ИЗКУСТВЕНИЯТ ИНТЕЛЕКТ – ОТ ПЕРЦЕПТРОН ДО CHATGPT

Ваня Лазарова

Доцент, доктор, катедра „Информационни технологии и комуникации“ УНСС,
e-mail: vlazarova@unwe.bg

Резюме

В тази разработка, ще се опитаме да проследим пътя на изкуствения интелект от неговата поява, до най-новите му проявления. Ще разгледаме най-значимите литературни източници, свързани с науката за изкуствения интелект. Ще се постараем да опишем основните етапи, по който се е стигнало до съвременното състояние на изкуствения интелект. Също така освен чисто информационни въпроси, ще засегнем и някои въпроси от други научни области, тъй като развитието на изкуствения интелект е свързано с много области от човешкото знание освен информационните системи и технологии - философията, когнитивните науки, статистиката. В разработката ще се фокусираме върху някои ключови за развитието на изкуствения интелект теории и експерименти.

Ключови думи: *изкуствен интелект, перцептрон, генеративни модели*

ARTIFICIAL INTELLIGENCE - FROM PERCEPTRON TO CHATGPT

Vanya Lazarova

Abstract

In this paper, we will try to trace the path of artificial intelligence from its emergence to its latest manifestations. We will look at the most significant literary sources related to the science of artificial intelligence. We will try to outline the main stages by which the modern state of artificial intelligence has been reached. Also, in addition to purely informational, we will touch on some philosophical questions, since the development of artificial intelligence is related to philosophy. In the paper, we will focus on some key theories and experiments for the development of artificial intelligence.

Key words: artificial intelligence, Perceptron, generative AI models

JEL: O3, Z0

Изкуствен интелект (ИИ) и връзка с човешки интелект

Счита се, че за първи път понятието Изкуствен Интелект (ИИ) е употребено през 1956 г. от Джон Маккарти на семинар, проведен в Дортмут колеж в САЩ, където се събрали учени интересуващи се от използването на компютри за наподобяване на човешкия интелект.

Използването на понятието ИИ е в два аспекта - за означаване на способността на компютърните системи да извършват дейности, подобни на човешкия интелект; за означаване на научно-приложната област за изследване и изграждане на системи с ИИ. В този доклад ще се спрем на използването на ИИ при изграждане на системи.

През 2004 г. Джон Маккарти (McCarthy, 2004), дефинира понятието така: „*Това е наука и инженерни технологии за създаване на интелигентни машини, и по-специално интелигентни компютърни програми. Това е свързано със задачата за използване на компютри за наподобяване на човешкия интелект, но ИИ не трябва да се ограничава само с методи, които са биологично наблюдаеми.*"

Дефиницията показва, че начинът изкуственият интелект да се опише, е чрез човешкия интелект, макар да не трябва да се разчита само на биологичните характеристики. Тъй като интелектът отразява само умствените възможности на даден човек, обикновено се използва

понятието интелигентност, което е по-широко, има множество проявления и включва възможността да се чувства, да се разбират взаимовръзки между събития, да се вземат решения.

Компютрите придобиват огромни възможности - изчислителна мощ, милиони операции за секунда, технически характеристики, които ги направят по-бързи от човека. Но дали са по-интелигентни? Маккарти задава въпроса: „Какво е интелигентност?“, защото без отговора на този въпрос, дефиницията не е пълна. Остава неясно коя компютърна програма да се счита за интелигентна.

Интелигентността е понятие, което традиционно се свързва с човека и неговите способности да мисли и да се учи от опита, да решава проблеми и взема решения, да се адаптира към околната среда и да притежава разбиране и контрол върху собствените мисловни процеси. Човешката интелигентност се проверява чрез различни тестове, който формират коефициент за интелигентност (IQ), измерител, предложен още през 1912 г. от Уйлям Стърн (Стърн, 1912).

