ВИДОВЕ ИЗКУСТВЕН ИНТЕЛЕКТ – ТЕСЕН, ОБЩ И СУПЕР ИЗКУСТВЕН ИНТЕЛЕКТ. РИСКОВЕ И ЕТИЧЕСКИ ПРОБЛЕМИ

МАРИАНА ТОДОРОВА

Институт за изследване на обществата и знанието, БАН

mariana_g_todorova@yahoo.com

TYPES OF ARTIFICIAL INTELLIGENCE - NARROW, GENERAL AND SUPER AI. SOME RISKS AND ETHICAL ISSUES

MARIANA TODOROVA

Institute for the Study of Society and Knowledge, BAS

Abstract

The paper explains what are the current types of artificial intelligence-narrow, general and super AI. The narrow, primary AI is already achieved. On that base, it suggests what might be the dimensions of their developments and application. Then it provides the author's vision what might be the following risks of AI implementations. Among them are: loss of sovereignty and cognitive abilities, desocialization and dehumanization, refusal to accept responsibility, existential crisis, etc.

Keywords: types of Artificial Intelligence; narrow, general and super AI; risk; ethical issues; loss of sovereignty; dehumanization; existential crisis

За да оценим възможността за еволюцията на изкуствения интелект, трябва ясно да очертаем съществуващите концепции за трите форми на проявлението му. Първият – тесен, първичен изкуствен интелект, вече е постигнат. Предизвикателствата и потенциалните опасности се срещат още в реализацията на това ниво. Но по-големите и неуправляеми рискове биха дошли от генералния и супер изкуствения интелект. За разлика от всички индустриални революции до момента, човекът се среща не със „средство", а с конкурираща същност под формата на небиологичен интелект.

1. „Тесен, първичен, слаб изкуствен интелект“

Нарича се така, защото е създаден да адресира ограничени, единични задачи. Голяма част от сега функциониращите системи на изкуствен интелект са именно тесен такъв, който има ясно дефинирана функции. Това е технология, която позволява на високо-функционални системи да преповтарят, а дори да надминават човешките способности по отношение на възложени задачи [1]. Такива са: „Уотсън” [Watson-суперкомпютърът на АйБиЕм], „Сири” [Siri] на Епъл и Кортана [Cortana] на Майкрософт. Всеки софтуер, който използва достижения като: събиране на данни , машинно обучение, разпознаване на модели и обработване на естествени езици, за да извършва автономно прости решения, може да се възприеме за тесен изкуствен интелект. Към тази категория може да причислим и спам-филтри (нежелана поща и съобщения), автономните коли и новините на Фейсбук. Програмата „Уотсън” комбинира в себе си метода на експертните системи, машинното обучение и обработката на естествени езици. Това, което ни демонстрира, дори побеждавайки в играта „Jeopardy”, или като подсигуряване на лекарски и адвокатски консултации, е определена степен на интелигентност в дадено поле, но без пълнотата, сложността и асоциациите при разсъждение, на които човек е способен. Друга важна отличителна черта е, че този тип интелект няма съзнание, самоосъзнаване като същност (не може да мисли сам за себе си-саморефлексивно) и автентична интелигентност, която да съответства на човешката. Този тип програми не могат да оперират добре и да се адаптират към реалности, извън фиксираната задача и цели.

Теснията изкуствен интелект вече е реалност, която все повече ще се разширява, в смисъла на функционалност и обхват. Позитивното при него е, че той към момента е напълно контролируем, но същевременно притежава достатъчно капацитет да извършва безброй различни задачи. Именно тази форма на изкуствен интелект е най-подходяща за сътрудничество, чрез която човешките способности се подсилват, но отпада тревогата за превъзходството над хората и възможните последици от това.

„2. Общ/ силен изкуствен интелект“

Той описва системи, с разбиращо знание и когнитивни способности, които го правят неотличим от естествения интелект, въпреки, че скоростта и способността на

обработка на данните е доста по-бърза и с несравним обем и капацитет. Такъв изкуствен интелект ще може да прави систематични обобщения, да разсъждава и да се учи от опита си [2]. Той ще притежава разбиране с преживяване на заобикалящата го среда, също като хората, но с много по-бързи реакционни възможности. По този път би могъл да се появи специфичен вид от нови (небиологични) същности.

