Гл. ас. д-р Мариана Тодорова

ИЗКУСТВЕНИЯТ ИНТЕЛЕКТ

КРАТКА ИСТОРИЯ НА РАЗВИТИЕ И ЕТИЧНИ АСПЕКТИ НА ТЕМАТА

София, 2019

© Издателство „Изток-Запад“, 2019

Всички права запазени. Нито една част от тази книга не може да бъде размножавана или предавана по какъвто и да било начин без изричното съгласие на автора и на издателство „Изток-Запад“.

© Мариана Тодорова, автор, 2019
© Деница Трифонова, оформление на корицата, 2019

ISBN 978-619-01-0560-2

Мариана Тодорова

ИЗКУСТВЕНИЯТ ИНТЕЛЕКТ

КРАТКА ИСТОРИЯ НА РАЗВИТИЕ И ЕТИЧНИ АСПЕКТИ НА ТЕМАТА

ИЗТОК ЗАПАД

Съдържание

II. МОРАЛНИ И ЕТИЧЕСКИ АСПЕКТИ, СВЪРЗАНИ С ИЗКУСТВЕНИЯ ИНТЕЛЕКТ / 103

III. ЕТИЧЕСКИ АСПЕКТИ, КАСАЕЩИ „ТЕСНИЯ, ПЪРВИЧЕН ИЗКУСТВЕН ИНТЕЛЕКТ“ / 115

IV. МОРАЛНИ И ЕТИЧЕСКИ АСПЕКТИ: ОБЩ И СУПЕРИЗКУСТВЕН ИНТЕЛЕКТ / 163

ВМЕСТО ЗАКЛЮЧЕНИЕ / 193

ИЗПОЛЗВАНА ЛИТЕРАТУРА / 201

Увод.
Актуалност на изследването

Светът е в очакване на няколко ключови трансформиращи мегатренда на големи, възприети като еволюционни стъпки, които ще променят изцяло концепцията за реалността и нормалността такива, каквито ги познаваме. Всяка радикална промяна досега при човечеството е предизвикана от способстващите за това фактори – от дървото, камъка, бронза, желязото, през парата, електричеството, автоматизацията, до нашето съвремие, което е силно повлияно от наличието на интернет и навлизането на технологиите. Скоро те ще предначертаят и изменят в още по-голяма степен не само политиката, икономиката, търговията, културата и обществата, но и човешката биология и нашата етика.

Изкуственият интелект във всичките му измерения – тесен, общ и суперизкуствен – ще бъде сред най-въздействащите върху живота ни феномени. Неслучайно много изследователи го определят като последното човешко изобретение[1]. Под „последно" може да разбираме както оптимистични, така и песимистични проекции и вложени очаквания. Своеобразният край на човешките изобретения може да се дешифрира в ракурса на прогнозата, че човек и изкуствен интелект от един момент насетне ще сътворяват всичко заедно, допълвайки се взаимно, за да неутрализират всички свои слабости и недостатъци. Но усетът за финал би могъл да означава и че човекът няма да притежава необходимия капацитет да бъде конкурентен на изкуствения интелект и ще предостави доброволно, а може би и принудително територията на научно-развойната дейност изцяло в негово владение. Някои изследователи дори стигат отвъд това твърдение, като заявяват, че всички важни решения за

[1] Barrat, J. Our Final Invention: Artificial Intelligence and the End of the Human Era. St. Martin's press: New York. 2013.

здравето, свободата и живота в цялост ще се генерират от изкуствения интелект.

В България все още липсва систематичен научен труд, който да разгледа множеството концепции за изкуствения интелект и да проследи хронологичното развитие на теориите за него. Отвъд описателния наратив липсва и критичното етично – морално осмисляне на последиците от него във всички или поне в няколко измерения, в които той ще има проявление.

В увода на докторската си дисертация[2] си отбелязвам, че днес съществува и друго своеобразно противоречие, което се откроява чрез изследването на методите за прогнозиране и откриването на техните несъвършенства. Към настоящия момент е все по-трудно да се правят точни и всеобхватни научни прогнози и същевременно с това нуждите от такива са по-големи от когато и да било поради огромната скорост на промените, включително и тези, които ще настъпят с идването на изкуствения интелект, „шока от бъдещето“ и липсата на институции, стратегии, социални технологии, научни общности, занимаващи се специализирано с тази проблематика. Подобна констатация още веднъж подчертава необходимостта от обогатени и отговарящи на усложнената реалност прогнози и последващи ги действия, които не се сливат с наратива за самия изкуствен интелект. Към днешна дата сред част от обществото е налице объркване и смесване на понятията за иновации и технологии, от една страна, със самото научно поле на футурологията и науките за бъдещето, от друга, които всъщност ни предоставят критична рефлексивност и инструменти за тълкуване.

Изходен ъгъл и мотив на труда

В общ план работата е структурирана в няколко части: 1) кратка история на опитите да се създаде изкуствен интелект – проследява се как се развива концепцията за изкуствен интелект в ис-

[2] Контрафактичен анализ и прогнозиране: възможности за съчетаване на контрафактичния анализ със сценарния метод на прогнозиране, защитен през февруари 2013 г.

торически план и какъв е смисълът на понятието в съвременен контекст. Хронологичният израз на разказа дава възможност в най-голяма пълнота да се разберат както етапите на развитие, така и неравномерният почти отсъстващ или пренебрегнат прочит на морално-етичните аспекти; 2) анализира се възможен ли е изобщо генерален и суперизкуствен интелект от гледна точка на теориите за съзнанието (Стивън Пинкър vs. Ювал Харари и др.); 3) проследява няколкото подхода за постигане на изкуствен интелект (символен и невронен); 4) очертава позитивни и негативни сценарии за реализацията на изкуствения интелект; 5) представя визията за видовете изкуствен интелект: а) тесен, първичен; б) генерален, общ и в) суперизкуствен интелект; 6) дискутира моралните и етични измерения, които са разгледани като следствие от въвеждането на тесния, първичен изкуствен интелект – относно бъдещето на труда и професиите, политиката, религията, медиите, образованието, независимостта и процеса на взимане на решения и редица други; 7) описва различни визии за етичните измерения на изкуствения интелект и др. В заключение е направен кратък прогностичен анализ относно моралните и етични последици от действащия вече тесен изкуствен интелект, както и след евентуалното възникване на генералния и суперизкуствен интелект.