Компютрите, съвкупност от взаимообвързани хардуер и софтуер, работят, като програмите (софтуера) дават инструкции на хардуера как да реагира на въведената информация. Тези инструкции могат да бъдат от прости, до много комплексни – да решават задачи, да дават указания за движение, да съхраняват, обработват и визуализират данни за стоки и услуги, да предсказват въз основа на изчисления върху данни и пр.

Какво ще ни накара да опишем една компютърна програма като интелигентна?

За първи път отговор на този въпрос се опитал да даде Алън Тюринг, през 1950 г., в публикацията си „Компютърни машини и интелект“ (Turing, 1950). Той предложил, вместо да се дава отговор на въпроса дали една компютърна програма е интелигентна, да се подходи по друг начин, чрез тест. Идеята му се състои това: наблюдател човек, да прецени от разговор, чрез размяна на текст, дали срещу себе си има друг човек или машина, без да ги вижда. Компютърната програма ще издържи „теста на Тюринг“, ако разпитващият не е способен да я различи от човек. Тюринг е считал, че ако поне 30% от тестваните не разпознаят, че това е програма, тя минава теста.

Днес вече сме убедени във възможностите на ИИ, но както всяка наука, и тази преминава през върхови моменти на развитие и моменти на спад. Има един дълъг период през 70-те до средата на 80-те години, когато в тази област са правени малко изследвания, в сравнение с останалите сфери на науката.

През 1957 г. Франк Розенблат прави фундаментално изследване и предлага "Перцептрон: възприемащ и разпознаващ автомат" (F., 1957) като клас изкуствени невронни мрежи, имитиращи аспекти на мозъка и рецепторите на биологичните системи. На база на това изследване през 1967 той изгражда компютър Mark 1 Perceptron базиран на невронна мрежа, която се „обучава“ на принципа проба - грешка.

През 1969 Марвин Мински и Сиймор Пейпърт публикуват книга, озаглавена „Перцептрони“, която е забележителната разработка върху невронните мрежи. Тази разработка за известно време, е главният аргумент срещу бъдещи проекти за изследване на невронни мрежи. (Minsky, 1969) Мински и Пейпърт атакуват Перцептрона на Розенбалт, като посочват, че неговата двуслойна структура, не може да обхване всички асоциации в една невронна мрежа.

Единствената област от ИИ, която се развива успешно през 70-те и началото на 80-те години са експертните системи. Те са предназначени да решават проблеми в някоя специфична област от човешкото познание. Днес те се наричат от някои учени „предишна ера на ИИ“, именно защото техния бум е по времето, когато беше замряло развитието на другите области на ИИ.

Експертните системи са изградени на основа на факти, извлечени от знанията на експерти от конкретна област, от които чрез логически правила се достига до ново знание. Типичен пример за такава система е MYCIN - експертна система за диагностициране на бактериални болести. MYCIN е разработена в началото на 70-те години на миналия век в Станфордския университет от Едуард Шортлиф (Shortliffe & B.G., 1975). Програмата е била разработвана в продължение на 5 години. Било необходимо да се натрупат и съхранят като факти, знанията на експертите в областта на бактериалните болести. Въз основа на фактите, чрез съчетаване на логически правила, се достига до заключение. Ето един прост пример от тази експертна система:

Ако Микроорганизмът е грам-отрицателен,

Морфологията на организма е пръчица и
Организмът е аеробен.
То Данните сочат, че организмът е бактероид.

Развитието на експертните системи днес не е спряло, макар че сега те са изградени съвсем на други принципи.