Критериите за установяване на общ/ широк изкуствен интелект са [Goertzel, Penachin, 2007]:разсъждаване и обосноваване (използване на стратегии, решаване на сложни задачи и отсъждане в ситуации на несигурност;

- ✓ пресъздаване на знание, включително и на такова, което демонстрира здрав разум;
- ✓ планиране;
- ✓ обучение;
- ✓ общуване на естествени езици;
- ✓ интегриране на всички умения за постигане на различни цели;
- ✓ способността да се чувства и да действа;
- ✓ въображение;
- ✓ автономност;

Тестът на Тюринг е от изключителна важност, който и до днес се възприема като ключов параметър за доказването на наличието на общ, генерален интелект. Неговият създател го предлага през 1950 г. като условие, констатиращо дали компютърът мисли. Критериите, които въвежда са компютърът/ програмата да действа, реагира и взаимодейства като същество, което осъзнато може да изпитва чувства, тоест да притежава „съзнателна чувствителност“ или „чувствителна осъзнатост“ [3]. За да се избегнат преднамереност и предразсъдъци от страна на отсъждащия, Тюринг предлага имитационна игра, в която човек от дистанция задава въпроси за фиксирано време и след това трябва да реши дали отговарящият е човек или компютър, според дадените на разнообразни въпроси отговори. Тестът би бил „успешен”, ако компютърът/ софтуерът бъде възприет за човешко същество.

Освен него, няколко други учени са изработили алтернативни критерии за удостоверяване на генерален, широк изкуствен интелект. Сред тях е „кафе-тестът“ на Стивън Возняк [съоснователят на Епъл, заедно със Стив Джобс], който гласи, че машината би трябвало да извърши самостоятелно всички стъпки, за да направи кафе.

Тоест, да намери помещението, кафе машината, капсулата за кафе и пр.

Известни са много други своеобразни тестове, валидиращи наличността на интелект, подобен на човешкия. Сред тях са: от теоретика по изкуствен интелект и един от създателите на робота „София"- Бен Гьотцел "тестът на робота – студент" (който трябва да е способен да премине всички университетски тестове, подобно на истинските студенти), тестът на Нилсон (за робот, упражняващ равностойно на хората професия), на Тони Северинс за машина (която трябва да разопакова всички части за мебели и да може да ги сглоби, следвайки инструкциите).

Смисълът на всички подобни изпитания се състои в това, че компютърът трябва да може да демонстрира самостоятелно причинно-следствено мислене, анализ, оценка, комбинаторика, интерпретация и импровизация и то при непредвидими и случайни ситуации и обстоятелства. Освен, че ще притежава равностойна на човека интелигентност, типът изкуствен интелект от второ поколение ще има сетивна чувствителност, съзнателност и способност да (се) усеща и придобива опит на базата на емоции и самосъзнание. В този ракурс се очаква и генералният/ общият изкуствен интелект да може да мисли абстрактно (подобно на хората), да прави стратегии и нещо, което вероятно ще е различно от човешкия мозък, който се позова на опит и спомени, да взима решение на базата на безбройни проиграни комбинации от информацията от големите данни (big data). Същото се отнася и по отношение на творческите и иновативни идеи.

Сред най-известните лидери на мнението по тази тема е Рей Курцвейл (ключова фигура в Гугъл и основател на "Singularity University"), който всъщност е един от създателите на тренда за необходимостта и възможността на общия изкуствен интелект. Той предвещава, че през 2029 г. изкуственият интелект ще премине теста на Тюринг и ще постигне човешкото ниво на интелигентност. А през 2045 г., отново според него, ще се случи технологичната сингулярност [singularity] – единение на мозъчната кора с базиран на изкуствен интелект усилвател на човешките способности [Kurzweil, 2005]. Като аргумент за неговите твърдения, служат редица принадлежащи му краткосрочни прогнози, които вече се сбъдват (над 200). Курцвейл е изразител на изключително мощния тренд в САЩ на технологичен абсолютизъм, детерминизъм и оптимизъм, който се опитва да изгради смела утопия за нов световен ред, в който технологиите се

смятат за финално решение във всички човешки области, в които все още има нерешени въпроси и предизвикателства пред човечеството. По този път ще се постигне и концепцията за свръхчовек, като разликата между индивид и машина няма да е видима, а човекът и изкуственият интелект ще се слеят във физически и ментален аспект.

