствения интелект. Първата методология е функционализмът, според който съзнанието и физическото са различни, но не могат да съществуват отделно. Той се опитва да открие и обоснове конституцията на физическото и къде е взаимосвързано с менталните състояния и процеси, така че да породи съзнанието. Функционалистите твърдят, че поведението е изходна реакция, т. е. резултат от сетивната входяща информация от околната среда, която активира нервната система. Мозъчните програми приличат на софтуер по отношение на това, че са последователни алгоритми за вземане на решения и предприемане на действия. С други думи, те са кодирани инструкции, вградени в мозъка. Съгласно функционализма мозъчните алгоритми (алгоритмите, според които мозъкът управлява системата и които общо взето представляват начини за вземане на решения и предприемане на действия) гарантират промените в системата с цел постигане на ново състояние, т.е. адаптиране. Мозъкът избира коя програма да изпълни, като този подбор е различен всеки път, тъй като е продукт от взаимодействие на различни части на мозъка, активирани от стимулите на околната среда. Перцепциите са структуриран модел на света, който позволява на мозъка да предскаже бъдещето и да вземе възможно най-доброто решение, така че да гарантира физическото

оцеляване. Наборът от мозъчни алгоритми се сравнява с компютърния софтуер и точно от това се интересува функционализмът. Особено важна постановка е, че софтуерът, програмата, алгоритмите, изпълнявани от мозъка, може да бъдат изпълнявани и от друг тип системи.

Според неврофилософията съзнанието е продукт на мозъчната структура и дейност, на невронните активност и мрежи. Предназначение на цялата нервна система е да обработва информацията, за да се произведе движение според обстоятелствата в околната среда. Неврофилософите смятат, че съзнанието ще се окаже процес подобен на метаболизма и всеки друг физиологичен процес, който познаваме. Неврофилософията може да открие връзката между физическото и феноменологичното, което убягва на функционализма. Според нея трябва да се направи и да се разбере разликата между съзнателните и несъзнателни състояния. Затова се използват два подхода – директен и индиректен.

Директният подход цели да идентифицира физическия субстрат, където възниква съзнанието, т. е. коя и каква мозъчна дейност характеризира съзнателните състояния. Невроучените провеждат редица и различни експерименти. Един от експериментите е следният: доброволец, поставен в непрозрачна кутия, възприема снимка на изгрев и снимка на лице, съответно с лявото и дясното око през два отвора, по отделно за всяко око. Оказва се, че доброволецът вижда поредица от залези и лица. Снимките се сменят с честота от 1, 5 до 10 милисекунди. Това означава, че мозъкът възприема стимулите постоянно, но не осъзнава образите едновременно – съзнателните състояния се редуват. Този експеримент показва връзката между невронната активност и осъзнаването. Може да се окаже, че двете явления протичат паралелно без свързаност помежду си. Целта на неврофилософията е да намери невронната идентификация на съзнанието. Това обаче изисква още много проучвания, изследвания и експерименти.

Според индиректния подход съзнанието се разглежда като глобално работно място (global workspace) по аналог на стените за обяви, където всеки

оставя информация, с която разполага и съответно може да намери, каквото му трябва. Това е само метафора, тъй като е известно, че мозъкът работи по друг начин. Информацията се предава от един неврон на друг, от един невронен слой на друг. Предполага се, че тук роля имат т. нар. дългоаксонни неврони, които създават необходимите условия за глобалното работно място. Трябва обаче да се изследват особеностите на състояния като сън, кома, вегетативните състояния, за да се получи пълна картина и в крайна сметка да се дефинира съзнанието.

Важни са разликите между начините, по които мозъкът и машините обработват постъпващата информация. Мозъкът използва постоянно изчезващи и възникващи невронни връзки. Живите организми обработват постъпващата информация, за да гарантират крайната си цел - да оцелеят и да се размножават. Всеки нов фактор в околната среда, дори и най-малката промяна, кара организмите да се адаптират и да синтезират напълно нови решения. Биологически погледнато, поведението е гъвкаво, изчисленията са неточни, защото се ръководят от нестабилни, променливи алгоритми. Машините от друга страна получават входяща информация, обработват я и след това изпращат изходна информация. Те извършват този процес съгласно строги правила и алгоритми, което представлява проблем за адаптацията.

В хода на разработките на изкуствения интелект става ясно, че математическите изчисления или играта на шах нямат значение за оцеляването. Алгоритмите на биологично зададеното поведение са неясни. Въпреки това, мозъкът възприема данни и структурирана информация от околната среда и променя алгоритмите на поведенческите реакции в съответствие с непрекъснато разширяващия се опит.