Основна цел

Трудът цели да запознае читателите с най-новите дискусии за изкуствения интелект, както и да очертае опасностите и възможностите, които ще се открият чрез него във функционален и морално-етичен аспект. Освен самият дискурс за историята на развитието на изкуствения интелект налице е опит да се обясни защо така мощно се налага тази тенденция, която е на път да се превърне в самоосъществяваща се прогноза. Целта ни е да превърнем темата в достъпна не само за научната общност и професионалната гилдия, занимаваща се с разработката на изкуствения интелект, но и за всеки, който се интересува от тази проблематика и казусите, пред които е вероятно да се изправим един ден в по-близко или по-далечно бъдеще.

Изследователска теза

В текста се открояват няколко ключови тези. Една от тях е, че изкуственият интелект е хилядолетен наратив и проект с променлива наситеност и популярност, но жизеноспособен отвъд границите на XX и XXI век. Идеята за него се среща под различни форми още в античните писания и митове. В късното Средновековие и Ренесанса започват и опитите за материално сътворяване на машини с интелект.

Макар и имплицитно, загатваме, че всеки от подходите да бъде постигнат изкуствен интелект страда от дефицити, които биха могли да се преодолеят, ако се преосмислят техните граничности и се направи стъпка в усилието те взаимно да се допълнят. Друг неизяснен казус според нас е липсата на ясно становище какъв феномен е съзнанието и дали е то необходим компонент за (само)възникването на изкуствения интелект. Тази теоретична празнина, която неминуемо ще получи отражение в самия процес на програмиране и създаване на изкуствен интелект, вероятно ще даде резултати в самия краен продукт. В емергентните теории за съзнанието то се приема като спонтанно възникващо само от факта на съществуване на адекватна физическа база, като например „мозък“ – хардуер с достатъчна и подходяща структура и възможност за негово „действие“ на физическо ниво. В подобни сценарии се очертава хипотезата, че ако изкуственият интелект се реализира във физически измерения и на функционално негово, то „неговото“ съзнание ще възникне автоматично поради наличието на „субстрата“.

Чрез текста правим опит и да онагледим новата ера на технологичен абсолютизъм, вдъхновена и ръководена от една глобална общност, която изповядва изначалното си убеждение, че човечеството е достигнало своя физически и интелектуален лимит в същността, която е днес. Затова то (човечеството) се нуждае от подобряване чрез сингулярност (сливане на естествения и изкуствения интелект и дори „изнасяне“ на човешкия интелект извън тялото като екстракт – чрез милиарди комбинации на машинното обучение или чрез пресъздаването на цялостната мозъчна архитектура) или като субстрат чрез „свалянето“ му и неговия реален или виртуален пренос в други структури.

Според нас този тренд е водещ и толкова силен заради очакването, че изкуственият интелект ще отмени човека във всичко непосилно, обременяващо, монотонно и скучно като дейност и ще го подсили в неговите липси и бариери, за да го превърне в значително по-способна версия на това, което е днес. Тези концепции намират израз и в представянето на човечеството като космическа раса или инженерното изобретяване на нов човешки вид, неограничен от хилядолетно градената си същност и идентичност.

Яростните противници на тези възгледи обрисуват апокалиптична картина, в която главен герой също е изкуственият интелект. Наличните крайни мнения по въпроса свидетелстват за факта, че човечеството комплексно изхожда изначално с предразсъдъци по темата за изкуствения интелект. Наличието им е неизбежно, но то изкривява до крайности възможните стратегии и действия към изкуствения интелект. Очевидно това засега не би могло да се преодолее и ние трябва да сме готови да възприемем разгръщането на изкуствения интелект точно в такъв контекст.

Споделеността на позитивни очаквания сред общността на технологичните детерминисти, сред които е и убеждението, че трите закона за функционирането на роботите на Азимов априори ще действат, обаче създава предпоставките за свръхподсилен и ненаучно мотивиран определен оптимизъм. Макар че исторически те представляват първата формална система на квазиаксиоматична основа, която има претенции да е нормативната система за изкуствен интелект. В този смисъл ние трябва да им отдадем необходимото уважение, тъй като на този етап те не го получават. Подходът на Азимов, макар и незавършен, незадълбочен и недостатъчно успешен и богат от формално-етична гледна точка, е именно същият като на учените, които „градят“ аксиоматични системи. Надеждният проект за етика на изкуствения интелект минава през създаването на богата и добра система от подобен тип. Тя трябва да е формално мощна, максимално консистентна и притежаваща редица други формализирани свойства. В противен случай няма да приляга към машините и няма да предостави необходимата функционалност от гледна точка на хората.

Тоест проблемът не е в самия Азимов, който от пиедестала на собственото си време е твърде адекватен, а в настоящите учени и разработчици, които се задоволяват със заложената от него рамка.

Именно поради изброените причини се опитваме да въобразим и да прогнозираме възможните последици от тесния, общия и суперизкуствения интелект, с което да се информира обществото и да започне подготовката за изграждане на визия и стратегии за посрещане и „колониализиране на бъдещето“ в този ракурс.

Относно работата по етичните аспекти, касаещи развитието на видовете изкуствен интелект, можем да констатираме, че макар вече да има разработки в тази посока, те са недостатъчни и не могат да покрият амбицията за глобален общочовешки консенсус по този въпрос.