През 1987 година Дейвид Румелхарт и Джеймс Макклеланд публикуват фундаменталния си труд „Паралелно разпределена обработка“ ((Rumelhart & McClellan, 1987), с който дават отново силен тласък на развитието на ИИ. Защо все пак хората са по-умни от компютрите, въпреки че отделни компютърни програми са способни да надминат човешкото изпълнение на дадена задача? Отговорът, според тези учени, се крие в мощната паралелна архитектура на човешкия ум. Те описват тази архитектура, чрез теория на познанието, наречена конекционизъм. Теорията предполага, че умът е съставен от голям брой елементарни единици, свързани в невронна мрежа. Психичните процеси са взаимодействия между тези единици, които се подбуждат и възпират една друга в паралелни, а не в последователни операции. В този контекст знанието вече не може да се разглежда като съхранявано в локализирани структури; вместо това се състои от връзки между двойки единици, които са разпределени в цялата мрежа. На тази теория основно се дължи напредъка в последно време на мрежите за дълбоко обучение (DL Deep Learning), и на големите езикови модели (Large Language Models LLM).

През 1995 г. Ръсел и Норвиг публикуват учебник (Russell and Norvig, 1995), наречен „Изкуствен интелект: модерен подход“. Тази книга става настолна за много поколения студенти, а през 2020 година, претърпява 4-то издание. Според наблюденията на университета в Бъркли (1547 Schools Worldwide That Have Adopted ИИМА (berkeley.edu) от този учебник учат повече от 1500 университета в света. В България, според този сайт, от учебника се ползват общо 5 университета.

Джон Сърл, професор във Философския факултет в Бъркли и дългогодишен изследовател на ИИ, създава понятието „силен изкуствени интелект (СИИ)” (Strong Artificial Intelligence SAI). Машина със силен изкуствен интелект ще има психически състояния като тези на хората, например, може да разбере разказани истории както човек ги разбира. Засега силният ИИ съществува само в теоретичен аспект - компютър има интелект, равен или надминаващ човешкия. Такъв ИИ ще има самосъзнание; ще може да предвижда; да се самообучава и да взема самостоятелно решения. Професор Сърл, създава мисловен експеримент, наречен „Китайска стая“, за да демонстрира невъзможността, според него, за конструиране на СИИ.

Същността на мисловния експеримент се състои в следната постановка. В една стая стои човек, който не знае китайски език и през една дупка в стената, комуникира писмено с китаец на китайски. Китаецът праща през дупката въпросите на човека, а той чете предварително написани инструкции на английски, в които са посочени отговорите на всички възможни въпроси на китаеца. Когато получи йероглифите, търси в инструкцията измежду всички възможни съчетания, кое да избере за отговор. Напомняме, че експериментът е мисловен, теоретичен. Един външен наблюдател ще заключи, че човекът в стаята знае китайски. Според така проведения експеримент, китаецът може никога да не разбере, че събеседникът му не знае китайски. А пък човекът вътре в стаята може никога да не проумее за какво е бил разговорът.

Ако в стаята вместо човек, имаше компютър, в който предварително са въведени инструкции за отговори и реакции на всички възможни въпроси и ситуации, то този компютър ще изглежда като човек, но в никакъв случай не може да се отъждестви с човек.

Есперимeнтът „Китайска стая“ ще илюстрираме като използваме изкуствен интелект (Фиг. 1). Компанията Майкрософт работи в насоката за създаване и обучение на приложение на ИИ, което генерира изображения, въз основа на текст Image Creator from Microsoft Bing. Текстът, който задаваме на ИИ да илюстрира е: „John Searle's science experiment called the chinese room with Chinese instructions and an outside observer“ Тази илюстрация е най-близо до представата за мисловния експеримент на проф. Сърл.

Нарисувал: Image Creator from Microsoft Bing https://www.bing.com/images/create/john-searle27s-science-experiment-called-the-chines/64bb8536c57746d1b544a94166da8440?id=3%2b6bnLNh1ksOiVU1MqnfQw%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP

Фиг. 1. Експеримент „Китайска стая“.

Този експеримент, наречен от някои учени „най-значимия мисловен експеримент на 20 век“ не е оборен теоретически, има много противници, които смятат, че е възможно да се създаде свръхинтелект. Основните критики се отнасят до това, че експериментът е само мисловен, не може да бъде проведен експериментално. На което професор Сърл отговаря, че технически невъзможно не означава, че е логично невъзможно (Searle, 1980).