Трябва да се постави въпросът дали първо не трябва добре да познаваме естествената интелигентност и съзнанието както в широк, така и в тесен смисъл, за да го имитираме, или имаме нужда от технологията на изкуствения интелект, за да опознаем по-добре същността на човешкото съзнание. Първото твърдение е логично и може би дори естествено, докато второто може да се приеме за

рисковано толкова доколкото липсата на изчерпателно познание за човешкото съзнание може да поведе разработчиците по грешен път. Мисля, че двата подхода и гледни точки за перспективите за развитие на въпросните теории се допълват, защото колкото повече знаем за естествената интелигентност, толкова по-успешна имитация ще произведем. От друга страна, развитието на изкуствения интелект разкрива много характеристики и механизми на естествената интелигентност. В този смисъл, изучаването на естественото съзнание и разработването (в най-общ смисъл) на изкуствен интелект са двата противоположни пътя за достигане на една и съща крайна цел – да разберем и да знаем какво е съзнанието.

Въпреки че философията все още изучава човешкото съзнание и дебатира какво е да осъзнаваш, науката, в лицето на инженерите и програмистите, успешно въвежда изкуствено интелигентни машини (дори и в хуманоидна форма) във всекидневието на хората. Лабораторните проекти вече започват да стават достъпни за всички. Ето защо държавите, както и Европейският съюз разработват стратегически документи, очертаващи приоритетите и перспективите за бъдещото развитие на изкуствения интелект и роботиката (които вървят ръка за ръка). Навлизането на интелигентни машини в ежедневието ни ще повдигне въпроси: какво е мястото им в нашата действителност?; достатъчно ли са интелигентни, за да се вземат решения за себе си и дори за хората?; следователно, ще трябва ли да имат права?; какво трябва да бъде отношението между хората и роботите? Хората ще бъдат изправени пред предизвикателството да се замислят за същността си в перспективата на новото присъствие в живота им. Футуристите прогнозират различни сценарии за ролята на роботите в човешкото общество и е настъпил моментът, когато трябва да решим какъв вид интелигентни машини искаме – ще останат ли инструменти, каквито винаги са били, или ще създадем машина, притежаваща интелект подобен на човешкия и която ще има потенциал да се превърне в равна на човека, в конкурент или дори в нещо по-висше. Как ще да се подготвим за отговорността и последствията?

Във втората глава на дисертацията са разгледани два стратегически документа, издадени от Европейския съюз. Това са "Роботика 2020 - Програма за стратегически изследвания за роботика в Европа" в рамките на Хоризонт 2020 и ЕУРОН Пътна карта по Робоетика.

Очакваните резултати са положително развитие по отношение на конкурентоспособността на индустрии като селско стопанство, транспорт, здравеопазване, сигурност и комунални услуги. Планирано е най-големите продажби на роботи в глобален мащаб да са именно в тези области. От настоящите 22 милиарда евро световни приходи роботиката трябва да постигне на годишни продажби на между 50 и 62 милиарда евро до 2020 г.

Думата "робот" е с чешки произход и означава "машина, която работи". Освен приоритетите Роботика 2020 уточнява видовете роботи, техните специфични функции и технологиите, които ще се използват и комбинират. Основните категории на класификацията са според околната среда, в която работят; според начина, по който си взаимодействат с потребителите; според физическата им форма и основната функция, която изпълняват.

Роботите ще бъдат разработени така, че да притежават способности, които ще определят конкретното им предназначение. Способностите са: Конфигурируемост, Приспособимост, Способност за интеракция, Надеждност, Способност за движение, Способност за извършване на манипулации, Способност за перцепция, Способност за вземане на решения, Познавателна способност.

Когнитивната способност представлява разбиране на средата, в която машината се намира, и планиране на съответните адекватни действия. Средата обикновено е неструктурирана и се възприемат фрагменти от нея, като мозъкът „добавя“ необходимите елементи. Поставяните задачи също ще бъдат неструктурирани – естественият език може да бъде мъгляв и двусмислен, но се очаква роботите да разбират правилно командите, отправени към тях. В когнитивната способност се включва и естествената интеракция с хората, които ще общуват с роботите, т. е. роботите трябва да разбират конкретните

настроения и емоции на хората. В крайна сметка роботите трябва да могат да се учат от миналия опит и допуснатите грешки. Когато става дума за познавателна способност, включваща разбиране, инженерите нямат предвид способност, съответстваща на херменевтичното разбиране на контексти у хората, а способност, представляваща трупане на опит от една страна, и от друга – използването на този опит при вземане на решение какво поведение е най-подходящо в конкретната ситуация.

Роботи ще се използват в много области – за цивилни цели (спасителни операции); в индустрията; в сферата на услугите. Най-важната в социално отношение разработка са роботите-асистенти. Те ще се грижат за малки деца, но по-голямата целева група са възрастните хора. Ползите от роботите-асистенти са несъмнени – възрастните хора ще запазят мобилността и независимостта си и в крайна сметка достойнството си като едновременно с това ще получават грижите, от които имат нужда, включително когнитивна подкрепа и социални контакти. В допълнение, асистентите ще осигурят свободно време за семействата на възрастните.

Друго важно приложение е в сферата на edutainment – образование плюс забавление. Роботите ще мотивират децата да се занимават с наука и дори спорт, ще ги запознават с различни видове среда и ситуации, ще стимулират работата в екип и ефективната комуникация, творчеството и сътрудничеството.