Методология на изследването

Следвани са няколко методологически похвата: 1) дескриптивен метод, като е изграден стремежът да се поддържа позицията на изследовател, който описва: съществуващите дискурси, представящи феномена на изкуствения интелект; природата на най-дискутираната проблематика около него и същността на морално-етическите аспекти, базата на което се правят обобщения и изводи; 2) задълбочен анализ на всички съпътстващи дискусията тенденции; 3) сканиране и идентифициране на трендове; 4) описание на пазара на прогнози; 5) изграждане на сценарии. Изследван е голям обем от специализирана литература, като особеното е, че тя покрива напълно различни области – философия, история, логика, икономика, социална психология, математическо моделиране, прогнозиране, като това обособява интердисциплинарния характер на работата. Също така е извършен сравнителен анализ на възгледите на съвременните автори от различните научни дисциплини, като целта е да се намери консенсусен и достъпен подход за представяне на информацията.

Уводни думи

В тази книга ще проследим историческия път на развитие на изкуствения интелект, ще направим прогнози за няколко възможни негови проявления и на тази основа ще разсъждаваме за възмож-

ните етично-морални измерения, от които тази нова и безпрецедентна същност ще бъде съпровождана. В текста се избира именно този подход, тъй като сме привърженици на убеждението, че всяка научна рефлексия – и може би най-вече по поставената проблематика –има нужда да привлече хронологичната подредба на възникването и случването на даден феномен. Подобен изследователски ракурс освен че е в синхрон с каузалното (причинно-следствено) мислене, което удовлетворява чувството за подреденост, позволява да се покажат етапите в еволюцията на едно явление, критическия отклик на обществата и еволюцията или деградацията на морално-етичните норми. Поради това е важно да се обследва всичко случващо се по отношение на ценностите, мотивацията и движещите сили около проектите за създаването на изкуствения интелект. Проявление на какво е той? Дали е еманация на идеята за безкрайния човешки прогрес – или напротив, просто израз на умората от същия този прогрес и желанието да се смени субектът, който отговаря за него и в който той е въплътен? А може би този проект е опит да се сложи край на религиите, да се сътвори нова(та) религия или да се снеме, преобрази, преизобрети и дори замени върховната идея за Бог? Тук може да съзрем обърнатия знак на идеята за сътворението. Човекът се опитва да създаде интелигента система по свой образ и подобие, но той допуска подобието да надмине създателя си нещо, което досега е било невъзможно за човека – да надмине своя Бог. Дали сме свидетели на надменна дързост, или е илюзия за достигане на качество, което дори самият Бог няма – да представи божественото като надминуемо, тоест състезателно, в темпорален контекст? Вероятно е и опит да творим отново „по образ и подобие", за да сме самите ние божествени. Или напротив – това може да е еманация на висша форма на отрицание на автентичната човешка природа и желание да се трансформира и да трансцедира в нова преизобретена същност, което може да е израз на евгениката[3] на XXI век. Споменатите вече технологични оптимисти привиждат този процес като еволюция на Homo Sapiens, който може да про-

[3] Wilson. Ph. Eugenics. Genetics in Encyclopedia Britanica, https://www.britannica.com/science/eugenics-genetics

дължи своята адаптация към новите реалности само технологично променяйки се. В този смисъл може да не го разглеждаме само като евгеника, а като стандартна еволюция, където технологията е оптималната еволюционна стратегия за адаптация и оцеляване.

Течението на трансхуманизма може да бъде пречупено през призмата на позитивната евгеника – за пределното подобряване на човешкия род. Въпросът е дали и в съвременните си измерения този подход е отново йерархичен (достъпен за отбран привилегирован елит), или следва все по-пропагандираната визия за хоризонтални структури на децентрализация (известни с термините на английски „holacracy“).[4]

[4] www.holacracy.org

I. Кратка история

1. Кратка история на опитите да се създаде изкуствен интелект

Още в древността чрез митологията за пръв път се полага образът на надарени с разум човекоподобни машини, които са сътворени от боговете, за да им служат както на тях, така и на хората. В своята книга „Машините, които мислят“[5] Памела Маккордък разделя древногръцката и староеврейската сюжетна нишка за „мислещите машини“, които според нея битуват и днес. Маккордък смята, че именно в старинните древногръцки текстове се въвежда за първи път идеята за мислещите машини, докато староеврейските писания осъждат тези идеи като зли, дръзки, дори богохулнически.

Като първи аналогов компютър в древността някои учени възприемат механизма от гръцкия остров Антикитера[6]. Открити са три бронзови плоски парчета, останали под вода близо 2000 години, които наподобяват модерни машинни части – колела с триъгълни зъбци, кръгли като часовник и разграфени по градуси. Хипотезата е, че устройството е възпроизвеждало небесните движения на Слънцето, Луната и планетите с изключителна прецизност.

В елинската митология древногръцкият бог Хефест[7], покровител на огъня, ковачеството, занаятчийството и метала, изобретява златни слугини и триноди; златни и сребърни кучета пазители и бронзови бикове за цар Ет от Колхида, бронзов великан за цар Минос от Крит, който е жива бронзова статуя, чието задължение

---

[5] http://www.pamelamc.com/html/machines_who_think.html
https://cambridgelatintutor.wordpress.com/2016/03/04/did-homer-invent-robots/

[6] https://bg.wikipedia.org/wiki/Механизъм_от_Антикитира

[7] http://www.greek-gods.info/greek-gods/hephaestus/myths/hephaestus-robots/

е да пази острова. Виждаме, че още в древността тези способности са приписвани на Бога изобретател – този, който твори блага директно за хората. Тоест в творенето на машините е вложено самото божествено. Дали още в тези древни текстове са заложени основните принципи на европейската цивилизация за напредък и юдейския скептицизъм към дръзналите да предизвикат върховния Бог можем да гадаем или да отдадем по-скоро на посткултурална свръхинтерпретация.

В изследването си „Talking heads" („Говорещите глави")[8] авторът Дейвид Линдзи (David Lindsay) проследява появата на т.нар. говорещи глави (псевдоговорещи машини). През XIII век се счита, че Алберт Магнус (Велики) изобретява говорещи машини, които са унищожени от неговия ученик – Тома Аквински. Подобно изобретение се приписва и на английския монах и учен Роджър Бейкън. Дори Мигел де Сервантес описва в едноименното си произведение срещата на Дон Кихот с говореща глава.[9]

Интересно е да се анализира защо през епохата на Средновековието точно говорещите глави са били толкова популярни и е толкова видима наличната целеустременост да бъдат създадени. Може би тогава интелигентността в духа на времето се обвързва изцяло с главата, а тялото се е отрича като нещо изначално греховно и ненужно.