Генеративен изкуствен интелект (Generative Artificial Intelligence GAI)

Изкуственият интелект е научна област, която съчетава компютърни науки и набори от данни, за решаване на конкретни проблеми. ИИ се обучава - това е процес на избор за ниво верни резултати и постепенното достигане до това ниво. Един от най-новите клонове в дълбокото обучение е разработването на специфичен вид невронна мрежа, наречена Трансформър. Това е невронна мрежа, която е проектирана да генерира последователности от свързани елементи от данни (като изречение например). За първи път Трансформърът е бил предложен през 2017 г. в статията „Вниманието е всичко, от което се нуждаете“ (Vaswani, et al., 2017), резултат от усилията на редица учени. От 2017 г. до създаването на приложението ChatGPT времето е само 5-6 години.

Генеративен изкуствен интелект е наричан този изкуствен интелект, който е в състояние да генерира съдържание, което преди това не е съществувало в пространството – текст, изображения, книги, филми, дори части от научни трудове.

Стъпвайки на съществуващи технологии като Трансформър и големите лингвистични модели (LLM), моделите, използващи GAI са обучавани върху огромни обеми от данни (например цялата История на България и например всички речи и произведения на Стефан Стамболов) как да предскажат следваща дума в дадено изречение. Анализирайки хиляди съчетания от думи, могат да предположат, че след фразата „Не щеме злато, искаме…“, най-вероятното продължение е „свобода и човешки правдини“. (Следвайки известното стихотворение на Стамболов „Не щеме ний богатство,/ не щеме ний пари,/ а искаме свобода, човешки правдини!“ заедно с фразата „щеме“, която очевидно се свързва с някакъв специфичен изказ.). GAI може да създаде ново стихотворение в стила на Стамболов, което не съществува до този момент, само въз основа на подаден на вход първи ред, Примерите могат да бъдат върху различни данни, но главната идея на генеративния изкуствен интелект е, че може даден модел

да се „научи“ да генерира статистически вероятни резултати, когато крайният потребител му даде задача.

Генеративните модели се използват от години в статистиката за анализ на числени данни. Възходът на дълбокото обучение обаче направи възможно разширяването им извън числата; до текст, изображения, реч и други сложни типове данни. Бъдещето е в тези модели, които се обучават върху широк набор от данни, които могат да се използват за различни задачи, с минимална фина настройка. Системите, които изпълняват конкретни задачи в един конкретен домейн, отстъпват място на генеративен ИИ, който учи по-общо и работи в различни домейни и върху различни проблеми. Базовите модели, обучени на големи масиви от данни и после фино настроени за набор от приложения, движат тази промяна.

Най-популярните примери на такъв тип приложения с GAI днес са ChatGPT, Bing AI, Bard и др. Силата на тези приложения е в това, че в резултат на анализа на огромни обеми от текстови данни и изображения, могат да съставят смислен отговор на който и да е въпрос, да генерират съдържание, което преди това не е съществувало.

Същността на GAI моделите, както посочихме, но нека пак да повторим, защото е много важно е, че те са вероятностни модели, генерират статистически най-вероятното съчетание на изрази, което следва от поставената им задача, но не и непременно 100% вярното.

На българския телевизионен екран в момента може да се гледа една популярна игра „Попитахме 100 човека…?“, в която играещите се опитват да отгатнат какво са отговорили тези 100 човека на въпроса, а не кой е правилния отговор. Постановката на тази игра много прилича на постановката, при която човек задава въпрос на някое от приложенията с генеративен изкуствен интелект. GAI ще даде най-вероятния отговор, който е извлечен от анализ на коментарите, блоговете, уикитата, чатовете и пр., но не задължително 100% верния отговор. Много вероятно е също така, при липса на достатъчно данни по даден въпрос, генеративния изкуствен интелект да си „измисли“ изрази, факти, т.е. да генерира информация, която изглежда най-правдоподобна като отговор на даден въпрос. Това явление има име - „халюцинации на GAI“. Компаниите, които разработват генеративни модели, винаги много ясно декларират отказ от отговорност, като предупреждават многократно ползващите моделите да не се доверяват изцяло на тях. Днес много хора не са в състояние да разберат, че срещу тях няма мислеща като човек машина, нито свръхинтелект. Засега.