Според стратегическите документи икономическите ползи са много. Ще се създадат работни места; ще се намалят производствените разходи; приходите ще се увеличат почти тройно до 50-62 млрд евро при настоящите 22 млрд. Социалните ползи включват разтоварване от ежедневната домакинска работа; намаляване трафика на хора; преодоляване на сексуалната експлоатация.

Необходимо е да се предприемат превантивни мерки, за да се преодолеят и намалят рисковете от още по-широкото навлизане на интелигенти машини в ежедневието ни. Роботика 2020 твърди, че отговорността е на онези, които създават роботите. Айзък Азимов предлага роботите да са собственост на производителите, дори когато се използват в домакинствата, което означава, че

юридическата отговорност за вреди и щети, причинени от роботи, се носи от компаниите. Това решение трябва да бъде уредено много добре, тъй като съществува реалната възможност за недоброжелателни потребители. Първата стъпка е да се прецени какви са рисковете, кой ще носи отговорността и как да се регулира новата социална ситуация, включваща нови агенти. Трябва да се мисли за национално, наднационално и международно право и стандартизация, като се вземат предвид социалните, културните, етическите и други особености и очаквания.

Социалните рискове са свързани с очакването, че роботите ще навлязат в ежедневието, което от своя страна може да доведе до безработица и трансформиране на моделите на заетост. Очаква се роботите да станат основната работна сила, която ще се изпълнява опасни и нежелани задължения не само в промишлеността. Друг риск е някои части от обществото да бъдат лишени от достъп до тази технология, поради прекалено високата крайна цена.

В системното разработване на роботите трябва да се включат стандарти, норми и закони, гарантиращи безопасност. Трябва да се установят протоколи за сигурност, предотвратяващи незаконна или неподходяща употреба на роботите. Самите роботи трябва да са проследими по аналог на черните кутии на самолетите, както и разпознаваеми чрез серийни и идентификационни номера. Последно, но не и по важност роботите трябва да осигуряват неприкосновеността на личния живот, а именно гарантиране сигурността на потребителската парола и на личните данни и информация, придобита по време на изпълнението на задачите.

Поради явното навлизане на роботите в ежедневието ни, в дисертацията са разгледани алтернативи на възможните отношения между хората и роботите. Една възможност е роботите, макар и притежаващи интелект, да си останат предмети, инструменти, уреди, които ни помагат в ежедневието и го улесняват. Машините ще останат „то“. Възможно е обаче да се превърнат в „той“ или „тя“. Да започнем да ги възприемаме като квази-други. Предпоставка за такова развитие е фактът, че говорим на растенията у дома, избираме име на колата,

привързваме се към предмети. Вече се говори за секс-роботи, което е още една предпоставка за възникване на привързаност към машините. Можем да започнем да приемаме роботите като достойни за партньорство, дори за равни. Това ще промени обществото, защото ще се появи нов социален агент. Роботите може да се обособят в отделна група със своя култура и традиция, а защо не и религия и философия. Тогава има риск от опозицията „ние – те“ като предпоставка за конфликт. Възможно е роботите да бъдат възприемани като заплаха по начина, по който е възприет тъкачният стан и последвалите бунтове, характерни за индустриалната революция.

Днес виждането за интелигентните роботи съвпада с възгледа на Аристотел за робите. Робите са устроени за работа и им липсва далновидност, не могат да разсъждават, те са инструменти за употреба на други инструменти, не могат да изпълняват повече от една задача. Разбира се, робите нямат права, не може да бъдат граждани и са собственост на господарите си. В контекста на отношение господар-роб не може да не се приложи Хегеловата диалектика. Двете страни на отношението зависят една от друга, но робите са тези, които притежават възможността да посочат какво не са. По този начин могат да отхвърлят господарите, да се определят за достойни. Това води до обръщане на ценностите. В исторически план такова събитие е възникването на християнството и приемането му за официална религия на Римската империя. При тези условия според Ницше робите са добри, защото просто са обратното на злото. Те не се самоутвърждават, а отрицават другия. Злото е другият, който не съм аз и не принадлежи към моята група.

В трета глава разглеждам отношението човек-робот и по-конкретно дали роботите имат морална значимост. Въпросът трябва да се мисли от гледна точка на факта, че ние хората сме спиецисти, т. е. нашият вид е готов да жертва интересите на всички останали биологични видове в името на задоволяването на собствените си интереси. В хода на историята човечеството все повече осъзнава, че не може да има подобна нагласа, както и значението на природата и живите организми и планетата като цяло най-малкото заради отговорността, която носим към бъдещите поколения.

За Кант само съществата, които имат съзнание и особено самосъзнание, имат морална значимост. Единствено хората имат морални качества и значимост поради простия факт, че в края на XVIII век само хората са разглеждани като самосъзнателни. Подобна позиция оправдава спиецизма, но днес знаем, че тя няма основания. Но... е приложима за машините, следователно и за роботите.