Линдзей отбелязва разцвета на подобни творения и през XVIII век, давайки пример с изобретатели като Фридрих фон Кнаус, Кратценщайн, барон Волфганг Ритер фон Кемпелен, Джоузеф Фабер и др. Макар тази сюжетна линия да не е така добре развита както в древногръцките митове, в юдаизма също е известно съществото Голем, сътворено от нежива материя, с вдъхнат живот от човека, за да работи.

Отвъд митовете и фолклора през 1642 г. Блез Паскал изобретява прототип на първата изчислителна машина[10], която Лайбниц подобрява[11] (Step Reckoner) през 1673 г. През този век (XVII в.)

---

[8] https://www.inventionandtech.com/content/talking-head-1

[9] https://cervantes.thefreelibrary.com/Don-Quixote/118–1

[10] https://www.britannica.com/technology/Pascaline

[11] http://www.gwleibniz.com/calculator/calculator.html

мотивът за сътворената машина официално навлиза в литературния жанр с произведението на Томас Хобс – „Левиатан“[12]. Не по-малка популярност придобива творбата на Мери Шели „Франкенщайн, или новият Прометей“[13], написана през 1823 г. В синхрон с идеите на Ренесанса този тъп механизми все по-често се появяват като цял корпус с тяло. Можем да говорим за насочен фокус именно към него. Вероятно точно този ракурс намира естествено продължение в литературата за роботи и създаването на реални прототипи.

В древността в колективното творчество, а по късно и при древногръцките творци тази фабула намира трайно място. Тенденцията продължава през Средновековието, Ренесанса и до днес в произведения от научната фантастика и антиутопиите. Чрез тези похвати творците проектират потребностите на човечеството да създава и вдъхва живот в негова служба и подчинение, подобно на боговете. Въпросът е кое налага смяната на оптиката – от подчинение към стремеж за сливане и проекция на недостижимото за човека превъзходство? Кога започва залезът на вярата в собствените сили?

В този смисъл създаването на конкуриращ или превъзхождащ ни изкуствен интелект може да се окаже нова съвременна версия на тази изконна и вътрешно присъща за човека проекция на своята същност в по-висши, създадени от него същества, които просто ни предизвикват да разгърнем потенциала си, включително и чрез тяхното изобретяване. Известно ни е, че в алхимията създаването на „хомункулус“ (изкуствено създаден малък човек) е определено като най-висше достижение.

Разбира се, съществуват много основани на доказателствата научни пробиви, които канализират мисленето в това направление, но и самите очаквания подсилват и провокират мощни тенденции и научни търсения в тази област. Интелигентните машини, роботите по същество нагледно представят концепцията за самоосъществяващо се пророчество. Трудно бихме могли да кажем дали писателите научни фантасти по необясним начин отгатват технологиите на бъдещето, или просто въплъщават своите фантазии в своите тек-

---

[12] Хобс, Т. Левиатан. Лист. С. 2017.

[13] Шели, М. Франкенщайн. Труд. С. 2012.

стове и така предзадават модели за следване и изпълнение. Затова и четвъртата индустриална революция е мислена и обоснована през такава призма – като технологична, която фундаментално ще промени начина ни на живот, работа и взаимоотношения в социума. Прогнозите са, че по мащаб, обхват и комплексност тя няма да наподобява нито един период, преживяван до момента. Дали тя ще доведе до дехуманизация на общностите, отношенията, производството и професиите ще стане ясно от т.нар. последващ адаптационен отговор[14] (понятие, въведено от Алвин Тофлър по повод 3-тата индустриална революция) на човечеството. Ако адаптационният отклик на човечеството е интегриращ и отчитащ цялата сложност, разбиращ и управляващ процесите, ангажиращ всички страни в процеса на глобално равнище, а не само на Западния свят, както от публичния и частния сектор, а така също и от академичните и граждански среди, то той би бил успешен. Той ще е необходим, за да се обхване цялата палитра от многообразието на наличните към бъдещия момент проблеми.

През XVIII и XIX век тенденцията за сътворяване на механични същности, не стихва, а напротив, дори се засилва. Унгарският инженер барон Волфганг фон Кемпелен изобретява фалшива шах машина, наречена „турчина“[15], която съвременниците ѝ възприемат като автомат. Паралелно с това се появяват няколко значими открития, които са в основата на човешкия напредък при машините. Изобретен е жакардовият стан, а в научен план Булевата алгебра, публикувана в книгата на Джордж Бул „Законите на мисленето“ (1854). Съвременните компютри ползват логически схеми и устройства, които извършват различни по вида си логически операции. Те са базирани именно на формалния логически апарат на Булевата алгебра. Малко по-късно Готлоб Фреге поставя начало на модерната формална логика, с което силно повлиява на мислители като Бъртранд Ръсел, Лудвиг Витгенщайн, Рудолф Карнап, Алфред Тарски, Курт Гьодел др. Всички тези открития пряко или косвено спомагат в един по-късен етап развитието на съвременните компю-

[14] Tofler. A. The third wave. Bantam books. 1980.

[15] https://en.chessbase.com/post/von-kempelen-s-che-turk-recreated

търни логически операции. Цялата история на модерната логика, заедно с разработките в математиката, създава основата за възникването на полето на изкуствения интелект, но ние умишлено няма да се фокусираме върху тази перспектива с цел да поставим акцент върху плоскостта на морално-етичните проблеми.

Всъщност още в далечната 1832 г. математикът и изобретател Чарлз Бабейч с помощта, твърди се, на Ейда Байрон (Ada Lovelace) проектира първата програмируема механична изчислителна машина, което на практика е първичен прототип на съвременния автоматичен дигитален компютър[16]. Бабидж възприема проекта си като аналитичен двигател (analytical engine), но така и не го завършва. От забвението си идеята излиза едва през 1931 г., когато са намерени непубликувани бележки на автора. Много по-късно, в периода между 1991–2002 г., е реализирана концепцията в нейния цялостен замисъл[17].