Фактът, че Курцвейл е технологичен оптимист силно повлиява на неговите теории и прогнози и в немалка степен ги прави еднопланови, субективни и лишени от критична рефлексия, тъй като ни представят само една (желаната) перспектива. Основателят на концепцията за „сингулярността“ пропуска да даде обстойна визия за това как само за няколко десетилетия сегашните системи експоненциално ще се трансформират изцяло в услуга на човека, отхвърляйки настоящата пазарна логика на бърза печалба и поощряване на консуматорството. Курцвейл вярва, че технологиите ще са достъпни поради факта, че ще са масови, но не прави анализ и паралел с историята на ползването на всякакъв тип блага и привилегии (материални и нематериални) досега от обществата. Все пак, неговата концепция остава интересна за един от възможните сценарий на развитие на общия изкуствен интелект, който още не е разработен и чисто експертните оценки дали това изобщо ще се случи, се различават много.

3. Супер/ превъзхождащ човека изкуствен интелект

Ник Бостром в своята книга: „Суперинтелигентност. Пътища, опасности, стратегии“[Бостром, 2014] описва последното според класификациите измерение на изкуствения интелект като: „всеки интелект, които в голяма степен превишава човешкото когнитивно представяне и постижение фактически във всички области и полета”.

Суперинтелектът ще надмине човешката интелигентност по всички параметри- от творчество до проява на мъдрост и решаване на проблеми. Той ще бъде такъв тип интелигентност, която не сме наблюдавали сред най-ярките човешки гении. Но феномен от такъв тип буди тревога сред визионери като Илън Мъск и Стивън Хокинг, които считат, че именно това откритие може да доведе до изчезване на човешката раса.

Свидетели сме на многообразие от мнения по темата. Започвайки със заключението, че подобен исторически момент е недостижим, през мнението, че може да се постигне твърде близо до нашето настояще, до възгледите на Курцвейл, че

суперинтелигентността ще се реализира благодарение на единението между човек и компютърен интелект. Широко разпространеното убеждение е, че появата на този вид интелект непосредствено ще последва след генералния, общ изкуствен интелект. Дейвид Чалмърс [Chalmers, 2010] застъпва подобна теза и възгледа, че мозъкът е механична система, която може да бъде наподобена от синтетични материали на принципа на еволюиращите алгоритми, симулирайки естествената биологична еволюция.

4. Рискове

Още с навлизането на първичния тесен изкуствен интелект сме изправени преди редица рискове от морално-етическо естество.

На първо място можем да определим **заплахата от загуба на човешки суверенитет** чрез делегиране на все повече права на изкуствения интелект. Известно е, че теснияят изкуствен интелект се усъвършенства в посока на това да бъде персонален „личнен асистент“. Такъв продукт в координация със специални HR програми ще може да селектира най-добрата професия за клиента си, ще може да следи емоционалното състояние на индивида, ще подбира подходяща според настроението му компания в социалните мрежи, както и здравословна храна, облекло, музика. Друга група от първичния интелект ще бъдат всякакъв тип роботи или системи за почистване, гледане на болни, грижа за възрастни, деца и пр.

Дилемата в морален аспект е доколко подобно всеобхватно подпомагане ще облекчи личността от рутинните и битови занимания и няма ли да я направи **ленива и социално нечувствителна към множество аспекти** от живота. „Допълнителната свобода“ би могла да подсили творческото начало, но и да „измести битието” в посока към виртуалната или разширена реалност, където ще властват новите принципи и форми на консумеризма.

В този ракурс, по-критично е становището на Ювал Харари, което той развива в главата „Религия на данните” [Harari, 2016: 428-462] от книгата му: „Homo Deus. A brief history of tomorrow”. Той заявява, че в настоящето едни системи се развиват все още линеарно. Сред тях са политика, социалните системи, културата и демокрацията в най-общ план, докато други, чрез технологиите, бележат експоненциално развитие като „био”, „техно”, „нано”, и „информационните” технологии. Именно заради това,

изкуственият интелект, чрез компютрите и роботите, според него ще се представя много по-добре във всички области.

Израелският изследовател прогнозира, че ще се промени целият обществен строй, като всички системи ще се реорганизират или някои от тях напълно ще изчезнат. Капитализмът ще бъде заменен от т.нар. от автора „даннизъм" – от "datism". Тази теория подчертава, че за разлика от трите предходни индустриални революции, днес технологиите не са просто инструмент и средство, чрез което подобряваме функционалността си, а променят духа и философията на времето. Хората доброволно и неосъзнато отдават своя суверенитет, без да осъзнават дългосрочните последици. За Харари в недалечна перспектива „неизменените", „неповишените" и „неподобрените" хора ще бъдат напълно безполезни. Той твърди още, че хората са застрашени да изгубят своята икономическа стойност и не само, защото „интелигентност" и „съзнание" се разделят в смисъла на взаимна обособеност, а тъй като такава ще бъде новата парадигма .