Комуникацията ни с други хора, връзките, взаимодействията и сътрудничеството също са изградени върху външната проява на съзнателност. Фактът, че този/ва, с когото/което общувам се държи и реагира много подобно на мен, ме кара да вярвам, че той/тя притежава съзнание като мен. В допълнение ако някой желае нещо, то той има правото да го притежава. Оттук следва, че ако роботите осъзнават и разбират какво е свобода, то те имат право да бъдат свободни и съответно ние трябва да признаем това право.

Как можем да сме сигурни, че всеки робот може да разбере какво е свободата? Скептиците винаги ще настояват, че роботите са неспособни да искат да бъдат свободни, те дори не могат да схванат идеята за свободата. Те са програмирани да се държат по определен начин и просто изпълняват вградените им алгоритми. Ето защо на роботите ще бъде отказана свобода.

Независимо от факта, че в края на XVIII век нямаме задължения към животните, както днес нямаме задължение към машините, етичното отношение към тях ни прави по-добри като човечество. Според Кант този, който е жесток към животните, става жесток и към хората, а напротив - нежното отношение към глупавите развива хуманни чувства към човечество. Включването на роботите в живота ни изисква да се отнасяме към тях етично и добре, защото в противен случай ще се превърнем в неморални същества. Това означава, че не трябва да се интересуваме дали машините имат морална значимост, а да приемем това като факт.

Когато мислим роботите като морални агенти, трябва да се вземат предвид и емоциите като елемент за вземането на морални решения. Ако изключим емоциите, следва, че моралната преценка се основава на разума и съществува ясен алгоритъм за взимане на морални решения. Емоционалността

обаче е фактор за свободата на избор. Липсата на емоции е по-опасна - някои престъпници са способни на безумни престъпления и психиатрите ги диагностицират като неспособни да изпитват емоции и съпричастност. В този смисъл, разработването на роботи без емоции ще доведе до опасни роботи „психопати“. Освен това те няма да бъдат морални агенти, защото ще следват правилата без никаква реална етична загриженост и ще им липсва способността да разбират какво са ценностите.

Хората са емоционални, но това може да се разглежда като слабост, тъй като емоциите пречат на способността за взимане на морални решения и зачитане правата на другите членове на обществото. Съществува напрежение между емоцията и взимането на морални решения. От една страна, трябва да сме съпричастни към другите, да бъдем чувствителни и да разбираме страданието и щастието им. От друга страна, емоциите ни пречат да взимаме правилни решения, защото попадаме под тяхното влияние и те замъгляват преценката ни. Хората винаги намират баланса между субективните си емоции и чувства и обективната реалност, изискваща спазване на моралния закон, или следването на моралните правила и принципи. Наличието на емоции у роботите ще бъде ли достатъчно, за да ги приемем като хора?

Истинският проблем е технологичен - може ли да се създаде морален робот? Човешкият морал е сложен, но разчита на феноменологичния аспект на съзнателната дейност и ако инженерите успеят да произведат роботи с емоции, ментални състояния и морална загриженост, тогава има реална възможност хората и роботите да живеят заедно и да се допълват взаимно.

Ако роботите ще бъдат част от човешкото ежедневие, ще се очаква те да имат етично поведение и да носят отговорност за действията си точно като хората. Истинският проблем е как да програмираме робот, който взема обосновани и адекватни морални решения на етична основа, съвместими с човешките. В случая на изкуствено интелигентните, етично действащи машини по-скоро трябва да се говори за функционална артефактуална отговорност.

Възможни са два подхода към отговорността на артефактите. Според класическия подход на предметите не може да се приписва отговорност. От друга страна, според прагматичния подход артефактите трябва да носят отговорност в съответствие с тяхната компетентност и функция. Аргументът срещу очакването на морална свобода на избор и съответно носенето на отговорност е, че изкуствено интелигентните агенти няма да откриват никакъв смисъл във взимането на решения и съответното поведение. Причината за това е липсата на емоции. Емоциите са важна част от моралната преценка, но тяхната особена роля все още е неясна. Независимо от това развитието на синтетичните емоции бележи бърз напредък. Друг аргумент е, че никоя машина не може да бъде морална, докато не придобие ментални състояния, тъй като моралът се счита за проява на съзнанието.

Моралното поведение се ръководи интуитивно от определени закони и правила. Но при по-дълбок поглед се оказва, че хората могат да намират дори и най-малките нюанси и да взимат съвсем различни решения в привидно подобни ситуации. Това е предизвикателството за учените, решени да разрешат загадката и да предоставят на инженерите алгоритъм, с който да изградят етични антропоморфни машини. Ако ограничим поведението на роботите, поемаме риск машините да взимат праволинейни, но погрешни решения.

За да се избегне такъв проблем, възможно решение е вграждането на етическа теория у роботите. За тази цел разглеждам сартровата екзистенциална етика като консеквенционалистка етическа теория. Основният постулат е, че съществуването предхожда същността, от който следва свободата на човека да реализира себе си съгласно собствения си избор. Изборът обаче трябва да се ръководи от това, че избирайки себе си, човек избира цялото човечество. От тук следва и отговорността, която човек сам носи за своите решения.