В този аспект е важно да уточним, че категориалната разлика между изчисление, от където произлиза – „computing-computer", и анализ (analysis), което е логическо и езиково действие, което не е редуцируемо до изчисления, според повечето хора няма, но не и според специалистите по информационни технологии.

Миналият век е емблематичен за полагането на основите на съвременното понятие за изкуствен интелект. Бъртранд Ръсел и Алфред Уайтхед публикуват забележителния труд „Принципи на математиката"[18]. Освен че представя теориите за модерната математическа логика, работата съдържа богат набор от философски понятия, като пропозиционална функция, логическа теория и теория на типа, които също допринасят за развитието на програмирането на компютърни езици и системи.[19] Върху тези научни постулати стъпват мислители като Курт Гьодел, Алонсо Чърч и Алфред Тюринг .

---

[16] https://www.britannica.com/biography/Charles-Babbage

[17] http://www.computerhistory.org/babbage/

[18] https://plato.stanford.edu/entries/principia-mathematica/

[19] Reynolds, J. „Types, Abstraction and Parametric Polymorphism", Proceedings of the IFIP 9th World Computer Congress. Paris. 1983. 513–523, https://books.google.bg/books?id=ZMpsCQAAQBAJ&pg=PA58&lpg=PA58&dq=1983,+%E2%80%9CTypes,+Abstraction+and+Parametric

В този период се разиграва и полемиката между Хилберт и Гьодел. Идеята на Хилберт е да формализира математика и теория на числата в частност, в система от аксиоми, чрез които всички истинни теореми да могат да бъдат доказани. Но Гьодел оспорва това.

Той описва разликата между теорията, която е синтактична единица, и нейните модели, които са семантични единици. В теоремата си за пълнотата (completeness theorem) той демонстрира, че нещо може да бъде доказано в първична теория тогава и само тогава, ако е истинно във всеки неин модел. В своята теория за непълнотата (incompleteness theorem) Гьодел постулира, че няма единствен модел за която и да е теория на числата. Освен стандартни има и нестандартни модели за числата.

В началото на XX век, през 1912 г., е изобретен магнитен механичен уред за игра на шах[20], който често се бърка с първата компютърна игра. Само осем години по-късно чешкият писател и драматург Карл Чапек публикува пиесата антиутопия „Р.У.Р“ (Росумски универсални роботи) – (Rossum's Universal Robots), в която за първи път се използва думата „робот“ (от чешки – „robota“, „робство, принудителна работа“), навлязла в своя оригинал във всички езици по света. Генеалогията на думата предполага извършване на тежък, обременяващ труд. Има известна ирония в това, че към днешна дата роботите не просто ще отменят, но ще заменят хората в техните професии, което изисква създаването на нова концепция за труда и доходите, придобивани от него.

През 30-те години на XX век Алън Тюринг предлага на вниманието на научната общност универсалната Тюрингова машина

+Polymorphism,%E2%80%9D+Proceedings+of+the+IFIP+9th+World+Computer+Congress,+Paris,+513%E2%80%93523.&source=bl&ots=uBLP21_hiu&sig=FVZaSvP_f8jHaGkgNrSHx5ilc94&hl=bg&sa=X&ved=2ahUKEwiKh6qPloffAhXFECwKHYaWBnYQ6AEwAnoECAUQAQ#v=onepage&q=1983%2C%20%E2%80%9CTypes%2C%20Abstraction%20and%20Parametric%20Polymorphism%2C%E2%80%9D%20Proceedings%20of%20the%20IFIP%209th%20World%20Computer%20Congress%2C%20Paris%2C%20513%E2%80%93523.&f=false

[20] https://en.chessbase.com/post/torres-y-quevedo-s-rook-endgame-automaton

(1936–1937), която всъщност е абстракция на първото определение на компютърен алгоритъм. Тя е и идеализиран математически модел на всеки универсален компютър.[21]

Тестът на Тюринг е друг параметър от изключителна важност, който и до днес служи като ключ за доказването на съществуването на изкуствен интелект. Неговият създател го предлага през 1950 г. като условие, констатиращо дали компютърът (може да) мисли. Критериите, които въвежда, са компютърът/програмата да действа, реагира и взаимодейства като същество, което осъзнато да може да изпитва чувства, тоест да притежава съзнателна чувствителност или чувствителна осъзнатост.[22] За да се избегнат преднамереност и предразсъдъци от страна на отсъждащия, Тюринг предлага имитационна игра, в която мъж и жена се намират в една стая и които си разменят бележки с водещ, намиращ се в друго помещение, натоварен със задачата да определи кой от двамата е жената. Мъжът трябва всячески да се стреми да обърка човека в ролята на разпознаващ, а жената съответно да му помогне. Създателят на текста се пита дали би могло мъжът да се замени с машина?

Тоест очаква се да ни бъде демонстрирана такава интелигентност, която да познава изначално мъжката и женската природа в такива детайли, че да „накара" и убеди някого да повярва, че срещу него е представител именно на женския пол. Почти никой от критиците на Тюринг не обръща внимание на този важен нюанс, че ние бихме идентифицирали някого лесно като човек, ако можем да идентифицираме една от двете полови идентичности. Днес тестът се редуцира по-скоро до алгоритъм, в който от дистанция се задават въпроси за фиксирано време и след това питащият трябва да реши дали отговарящият/отговарящата е човек, или компютър според дадените отговори на разнообразни въпроси. Тестът би бил „успешен", ако компютърът/софтуерът бъде възприет за човек същество[23]. Към днешна дата няма данни за успешно преминаване на теста.

---

[21] https://www.cl.cam.ac.uk/projects/raspberrypi/tutorials/turing-machine/one.html

[22] Преводът на думата „sentient" от английски предполага комплексност на превода.