Реални опасности от развитието на ИИ

Сред учените се води неразрешен и до днес диспут „Може ли да се създаде свръхинтелект, който да чувства като човек?“. Към този въпрос се връщат най-често учените от когнитивните науки, за които е важно дали може да се направи ИИ, който да наподобява човешкия.

Дали въобще някога ИИ ще може да изпитва истински емоции, без да ги симулира? Това е въпрос, който засяга науката за човешката психика и мозък. Засега не е ясно. Има много неща, които не знаем за възможностите на ИИ. Учените изследват развитието на ИИ и предупреждават, че в бъдеще може да се стигне до състояние на „сингулярност“ (Kurzweil, Ray, 2005), точка от която ИИ, може да се развива самостоятелно, без участието на човек и няма да може да се върне обратно.

Но дори и сегашното развитие на ИИ е такова, че доведе до спорове, дискусии относно вредите, които може да нанесе на хората и дали тези вреди са съпоставими с ползите от него.

По-горе в текста споменахме за „халюцинациите“ на генеративния изкуствен интелект – състояние, при което програмата генерира измислици и неверни изречения и твърдения. Това състояние не е умишлено, вече подчертахме, че засега AI само симулира емоции и съпричастност, но това води до заблуда много хора, което е вредно и особено опасно. Проблемът е, че софтуерът с изкуствен интелект може да заблуди всеки - от високо образовани специалисти до хора с лабилна психика, готови да посегнат на живота си.

Много популярен стана случаят от юли 2022 г., със софтуерния инженер на Гугъл, който твърдял, че чатботът, с който работи е личност и има съзнание. Сензацията толкова се разширила, че било необходимо да се намесят психолозите на компанията, да проведат и

разпространят диалог с чатбота, за да покажат, че това е просто програма, която имитира човек – чатботът казал че има семейство, деца и пр.

Защо е толкова трудно да се разпознаят „халюцинациите“ на AI?

Проблемът идва най-вече от ограниченото ни познание по даден проблем. Ако проблемът е илюстриран от AI, ние лесно ще видим грешките и несъответствията и ще се разберем, че дадена картинка е направена от AI. Увеличите многократно Фиг. 1, по-нагоре от този текст, и ще видите едно гротескно изображение на човек - несъразмерно око, недовършени ръце, изобщо не особено умел опит на генеративния изкуствен интелект да направи илюстрация на Джон Сърл и то такава, че да не повтаря никоя негова снимка. Все пак изображението трябва да е уникално. Но вече системите за генериране на изображения дотолкова се усъвършенстваха, че наистина могат да създадат изображение, което да е поне толкова достоверно, колкото би го рисувал художник.

Как да разпознаем дали даден текст съответства на истината, при положение, че нямаме достатъчно познания по дискутирания въпрос. Това е трудно. Например, ученици трябва да напишат есе за Че Гевара или Васил Левски... Имат някакви познания по въпроса, но не дотолкова, че да преценят дали е вярно всичко, което им е написал чатбота.

Ето един реален пример… Google разработва свой чатбот, наречен Bard, който е в експериментален етап. При зададен въпрос: „Как е загинал Васил Левски“, чатботът отговаря с няколко смислени изречения, че „Левски е един от най-великите български революционери и национални герои“. Заедно с това „халюцинира“ версията, че е „обесен на Витоша, като при обесването му е използван механизмът на бялата лястовица“… Как се е намесил разказът на Йовков в този въпрос, никой не може да отговори. Но пък написаното звучи толкова достоверно, че заблуждава дори хора, които знаят от училище българската история. Ако този текст, беше попаднал при чужденец, който нищо не знае за Левски, но все пак е чувал нещо по първата част от текста, че това е наш национален герой, ще повярва и на останалата част. Текстът можеше да попадне в ученическо есе на китайски, немски, японски и пр. и тази небивалица бързо да се разпространи… В даден момент можеше да съществува в толкова версии в интернет, че да бъде приета за много вероятна от следващия чатбот, който анализира данните по този въпрос.