От друга страна, деонтологичната етическа теория на Кант поставя в центъра дълга. Моралните понятия са а приорно присъщи на съзнанието. Моралният акт се ръководи от императив, който е насочен срещу склонностите. Всеки акт, подчинен на необходимостта и на дълга, има морална стойност.

Волята е свободна доколкото е в съгласие с моралния закон и дълга, но едновременно с това е и несвободна, защото е в конфликт с инстинктите. В това се състои моралната ситуация. Моралната постъпка е резултат на волята, мотивирана от инстинктите, но подчинена на дълга.

Пред инженерите и програмистите стои проблемът да създадат именно такъв интелект – притежаващ програма, механизъм, според който да ръководят поведението си. Но от друга страна, трябва да имат „малкия недостатък“ да притежават склонности или инстинкти, за да съществува необходимият за моралното поведение конфликт. Един от проблемите на деонтологичните теории е, ако принципите и правилата конфликтуват, как трябва да постъпи субектът.

Айзък Азимов формулира трите закон на роботиката:

1. Роботът не може да причини вреда на човек или с бездействието си да допусне на човека да бъде причинена вреда.

2. Роботът е длъжен да се подчинява на човека, ако това не противоречи на Първия закон.

3. Роботът е длъжен да се грижи за собствената си безопасност, ако това не противоречи на Първия и Втория закон.

Според Азимов трите закона са проява на повечето, ако не и всички етични норми и съответно системи. Третият закон е свързан със самосъхранението, което в края на краищата е от съществено значение за всички. Всеки човек е длъжен да се подчинява на правилата и законите и да спазва социалните норми. Всеки спазва традицията, дори и ако животът му е поставен в опасност. Ние уважаваме авторитета на лекари, началници, правителството и приятели. И това е вторият закон на роботиката. Накрая, ние хората сме приятели и се обичаме и защитаваме; рискуваме собствения си живот, за да спасим близките си. Именно такова поведение е заложено в Първия закон на роботиката.

Като цяло разказите в сборника "Аз, роботът" са позитивни и оптимистични за запазването подчинената позиция на роботите. Хората са създателите, а роботите са създадените. И така трябва да бъде. Азимов постоянно демонстрира превъзходството на хората, тъй като те винаги надхитрят машините, въпреки

известното високомерие, недоверие и чувство за превъзходство у роботите. Според Азимов хората са производители или дори творци на роботите и затова правят всичко възможно да поддържат статуквото – машините извършват тежък физически труд и могат да бъдат определени като робите в едно бъдещо общество. Основният въпрос тук е дали хората ще успеят да удържат превъзходството си и дали ще измислят стратегии за реакция при хипотетичен бунт на роботите.

Азимов предлага законите да бъдат преведени на математически език и да бъдат част от целия програмен пакет, „качен“ в позитронния „мозък“ на робота. Това означава, че роботите ще получат наготово всички знания, които са им нужни и следователно няма да има необходимост да учат. По този начин законите ще са непроменяеми и няма да се развиват. Предложението на Азимов за неизменими закони противоречи на съвременните концепция и амбиции за създаване на изкуствен интелект и възможно следствие би било да се откажем от идеята за изкуствения интелект или пък да търсим друго решение, което може да се състои във възпитание в социално приемливо етично поведение.

От една страна, трябва да се има предвид възможността интелигентните роботи да бъдат третирани като умни и полезни машини. Ако те са просто машини, каквито ги познаваме днес, няма да има никакъв проблем от гледна точка на пряка физическа заплаха за хората. От друга страна, има възможност машините с изкуствен интелект да бъдат приети като интелектуално равни на хората. Ако изкуственият интелект е затворен в „кутия“ (в настолния компютър, например) или е с форма, различна от хуманоидната, тогава няма да се приемат за сериозна заплаха, за разлика от такъв в хуманоидна форма. Апокалиптичната нагласа се свързва с андроидната форма на машините и тъй като притежават интелект, те са мислени като потенциална заплаха за хората.

Програмирането на идеи и алгоритми за поведение, водено от дълг/програма и желание за служба е начинът, по който можем да запазим превъзходството си. Настоящото проучване показва, че деонтологичните теории са по-подходящи за контролиране на интелигентните машини, тъй като

консеквенционалистката теория на екзистенциализма се отнася за хората, които имат високо ниво на самосъзнание и следователно са напълно способни да носят отговорност за своите действия. Законите на роботиката, предложени от Азимов, са опит за преформулиране на деонтологичните теории за морала. Настоящото проучване показва, че изкуственият общ интелект няма да бъде постигнат скоро предвид това, че етичното и морално поведение у роботите е все още на дневен ред.