[23] https://www.britannica.com/technology/Turing-test

През 1981 г. американският философ Джон Сърл предлага аргумента „китайска стая“[24], който е остро възражение на това, че Тюринговият тест е способен да покрие целите си. Неговата обосновка е кратко аргументирана и ясна. Той обрисува хипотетичната ситуация, в която в заключена стая се намира човек, който не знае китайски, но разполага с инструкции/компютърна програма как да дава подходящи отговори на въпроси, зададени на езика. Той би заблудил питащия, симулирайки знание, без да го владее, ако борави добре с напътствията. Същата логика Сърл съотнася към валидността на теста на Тюринг. Според много хора от индустрията на изкуствения интелект тестът вече е преминат и никой не се отнася към него на сериозно заради слабостите му. По дефиниция програмата трябва да може да заблуди човек в над 30% от разговорите, за да го издържи. Хипотетично дори един чат робот от поддръжката на ресторант може да издържи този тест. Всичко е въпрос на дефиницията на „преминаване“ на изпита, без да се излиза от тесния домейн на знания.

Но въпреки убедително му опровержение и в настоящия момент тестът все още се възприема като благонадежден инструмент за проверка на нивото на изкуствения интелект, и то не само в научно-популярните четива.

Всъщност и двата теста онагледяват липсата на „основното звено“ при изкуствения интелект – фактът, че компютърът обработва информация, пресява, прави заключения, но по същество не разбира естествените езици, а засега само математическите.

Но тук може и да предположим, че не разбирането стои в основата на теста, а изграждането на такъв алгоритъм, който максимално да уплътни всякакви възможни реакции, така че хипотетично да е възможно чрез метода на „китайската стая“ да се предскаже наличие на човекоподобен изкуствен интелект.

Ричард Докинс в книгата си „Слепият часовникар“[25] („The Blind Watchmaker“) извежда няколко тези. Продължавайки идеите на Дарвин в нов ракурс, той твърди, че естественият подбор е сляп,

[24] https://plato.stanford.edu/entries/chinese-room/

[25] Dawkins. R. The Blind Watchmaker. Norton & Company, Inc. 1986.

несъзнателен, автоматичен процес, който е и обяснение на живота. Той (естественият подбор) няма никаква цел и ако му припишем ролята на часовникар в природата, то той е „сляп часовникар". Вероятно и самото заглавие на книгата е хвърлена ръкавица към теолога от XVIII век Уилям Пали[26], който използва аргумента за съществуването на часовника като доказателство за наличието на Бог, който според него е великият часовникар на природата. Пали загатва, че дизайнът предполага, че трябва да има дизайнер, а Докинс отрича това, като твърди, че еволюцията е случайна и зависи от конкретните условия. Самият той още през далечната 1986 г. прави експеримент, като залага в компютърна програма параметрите на съществуващата среда и наблюдава как тя генерира двуизмерни образи, много близки до истинските насекоми, прилепи, птици. Точно това негово откритие, както и споменатите емергентни теории, могат да бъдат отправна точка в убеждението, че изкуственият интелект и в частност някакъв тип негово съзнание могат да възникнат от само себе си на еволюционен принцип (както Докинс го тълкува).

Основите на роботиката също се полагат през XX век. През 1926 г. електрическата корпорация „Уестингхаус" създава „Телевокс" – първия робот, използван за полезна работа. През следващите години компанията създава няколко нови модела, сред които е хуманоидния „Електро", ползван за демонстрации заедно с неговото механично куче – „Спарко". Няколко години по-късно Уорън Маккълък, учен в областта на невронауките, и Артър Питс, логик, публикуват „Логическо смятане на идеите, вътрешно присъщи за нервната активност" („A Logical Calculus of the Ideas Immanent in Nervous Activity"), с което по същество започва работа по полагането на теориите за невронните мрежи (е модел за обработка на информация, вдъхновен от изучаването на биоелектричните мрежи в мозъка на човека и животните, образувани от неврони и техните синапси)[27].

В статията си двамата автори се опитват да обяснят как мозъкът произвежда високо комплексни модели, използвайки

---

[26] Paley, W. Natural Theology: or, Evidences of the Existence and Attributes of the Deity; Collected from the Appearances of Nature. R. Faulder, London, John Morgan, Philadelphia. 1802.

[27] https://bg.wikipedia.org/wiki/Изкуствена_невронна_мрежа

базисните мозъчни клетки – неврони[28]. През същата година – 1943 г. – Артуро Розенблут, Норбърт Винер и Джулиън Бигелоу (Arturo Rosenblueth, Norbert Wiener, Julian Bigelow) въвеждат понятието „кибернетика", а по-късно (1948) един от тримата автори – Винер – издава и едноименна си книга.

Отново през 1943 г. логикът Емил Пост публикува свое изследване, в което обосновава тезата, че базирани на правилата експертни системи, известни и като производствени системи, са най-простите форми на изкуствен интелект (rules based systems – also known as production systems or expert systems)[29]. Тези системи са от изключителна важност, тъй като са първият и може би най-значим опит да се постигне изкуствен интелект, като изчисленията се възприемат само като основа за възникване на базирано на правилата мислене и логическо разсъждение (rule based thinking and logic reasoning). Само 2 години по-късно Джордж Поя (George Polya)[30] публикува бестселъра си за евристичното мислене „Как да го решиш" („How to solve it"). Той въвежда термина „евристика" и с работата си повлиява на много учени в областта на изкуствения интелект. В духа на тези тенденции Ванибар Буш публикува студията „Както можем да си помислим" („As we may think"[31]), в която лансира тезите, че компютрите ще бъдат основни помощници на хората във всички техни ключови дейности. Опитите за развитие и съчетаване на невронауките с роботиката продължават да набират скорост. Невропсихологът и кибернетик Уилям Уолтър изобретява роботи костенурки, като залага на хипотезата си, че множеството връзки между малък брой от умни (мозъчни) клетки може да доведе до възникване на комплексен разум и разумно поведение,[32] подобно на идеите на Докинс за комплексност и ред.

---

[28] http://www.mind.ilstu.edu/curriculum/modOverview.php?modGUI=212

[29] Post, E. Formal Reductions of the General Combinatorial Decision Problem, in *American Journal of Mathematics* Issue **65** (2). 1943. pp. 197–215.