Отказът от дейности, засягащи социална и персонална активност може да доведе до загуба на емпатия, десоциализация и дехуманизация. В миналото медиите застават зад публикации с авторитета на своето името. Дори тези, които конюнктурно обслужват властта са известни и читателите, чрез тях, по-скоро проследяват идеологическата и пропагандна линия. Сега, а още повече в бъдеще, все по трудно ще се търси и намира истината. Благодарение на големите бази данни и множеството налична лична информация, всички новини и медийно съдържание ще бъдат персонализирани. Тоест, всеки от нас ще получава различен тип информация въз основа на своето поведение в мрежата и социалните медии. Така ни се разкрива едно от противоречията, които се зараждат сега и ще продължат и в бъдеще-наличие на все повече информация и все по-стеснен кръгозор на потребителите, които са „затворени" в „ехо-стаите си" [4]. „Ехо стаите" са феномен, при който определени идеи, вярвания и данни, се подсилват чрез повторяемост в затворени системи, които не позволяват движение на алтернативни или конкуриращи идеи да се появят. Чрез тях успешно се моделират внушения и убеждения. Поради важността да се генерира трафик в Интернет чрез мрежата, която стимулира своите потребители, ще ги „кара" да чуват и виждат това, което персонално им носи най-голямо удоволствие и удовлетворение

Размиване и прехвърляне на лична отговорност

Политика, която не може да предлага идеи и визионерство, в момента, в който изостане като сложен инструментариум за предоставяне на решения, става нарастващо неатрактивна за гражданите. Властта и отговорността се размиват, като властите предимно администрират, а не създават стратегии. Харари, твърди, че дори демократичните избори ще станат излишни [Harari, 2016: 394 – 395], защото Гугъл ще може да представлява личните политическия предпочитания по-добре от самия индивид. Бъдещето приложение за политически вот няма да забравя и да страда от късогледство, пренебрегвайки например намаляването на данъците или неспазените предизборни обещания.

Малцина знаят, че на последните президентски избори в Русия през март 2018 г. има кандидат с името „Алиса“, който получава 25 хил. гласа. „Алиса“[5] не е реална личност, а изкуствен интелект, създаден от Яндекс [Yandex], руският еквивалент на Гугъл. Предприемачът Роман Зарипов, който стой зад проекта, обосновава защо за президент на Русия не е необходим човек, а изкуствен интелект. В сайта, е публикуван политически манифест с шест точки, две, от които са:

1) Взимане на решения. Роботът се опира на логиката. Не се ръководи от емоции, не търси лична изгода и не дава оценки;
2) Реагира бързо. Изкуственият интелект обработва информацията много по-бързо от човешкия мозък. Мигновено анализира данните и прави изводи на тяхна основа.

Закърняване на когнитивни способности

Съобразно вече достигнатите нива на тесния изкуствен интелект Герд Леонард в книгата си: „Технологията срещу човечеството: идващият сблъсък между човека и машината“ прогнозира, че за децата няма повече да е необходимо да се учат как да пишат, защото компютрите ще слушат и записват и това, което им се казва. Няма да има нужда да се обучават да свирят на инструмент, тъй като взаимодействието между мозък и компютър ще може да произвежда музика, само чрез мисленето за нея. Отново, няма да има нужда и да се учат езици, защото ще има автоматичен превод от и на всеки желан език. Не на последно място, дори няма да е необходимо да общуваме с реално

хора, защото евентуално бихме имали достъп до информационен масив за тях [Leonard, 2017: 57 – 64].

Можем да станем „експерти" за кратко време чрез селектиране на най-представителната информация по даден въпрос. Нуждаем се само от правилния подход и правилната програма. Повече ще боравим с богатите данни, отколкото да намираме и запомняме както до сега. Но няма да е човешко да се търси краткия път във всяка една дейност, защото така се изисква самите ние да се превърнем в машини или поне частично. Даниел Канеман [Kahneman, 2011] неколкократно подчертава, че съзнанието е въплътено в (човешко) тяло. Според него ние мислим с цялото тяло, не само с мозъка. Човешкото за него е холистичен опит, чието изучаване е чрез взаимнозависми фактори. Ако премахнем цялата работа, която се изисква за пълноценни дискусии, размисли и емоции какво ще се случи също така и с колективното понятие за хуманност?