Заключение

Само за около 70-80 години, ИИ претърпя огромно развитие. Времето от научните открития, до тяхното прилагане в практиката става все по-кратко.

Проблемите свързани с потенциалния риск от създаването на изкуствен суперинтелект най-често се неглижират от разработчиците на ИИ. Причината е, че създаването на системи с общ изкуствен интелект (Artificial General Intelligence), е централна стратегическа цел на водещите компании в областта на ИИ, като Microsoft, OpenAI, Meta, Google, Amason, IBM. Да се приеме сериозно риска от създаването на изкуствен суперинтелект би означавало да се забави в определена степен създаването на системи с общ ИИ чрез въвеждане на допълнителни правила, контролни функции и други ограничения.

ИИ трябва да служи на хората, за подобряване на техните качества. Като преподаватели във висши училища, сме длъжни да запознаваме студентите с опасностите от безконтролното развитие на изкуствения интелект. Чрез съвместни разработки между преподаватели и студенти (Hristov Georgi, Markova M. (2022); Kovacheva M. (2020) трябва да насочваме тяхната креативност към полезните приложения на ИИ, към използването има главно за облекчаване и повишаване качеството на човешкия труд.

Хората от своя страна трябва да знаят, че срещу тях засега стоят обучени машини. Може би в близко бъдеще това ще се промени, но във всички случаи, човекът трябва да запази контрол върху машините.

Последната дума относно жизнено важни решения, свързани с ИИ, трябва да се взема от хората.

References

1. F., R. (1957). The Perceptron: A Perceiving and Recognizing Automaton.
2. McCarthy, J. (2004). WHAT IS ARTIFICIAL INTELLIGENCE? (Stanford University). Stanford: Stanford University.
3. Minsky, M., & Papert, S. (1969). Perceptrons: An Introduction to Computational Geometry. Cambridge: MIT Press.
4. Rumelhart, D. E., & McClelland, J. L. (1987). Parallel Distributed Processing. Explorations in the Microstructure of Cognition: Foundations. MIT Press.
5. Russell, S. J., & Norvig, P. (1995). Artificial Intelligence: A Modern Approach. Prentice Hall.
6. Searle, J. R. (1980). Minds, brИИns, and programs. Behavioral and BrИИn Sciences(https://web-archive.southampton.ac.uk/cogprints.org/7150/1/10.1.1.83.5248.pdf).
7. Shortliffe, E., & B.G., B. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences. 23 (3–4), pp. 351–379.
8. Turing, A. M. (1950). Computing Machinery and Intelligence.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention Is All You Need. Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
10. Wiki. (1966). Wikipedia. Извлечено от ELIZA: https://en.wikipedia.org/wiki/ELIZA
11. Стърн, У. (1912). Die psychologischen Methoden der Intelligenzprüfung: und deren Anwendung an Schulkindern [Психологическите методи за тестване на интелигентността]. От W. Stern, Монографии на образователната психология, бр. 13. Лайпциг.
12. IBM. What is artificial intelligence? https://www.ibm.com/topics/artificial-intelligence
13. Stanford University. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/chinese-room/#toc
14. Kurzweil, Ray (2005) The singularity is near: When humans transcend biology.
15. Kovacheva M. (2020) Storing Big Data in NoSQL databases compared to SQL -advantages and problems. In: 10th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education – 2020, Sofia.
16. Hristov Georgi, Markova M. (2022) Reporting of R&D Costs Under IAS 38 – The Case of Biopharmaceutical Companies. 4/2022, Научни трудове на УНСС.