[30] Polya, G. How to Solve It. A New Aspect of Mathematical Method. 1945.

[31] https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

[32] Walter, G. The Pioneer of Real Artificial Life, Holland, Owen E. Proceedings of the 5th International Workshop on Artificial Life, Christopher Langton Editor, MIT Press, Cambridge, 1997, pp. 34–44.

Още в зората на опитите да се върви по пътя на създаването на изкуствен интелект наблюдаваме усилието да се пресъздаде еволюционистки подход, като едновременно с това започва да се прави връзката с невронауките. Тази тенденция продължава и днес както със стремежа да се пресъздаде цялостната структура на мозъка, така и с невронните мрежи.

Клод Шанън посвещава свой труд на програмирането на компютър, който да може да играе шах. Във въведението на статията си „Програмиране на компютър да играе шах" („Programming a Computer for Playing Chess"[33]) Шанън аргументира, че това би могло да доведе до нови, по-комплексни функционалности на компютрите като: 1) машини за проектиране на филтри, изравнители, др.; 2) машини за проектиране на релета и превключващи вериги; 3) машини, които ще обслужват разпределянето на телефонни разговори според индивидуалните/отделните обстоятелства, а не по фиксирани модели; 4) машини, изпълняващи символични (нечислени) математически операции; 5) машини, които могат да превеждат от един на друг език; 6) машини, които да предлагат стратегически решения при опростени военни операции; 7) машини, които могат да оркестрират мелодия; 8) машини, способни на логическа дедукция (Shanon: 1).

Към настоящия момент шест от тези осем функционалности вече са възможни благодарение на феномените на машинното и дълбокото обучение (machine and deep learning). Тоест в определени направления ние вече сме свидетели на експоненциално развитие и постигнати резултати в рамките на 50–70 години от първоначално заложената цел. В същата година (1950) Айзък Азимов[34] публикува трите закона на роботиката:

1. „Роботът не може да навреди на човешко същество или чрез бездействие да причини вреда на човешко същество."

[33] http://archive.computerhistory.org/projects/chess/related_materials/text/2–0%20and%202–1.Programming_a_computer_for_playing_chess.shannon/2–0%20and%202–1.Programming_a_computer_for_playing_chess.shannon.062303002.pdf

[34] Asimov, I. I, Robot. Gnome press. 1950.

2. „Роботът трябва да се подчинява на заповедите, получени от човешки същества, освен когато тези заповеди влизат в противоречие с Първия закон."

3. „Роботът трябва да защитава съществуването си, освен когато това влиза в противоречие с Първия и с Втория закон."

Днес те имат стойността на библия за технологичните детерминисти, абсолютисти и оптимисти. Макар да звучат утопично и да са оспорими в много пунктове, законите остават базисно възприятие и очакване за правилата на действие на изкуствения интелект, които дори са придобили параметрите на евристична склонност (bias). Тоест те се приемат за доказани и неоспорими като презумпция, но всъщност никога не са реално прилагани и тествани. Те са много важни в исторически и дори технологичен ракурс, защото са първият опит за създаване на формална нормативна система за машините, колкото и несъвършена и рудиментарна да е тя. Всеки закон сам по себе си е нормативен и така те се опитват да вкарат една хуманна етика към човека като висше благо в поведението на същите тези машини.

В законите на Азимов може да уловим и няколко проблематични постановки. Те правилно изхождат от разбирането, че животът е най-висшата ценност. Но зад тази саморазбираща се аксиома се крият доста нюанси. За да се запази животът на множество невинни хора, ще бъде ли допустимо роботът/роботите да отнемат живота на тези, които го застрашават, като терористи, сепаратисти и пр.? Или как гледаме на избора на самопожертването, което може да се извърши заедно с роботи или от роботи, за да се постигне някаква по-голяма цел, като запазването на духа (обективно състояние на мирозданието) в Хегеловия смисъл.[35]

Другият въпрос, който вече съществува и дори правно е закрепен в някои законодателства (Южна Корея), се отнася до делегирането на права на роботите също като човешките. Така се предполага, че „животът" на самите роботи също ще има априори ценност, което хипотетично може да предизвика питането – кой живот е с

[35] Хегел, Г. В. Фр. Феноменология на духа. Изток-Запад. С. 2011.

по-висша ценност, с което автоматично ще отпадне правилото за подчинение и служене.

Днес някои от роботите са далеч по-прости, а други ще бъдат значително по-сложни от тези, които Азимов описва в книгите си. Законите са наречени така от създателя си, но те са обикновени предписания, пред които трябва да бъдат поставени и критерии за комплексността на функциите. Нанороботите не биха имали каквато и да е съзнателност, но притежават потенциал да навредят на човека, а бойните роботи не би трябвало изобщо да бъдат създавани.

Разбира се, трябва да изходим и от презумпцията, че ако бойните роботи са принципно въобразими, очевидно и „лоши агенти", и държави могат и ще ги въобразят, което е на една крачка от създаването им. „Добрият агент" трябва да стори това, преди да се е случило, защото иначе с глупостта си ще допусне победата на противниците и така сам ще се превърне в пасивен, но основен „сътрудник" за лоши събития. Затова негово етично задължение е да предвиди този сценарий и сам да създаде преди другите системи защита от бойни роботи, които иронично, но неизбежно ще трябва самите те да бъдат бойни роботи, защото в противен случай няма да могат да изпълнят задачата.

2. Каква е съвременната история на изкуствения интелект (започнала от XX век насетне)?

През 1956 г. Джон Маккарти измисля и прокарва официално в езикова употреба понятието „изкуствен интелект", предлагайки това да бъде основна тема на лятната школа на университета Дартмут в Хановер.[36] Като теми за обсъждане той поставя: 1) автоматични компютри; 2) как може да бъде програмиран компютър, който да използва език; 3) невронни мрежи; 4) теория за размера на изчисленията; 5) самоподобряване; 6) абстракции; 7) случайност и творчество (McCarthy, Minsky, Rochester, Shannon). Същата

[36] http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

година трима други учени (Нюъл, Саймън и Шоу) пишат първата (макар че има лек спор дали преди тях не е „шах играта“) т.нар. програма за изкуствен интелект, която трябва да подражава на граматическите способности на хората и да докаже част от теоремите на Ръсел и Уайтхед от техния труд „Принципите на математиката“. В този времеви отрязък можем да разпознаем началото на опитите за семантичния подход към изкуствения интелект.