ЗАКЛЮЧЕНИЕ

Резултатите от направеното изследване успешно доказват хипотезата на дисертацията. Изкуственият интелект е възможен и когато той е в хуманоидна форма (т. нар. андроиди), етическите и социалните проблеми се намират в отношението човек-машина, а не в отговорността, която лежи върху работата на учените, инженерите и програмните разработчици. Поради тази причина етическите теории, които могат да бъдат вложени в софтуерната разработка на изкуствения интелект, са деонтологичните.

Настоящата дисертационна разработка показва, че социалното участие на роботите с изкуствен интелект в ежедневието на хората не само на работните им места, но и в домовете им, неразривно поставя проблема за етичното поведение на машините към хората и на хората към машините. На този начален етап на научните постижения в областта на изкуствения интелект отговорността за поведението на роботите се носи основно от проектантите, инженерите, софтуерните разработчици, учените и изобщо от хората, участващи в създаването на изкуствения интелект и неговия носител (хардуер).

Настоящото изследване успешно е изпълнило поставените цел и задачи. Въпреки това, остават още области за проучване като кои конкретно деонтологични теории трябва да се адаптират, за да се превърнат в софтуер и да бъдат вградени в умните машини. Разбира се, част от проблема остава в сферата на софтуерния дизайн и разработки и затова философите и програмистите трябва да работят в тясно сътрудничество. Предложените от Азимов закони на роботиката са само пример, мисловен експеримент и трябва да бъдат подложени на емпиричен тест за недостатъци и по-нататъшни разработки. Последно, но не по значение, е, че дисертацията е с „отворен край“, тъй като философският анализ на откритията и постиженията на невронауката; на изкуствения интелект по отношение на софтуера и хардуера; както и разширяващата се политическа подкрепа, ще допринасят за продължаващи изследвания и проучвания.

ПРИНОСНИ МОМЕНТИ НА ИЗСЛЕДВАНЕТО

Изкуствен интелект е възможен като изкуствен тесен интелект на базата най-вече на функционализма, както и на приносите на невронауката. Основният проблем, за да се създаде изобщо някакъв интелект, подобен на човешкия, се състои в намирането на начин за програмиране и вграждане в изкуствения интелект на мотиватори, равностойни на тези, които движат човешкото поведение. Става дума за биологичните характеристики и потребности на човешкото тяло като стремеж към оцеляване (намиране на храна, избягване от хищници) и размножаване, поведение, следващо принципа на удоволствието и неудоволствието.

Доказано е, че машини, носители на изкуствен интелект, ще навлизат все повече в ежеднението на хората в областта на домакинството, обслужването, възпитанието и образованието, медицината. Навлизането на роботите в ежедневието ни ще създаде работни места и ще реши проблеми като грижите за старите хора и децата, ще се осигури повече свободно време за хората в активната, най-полезната за обществото възраст. От друга страна, трябва да се обърне внимание на рискове като отчуждение, размиване на границата между реално и виртуално, прекалена привързаност към машините, които ще придобият ново измерение.

Настоящата дисертационна разработка показва, че социалното участие на роботите, носещи изкуствен интелект в ежедневието на хората не само на работните им места, но и в домовете им, императивно поставя проблема за етичното поведение на машините към хората и на хората към машините. Но на този етап научните постижения в областта на изкуствения интелект са в начален стадий, което означава, че отговорността за поведението на роботите се носи основно от проектантите, инженерите, софтуерните разработчици, учените и изобщо от хората, участващи в създаването на изкуствения интелект и неговия носител (хардуер).

Проведеното изследване на консеквенционалистката етическа теория на екзистенциализма и деонтологичната теория на Кант показва, че

деонтологичните теории са по-подходящи за превръщане в алгоритми и протоколи за поведение и следователно за вграждане и програмиране на машини, носители на изкуствен интелект. Трите закона на роботиката са конкретен пример за деонтологично базирани правила за поведение, като в литературата на писателя-фантаст се изследват редица възможности за развитие на отношенията между хората и роботите, както и на еволюция на мисленето – продукт на изкуствения интелект.

БИБЛИОГРАФИЯ КЪМ ДИСЕРТАЦИОННИЯ ТРУД

News 7, 21.03.2014. *Робот от БАН носи лекарства и помага на възрастни хора,*

Anderson, Susan Leigh, 2008. Asimov's ''Three Laws of Robotics'' and Machine. *AI & Soc*, pp. 477-493.

Anon., 2015. *ИНСТИТУТ ПО ИНФОРМАЦИОННИ ТЕХНОЛОГИИ.* [Online] Available at: http://www.iit.bas.bg/IIT_bg/index_bg.html

Aristotle, 1920. *Politics.* Oxford University Press.

Asimov, Isaak, 1982. Lenny. In: *The Complete Robot.* London: Nightfall Cor, pp. 368-384.

Asimov, Isaak, 1982. Reason. In: *The Complete Robot.* Nightfall Inc.

Asimov, Isaak, 1982. Robbie. In: *The Complete Robot.* Doubleday, p. 557.

Asimov, Isaak, 1990. *Robot Visions*.

Asimov, Isaak, 2000. *The Bicentennial Man.* Gollancz.