Загуба на екзистенциален смисъл

Трудът все по-малко ще бъде основен фактор за осмисляне на човешкия живот, а разпадът на общностните и социални връзки може да придобие драстични размери. Отново футурологът Леонард прогнозира, че ни очакват по-голяма продуктивност и ръст, но по-малко работни места, както и повече на брой техномилионери и потъваща средна класа. Неговите възгледи относно заетостта са описани в главата „Автоматизиращото се общество" [Leonard, 2016: 47 – 65].

Сега, по негово мнение, се случва трансформация от „информационна икономика" и „икономика на знанието" към „икономика на машинната интелигентност". Той очаква заетостта да спадне, а несъответствието между производителността и средното заплащане да нараства в ущърб на трудещите се. Тоест, новият тренд според него е, че вървим към икономика на намаляване на работните места и растеж без заетост.

Изследователят застъпва и твърдението, че новата тенденция свидетелства за настъпил обрат, а именно, че технологичният прогрес вече не е катализатор на доходи и работни места, както е в индустриалната или ранната Интернет епоха. Ръстът на печалбите се увеличава, но той дава пример с шофьорите на камиони, които трудно биха се преквалифицирали в интерфейс дизайнери. В същото време, неравенството

нараства в глобален план. По негови прогнози около 50% от професиите, ще бъдат автоматизирани следващите десетилетия. Печалбите от предприемачество ще се увеличат многократно, но ще бъдат намалени заетите, като това явление ще се мултиплицира в много индустриални сектори.

Друг голям и неразрешен към днешна дата кръг от въпроси, е кой ще има правото да регулира доембрионалната селекция, подобряването на гените и опцията за „дизайн бебета“? Ще може ли да се тълкува по нов начин концепцията за „родителство“, ако хипотетично ще е възможно да се роди дете с гените на трима, четирима и повече родители? Тези казуси могат ли изобщо да са обект само на национално законодателство или по-скоро това ще породи необходимостта от създаването на една глобална правна система? Как данните от разчитането на генома, което вече е факт, ще бъдат ползвани така, че да не стигне до социална, трудова и здравна дискриминация? В момента тече ускорен процес по учредяване на банки за такава информация и събирането й, тъй като тя е необходима за развитието на изкуствения интелект чрез машинно и дълбоко обучение. Но информацията ще е и безценен източник за работодатели и всякакъв тип здравни и социални осигурителни фондове.

Очертахме само няколко възможни измерения на риска при имплементирането на изкуствен интелект, който е на първично ниво. Но сред очакванията за супер изкуствения интелект са, че той ще може да се препрограмира, чрез което непрекъснато ще подобрява себе си. Тоест, ще притежава функциите на повтарящо се (само)подобряване, което неизбежно ще доведе до интелигентна експлозия. При такъв сценарий, опасенията са, че суперинтелектът ще стане толкова силен и непредсказуем, че трудно би могъл да бъде контролиран или спрян. Затова спешно е необходимо, ако дори дебатът не е закъснял, да се проведе глобален разговор между правителства, институции, научни центрове, неправителствени организации, който да доведе до консенсусни аспекти какъв да бъде етическия кодекс на изкуствения интелект и как той може да бъде постигнат.

БЕЛЕЖКИ

[1] По дефиниция от: https://searchenterpriseai.techtarget.com/definition/narrow-AI-weak-AI

[2] Според: https://www.britannica.com/technology/artificial-intelligence

[3] Преводът на думата „sentient" от английски предполага комплексност между „чувствителност" и „съзнателност".

[4] https://www.techopedia.com/definition/23423/echo-chamber

[5] https://alisa2018.ru/y

ЛИТЕРАТУРА

Bostrom, N. (2014). *Superintellignece: Paths, Dangers, Strategies*. Oxford university press.

Chalmers, David. *(*2010).The Singularity: A Philosophical Analysis. (PDF). *Journal of Consciousness Studies. 17: 7–65.*

Goertzel, B., Penachin, C. (2007). *Artificial general intelligence*. Springer.

Harai, U. (2016). *Hommo deus. A brief history of tomorrow*. Vintage.

Kahneman, D. (2011). *Thinking fast and slow*. MacMillan.

Kurzweil, R. (2005). *The Singularity is near*. New York. The New York Times.

Leonhard, G. (2016). *Technology vs. Humanity. The coming clash between man and machine.* Fast Future Publishing Ltd.

Vikhar, P. (2017). Evolutionary algorithms: A critical review and its future prospects. *Computer Science.* IEEE Xplore, https://ieeexplore.ieee.org/document/7955308/metrics#metrics