Единият от екипа автори, Хърбърт Саймън, печели Нобелова награда за икономика през 1978 г. В „Ранд корпорейшън“, чийто консултант е, той среща Нюъл и Шоу (щатни учени там) и заедно решават да конструират машина, която да може да мисли. Между тяхната група и тази на Мински, Рочестър и Шанън се поражда видима надпревара, която проличава на конференцията в Дартмут, където екипът на „Ранд“ е посрещнат хладно и резервирано от организаторите.

Въпреки това Саймън и партньорите му очертават полето на евристичното програмиране, а програмата им „Логически теоретик“ успява да докаже първите 38 от общо 52 теореми в глава втора на „Принципи на математиката“ („Principia Mathematica“). Самият Бъртранд Ръсел високо оценява работата и е задоволен от работата на машината[37]. Добрите резултати насърчават екипа и скоро те изобретяват „General Problem Solver“ – програма за разрешаване на всички проблеми, които биха могли да бъдат изобразени в набор от добре постулирани формули от типа на клаузите на Хорн[38]. Програмата успява да даде решение на казуси като „Кулата на Ханой“ – математическа игра тип пъзел, но не е в състояние да разреши какъвто и да е реален проблем, тъй като търсенето се губи в комплексността и в множествата на възможни комбинации [39]. Все пак чрез това достижение се построява връзката към по-сложни програми за когнитивни умения на компютрите. Съществуват и други примери за опити за универсални доказателства на теореми, но няма да акцентираме върху тях.

---

[37] https://history-computer.com/ModernComputer/Software/LogicTheorist.html

[38] https://en.wikipedia.org/wiki/Horn_clause

[39] https://www.instructionaldesign.org/theories/general-problem-solver/

През 50-те се появяват езиците за програмиране (Fortran, Lisp), а малко след това и семантични гнезда за автоматични машинни преводи, разработени за първи път от Маргарет Мастърман[40]. Обособява се цяло поколение от докторанти и млади учени в Масачусетския технологичен институт, които създават редица нови програми и въвеждането на интерактивните графики. Под редакторство на Едуард Файгенбаум и Джулиън Фелдман излиза сборник от 20 статии „Компютри и мислене" („Computers and Thought")[41] на емблематичните за времето си учени, които дефинират и развиват полето на изкуствения интелект. Отново през този период се появява „Елиза", програма, симулираща разговор между компютър и фалшив психиатър. Тогава се появяват и редица програми, базирани на знанието (knowledge-based program) в химията, математиката и др.

През 70-те години на XX век са реализирани усложнените програми „SCHOLAR", „ARCH", „SHRDLU" за разбиране и разчитане на езици, както и първите експертни и планиращи програми. Тенденциите от следващите години, включително до нашето настояще, бележат бурно развитие в областта, като всяко предходно изобретение подпомага идващите разработки. Преди около 50 години е създадена ARPAnet, първата функционираща мрежа, предшественик на интернет. Разработена е от DARPA, правителствена агенция към Министерство на отбраната на САЩ. Както вече споменахме, през 1978 г. Хърбърт Саймън печели Нобелова награда за теорията за ограничената рационалност (bounded rationality), която въвежда „идеята, че при взимането на решения индивидите са ограничени в няколко аспекта – притежавана информация, когнитивните възможности на тяхното мислене и времето за взимане на решение". Тя е предложена като алтернативна основа за математическото моделиране на взимането на решения[42]. Саймън измисля и понятието „satisficing", комбинация от „satisfy" и „suffice" („задоволявам" и „сти-

[40] http://www.mt-archive.info/Aslib-1988-Wilks.pdf

[41] https://dl.acm.org/citation.cfm?id=601134

[42] https://en.wikipedia.org/wiki/Bounded_rationality
https://www.behavioraleconomics.com/resources/mini-encyclopedia-of-be/bounded-rationality/

гам“ [в смисъла на достатъчност]). Ученият употребява понятието си в случаите, когато „хората търсят решения или правят избори или съждения, които са достатъчно добри за целите им, но биха могли да се оптимизират“ (English: 2016). Същото понятие се превръща в крайъгълен камък за ефективността при изкуствения интелект.

През 1979 г. е създадена „MYCIN“ – програма, която на базата на заложената в нея интелигентност доставя данни според степента, до която може да бъде програмирано интелигентно поведение.

Свързващата машина (connection machine) е масивна паралелна конструкция от компютри, която бележи началото на суперкомпютъра, произлизаща от докторската дисертация на Дани Хилис от Масачусетския технологичен институт. Първоначално намерението му е било изобретението да се използва за приложение на изкуствения интелект и обработване на символите, но по-късните версии стават част от изчислителни машини[43].

Десет години по-късно (1989) е разработен ALVINN (An Autonomous Land Vehicle in a Neural Network) [44], компютър, който е способен да командва кола. Това е прототип на съвременните проекти на автономните, самоуправляващи се автомобили.

Базисните разработки за автономните коли са направени преди 30 години, но едва сега (2019) имаме увереност, че тази технология е на прага на пълното си осъществяване. По тази логика може адекватно да се запитаме дали експоненциалното развитие действително винаги е толкова „експоненциално“, или просто хората имат къса памет за предисторията на дадено изобретение. По този начин косвено се засягат и научните етични норми, защото желанието за реализацията на свръхинтерес, съответно на голяма печалба води до изкривено представяне на продукта – като резултат на само най-нови модерни технологии, без да се отдаде значимото на фундаменталнта наука и поредицата от предшестващи научни постижения. С подобни актове също се подсилва и спираловидното

[43] https://www.sciencedirect.com/science/article/pii/016727898490263X

[44] https://papers.nips.cc/paper/95-alvinn-an-autonomous-land-vehicle-in-a-neural-network.pdf