Block, Ned, 1980. Introduction: What is Functioanlism. In: N. Block, ed. *Readings in Philosophy of Psychology.* Harvard University Press, pp. 171-184.

Byrne, Alex, 2006. What Mind-Body Problem? Understanding consciousness may be easier then we thought. *Boston Review,* 04 May.

Capek, Karel, 2014. *R.U.R. (Rossum's Universal Robots).* Adelaide: eBooks@Adelaide.

Carter, Matt, 2007. *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence.* Edinburgh University Press.

Chalmers, David, 2007. Naturalistic Dualism. In: S. S. Max Velmans, ed. *Blackwell Companion to Consciosness.* Blackwell Publishing Ltd, pp. 358-366.

Chappell, Vere, 2004. *Locke on Consciousness.* draft.

Churchland, Paul M., 2007. Catching Consciousness in a Reccurent Net. In: *Neurophilosophy at Work.* Cambridge Univesity Press, pp. 1-17.

Churchland, Paul M., 2007. Into the Brain: Where Philosophy Should Go from Here. In: *Neurophilosophy at Work.* Cambridge University Press, pp. 232-238.

Churchland, Patricia S., 1986. *Neurophilosophy. Toward a Unified Sience of Mind/Brain.* Cambridge, MA: MIT Press.

Churchland, Patricia S., 2002. *Brain-Wise. Studies in Neurophilosophy.* The MIT Press.

Clarke, Roger, 1994. Asimov's Laws of Robotics: Implications for Information Technology. *Computer Milieux,* pp. 57-66.

Coeckelbergh, Mark, 2010. Moral Appearances: emotions, robots, and human morality. *Ethics and infomration Technology,* Volume 12, pp. 235-241.

Coeckelbergh, Mark, 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology,* pp. 209-221.

Coeckelbergh, Mark, 2010. You, robot: on the linguistic construction of artificial others. *AI and Society,* 10 August , pp. 61-69.

Davenport, David, 2013. The Two (Computational) Faces of AI. In: V. Muller, ed. *Philosophy and Theory of Artificial Intelligence.* Springer, pp. 43-58.

Dennett, Daniel, 1973. Mechanism and responsibility. In: T. Honderich, ed. *Essays on Freedom of Action.* Routledge & Keegan Paul.

Descartes, Rene, 1996. Meditations On First Philosophy. In: *The Philosophical Works of Descartes.* Cambridge University Press.

Ecclesiastes, *2:24.*

Floridi, Luciano and J. W. Sanders, 2001. *On the Morality of Artificial Agents,* Oxford: Oxford University Press.

Fodor, Jerry, 1981. *The Mind-Body Problem,* Scientific American, Inc., available at: http://www.lscp.net/persons/dupoux/teaching/QUINZAINE_RENTREE_CogMaster_2010-11/Bloc_philo/Fodor_1981_mind_body_problem.pdf

Frankfurt, Harry G., 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy,* 68(1), pp. 5-20.

Gert, Bernard, 1988. *Morality.* Oxford: Oxford University Press.

Gips, James, 1991. *Towards the Ethical Robot.* Perdido Key, Florida, MIT Press.

Gollwitzer, Peter M., 1990. Action Phases and Mind-Sets. In: *Handbook of Motivation and Cognition.* New York: The Guilford Press, pp. 53-92.

Crnkovic, Gordana Dadig, B. C., 2012. Robots: ethical by design. *Ethics and Information Technology*, pp. 61-71.

Gulick, Robert Van, 2007. Functionalism and Qualia. In: S. S. Max Velmans, ed. *The Blackwell Companion to Consciousness.* Blackwell Publishing Ltd, pp. 381-394.

Hegel, Georg W. F., 1977. *Phenomenology of Spirit.* Oxford: Oxford University Press.

Hoffman, Paul, 2008. The Union and Interaction of Mind and Body. In: J. B. a. J. Carreiro, ed. *A Companion to Descartes.* Blackwell Publishing, pp. 390-403.

Holton, Richard, *Introduction to Philosophy: Free Will 2*, [Online], available at: http://web.mit.edu/holton/www/edin/introfw/introfw2.pdf

Kant, Immanuel, 1997. *Critique of Practical Reason.* Cambridge: Cambridge University Press.

Kant, Immanuel, 1997. *Lectures on Ethics.* Cambridge University Press.

Kant, Immanuel, 1998. *Critique of Pure Reason.* Cambridge University Press.

Kant, Immanuel, 2004. *Prolegomena to Any Future Metaphysics: That Will Be Able to Come Forward as Science: With Selections from the Critique of Pure Reason.* Cambridge: Cambridge University Press.

Kant, Immanuel, 2011. *Groundwork for the Metaphysics of Moral.* Cambridge University Press .

Locke, John, 1999. *An Essay concerning Human Understanding.* The Pennsylvania State University.

Mandik, Pete, 2007. The Neurophilosophy of Consciousness. In: S. S. MAx Velmans, ed. *The Blackwell Companion to Consciousness.* Blackwell Publishing Ltd, pp. 418-429.

McCauley, Lee, 2007. AI Armageddon and the Three Laws of Robotics. *Ethics and Information Technology*, pp. 153-164.

Nagel, Thomas, 1974. *What is it like to be a bat?*. October, Volume 83, pp. 435-50.

Nagel, Thomas, 1998. Conceiving the Impossible and the Mind-Body Problem. *Philosophy*, July, Volume 73 No 285, pp. 337-352.

Nietzsche, Friedrich, 2006. *On Geneology of Morality.* Cambridge University Press.

Norhoff, Georg, 2004. *Philosophy of the Brain: The brain problem.* John Benjamins Publishing.

Nussbaum, M. C., 2001. *Upheavals of Thought: The Intelligence of Emotions.* 1st ed. Cambridge: Cambridge University Press.

Petersen, Stephen, 2006. *The Ethics of Robot Servitude,* Available at: *http://stevepetersen.net/professional/petersen-robot-servitude.pdf*

Robotics 2020, 2013. *Robotics 2020 Strategic Research Agenda for Robotics in Europe.*

Rohlf, Michael, "Immanuel Kant", *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), available at: http://plato.stanford.edu/archives/spr2016/entries/kant/

Rozemond, Marleen, 2008. Descartes's Dualism. In: J. B. a. J. Carreiro, ed. *A Comapnion to Descartes.* Blackwell Publishing, pp. 372-389.

Sartre, Jean-Paul, 1946. *Existentialism is a Humanism,* Available at: http://homepages.wmich.edu/~baldner/existentialism.pdf

Sartre, Jean-Paul, 2003. *Being and Nothingness.* London: Routledge.

Singer, Peter, 1990. *Animal liberation: All animals are equal* in [Online] Available at: http://www.uvm.edu/rsenr/wfb175/singer.pdf

Smart, J. J. C., "The Mind/Brain Identity Theory", *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Available at: http://plato.stanford.edu/archives/

win2014/entries/mind-identity/.

Sommervile, I., *Models for Responsibility Assignment.* [Online] Available at: http://archive.cs.st-andrews.ac.uk/STSE-Handbook/Papers/Models

ResponsibilityAssignment-Sommerville.pdf.

Tooley, Michael, 1972, Abortion and Infanticide In *Philosophy and Public Affairs*, Vol. 2, pp. 35-67, available at: http://eclass.uoa.gr/modules/document/file.php/

PPP504/Michael%20Tooley,%20Abortion%20and%20infanticide.pdf

Veruggio, Gianmarco, 2007. *EURON Roboethics Roadmap*

Vihvelin, Kadri, "Arguments for Incompatibilism", *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), available at: http://plato.stanford.edu/archives/fall2015/entries/incompatibilism-arguments/

Wallach, Wendell, 2010. Robots minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology*, Volume 12, pp. 243-250.

Wallach, Wendell and Colin Allen, 2008. *Moral Machines: Teaching Robots Right From Wrong.* Oxford: Oxford University Press.

Warren, Mary Anne, On the Moral and Legal Status of Abortion, In *Biomedical Ethics*. 4th ed. T.A. Mappes and D. DeGrazia, eds. New York: McGraw-Hill, Inc. 1996, pp. 434-440, available at http://instruct.westvalley.edu/lafave/warren_article.html

Waseda University, H. R. I., 2000. *Activities,* Tokyo

Young, John Zachary, 1988. *Philosophy and the Brain.* Oxford University Press.

Азимов, Айзък, 1993. *Аз, роботът.* сп. "Фантастични истории".

Азимов, Айзък, 1993. *Гоненица*. сп. "Фантастични истории".

Герджиков, Сергей, 2014. *Изкуствена и естествена интелигентност,* конференция "Съзнание", СУ.

Кант, Имануил, 1963. *Об изначальном злом в человеческой природе в Сочинения в шести томах.* Москва: Академия наук СССР Институт философии Издательство социально-экономической литературы „Мысль“.

Кант, Имануил, 1974. *Критика на практическия разум.* София: Издателство на Българската академия на науките.

Кант, Имануил, 1974. *Основи на метафизика на нравите.* София: Наука и изкуство.

Кант, Имануил, 1993. *Пролегомени към всяка една бъдеща метафизика, която би се явила като наука.* София: Лик.

Карагеоргиева, Анета, 2008. *Философия на съзнанието.* София: ИК „Библиотека 48".

Сартр, Жан-Пол, 1996. Екзистенциализмът е хуманизъм. In: *Философски текстове.* София: Философска фондация Минерва.

ПУБЛИКАЦИИ ПО ТЕМАТА НА ДИСЕРТАЦИЯТА

Robots as modern slaves

REVISTA PAPELES Volume 5 No 9, pp. 68-74, published on 4th Aug 2014 availabel at: http://csifesvr.uan.edu.co/index.php/papeles/article/viewFile/310/230, издание на Университет „Антонио Нариньо“, Богота Колумбия.