https://doi.org/10.53656/nat2023-2.04

СРАВНИТЕЛЕН АНАЛИЗ НА МЕНТАЛНИ МОДЕЛИ И ИЗКУСТВЕН ИНТЕЛЕКТ

Пламен Чергаров
Софийски университет „Св. Климент Охридски“ – София

Резюме. Менталните модели са репрезентация на външния свят. Те представляват хипотетични примери на релевантните за познаващия субект елементи, конфигурирани по аналогия със световните същности. Формирането на гарантирани вярвания може да бъде резултат на самата надеждност на този тип репрезентация. Статията аргументира такава позиция и защитава тезата, че светът е епистемично несигурен. Това означава, че няма обективна причина, сама по себе си, дадени елементи да бъдат предпочитани за сметка на други. Това представлява сериозен проблем за изследователите на ИИ, които намират трудност при пресъздаването на способността на човек да се ориентира визуално в света. Първите трудности при създаването на ИИ са били свързани с разпознаването на образи, докато това е естествен механизъм на визуалната система при хората. По тази аналогия менталните модели са с висока стойност в ориентирането ни в епистемичната среда. Тази статия аргументира в полза на по-високата епистемична надеждност на менталните модели, в сравнение с ИИ.

*Ключови думи***:** ментален модел; епистемична несигурност; изкуствен интелект; надеждност

Проблемът за епистемичната несигурност

Живеем в свят, където нашето познание, ръководещо поведението ни, достатъчно често зависи от индуктивни умозаключения. В голямата си част тези заключения вършат чудесна работа за това да се ориентираме и справяме с ежедневните предизвикателства. Те са с ключова роля за технологичния и научен прогрес. Въпреки тези постижения остава една мистерия около индукцията. Как знаем кои са релевантните елементи от света, за да ги включим като предпоставки в умозаключенията си?

За да отговорим на този въпрос, нека разгледаме развитието на изследванията в сферата на изкуствения интелект (ИИ). Логиката е неразделна част от създаването на алгоритмите, с които работят всички видове ИИ, затова и проблемите, пред които се изправя развитието на тази технология, ще бъдат показателни за недостатъците на принципите, около които е изградена.

Историята на ИИ започва с неговите предшественици като базови шахматни програми и изчислителни алгоритми и достига до съвременната па-

радигма на дълбоките самообучаващи се механизми за около 60-те години (Anon 2020). Мощта на обработка на данни се е увеличила, но и структурата на обработването на данни се е променила. От процесуална обработка в ранните дни, днес обработката е на много паралелни нива, които си комуникират. Това позволява разработването на цяла палитра от действащи алгоритми, които са както тясно специализирани, така и по-общи. Например тясно специализиран ИИ са програмите за шах като AlphaGo, а по-широки са модели като Gato, които могат да извършват повече неща, като чат, търсене на образи, игра на „Атари“ и др.

Може спокойно да се каже, че бързото израстване на сектора отразява не само добър инвестиционен модел, но и оптимизъм за това, че можем да пресъздадем най-доброто от човека в машината. Това обаче все още не се случва напълно. Не само че моделите за обща обработка не се справят толкова добре, колкото хората, но и някои тясно специализирани модели се провалят като предвижданията на специалистите далеч не се оправдават (Ismail 2021).

Много компании предвиждаха да започнат продажби на самошофиращи се автомобили през 2020 г. Това, разбира се, не се случи. В тази сфера наистина е налице голям прогрес, но той далеч не е достатъчен, за да имаме такъв тип ИИ на пътя. Оказва се, че шофирането е по-сложна способност, отколкото се е смятало преди, и сегашните ИИ не са готови да заместят човека дори в тази област.

Нещо повече, през 2021 г. по-малко компании се решават да тестват подобен тип коли (Bellan 2022). Въпреки това те са шофирали повече километри, отколкото преди. Преминати са общо над 6 милиона мили, като 40 хиляди от тях са били изцяло без шофьор. Това означава, че по-голямата част от изминатото разстояние е било с шофьор зад волана, който да бъде нещо като осигуряващ механизъм за безопасността на колата. В случай че колата, или по-точно алгоритъмът, вземе грешно решение, човекът вътре е трябвало да се намеси.

Тук идва и интересната статистика. Колко често се намесват шофьорите? Средно шофьорът в колата поправя решенията на ИИ веднъж на 12 553 км. Но това е само в компанията Waymo. От Cruise правят поправка веднъж на 67 000 км. Други компании имат различни показатели, но все още няма единен стандарт, който да е показателен за качеството на взетите решения на ИИ. Очаква се и че тези цифри с времето ще вървят към все по-малко необходими намеси от страна на шофьорите. Това, което можем да кажем, е, че статистическата извадка е нараснала достатъчно, за да можем да имаме някои прозрения за това как се справя алгоритъмът в сравнение с човека.

Основният въпрос е какво определяме като надеждност на пътя и как да я измерим. В общи линии можем да дефинираме надеждността като честота (в случая ни интересува тя да е ниска, т.е. можем да говорим за рядкост) на определен тип инциденти. Колкото по-рядко се случват, толкова по-надеждни

са решенията на колата или човека. Изследванията в тази насока ни изобразяват отрезвяваща картина по отношение на способността да изградим ИИ, който да се съревновава с нас в тази сфера на умение. В най-лошия случай ИИ има надеждност от 99.997%. В сравнение с това хората са неимоверно по-надеждни – около 99.9999999999% (Media, N.H.T.S.A. 2020). Измерването е направено в САЩ, където има един фатален инцидент на 160 милиона изминати километра.

Един шофьор взема над 100 решения за 160 км. При това тези решения трябва да се съобразят с решенията на ИИ. Решенията на най-добрия ИИ, който е създаден по новите технологии за дълбоко обучение, са 5 пъти по-ненадеждни от тези на човешките шофьори. Тук говорим за отворена среда, където няма предварително филтрирани данни. Това означава, че нито ИИ, нито хората са в предварително опозната среда (епистемично сигурна), където да могат да се ориентират по-лесно.

Изводите, които можем да направим, са, че хората са изключително по-надеждни от ИИ към този момент, и да зададем въпроси защо не можем да пресъздадем тази надеждност. Хората са надеждни и в редица други сфери. Например в САЩ стрелбата е популярно хоби. На година се купуват около 10 милиарда амуниции, а смъртните случаи са малко повече от 500 (Anon 2022). Говорим за случайни смъртни случаи, а не за умишлени. Това прави надеждна стойност от ~1:20 000 000 или 99.999995%.

Във военен контекст, има над 1 милиард изстреляни амуниции, а само около 10 смъртни случая, които са от случаен характер. По време на обучението за стрелба смъртта е изключително рядко явление (Ricks 2011). Това е надеждност от ~99.9999999%. Оказва се, че на множество фронтове случайността на инциденти е рядко срещано явление. Повечето хора се справят завидно добре в това област.

През 2016 г. е направено изследване сред експерти в сферата на ИИ за това какви неща ще са способни да вършат тези технологии (Anon 2021). Експертите смятат, че ИИ ще е способен да заема все по-голяма част от нашия живот, но към 2022 г. тези прогнози не оправдават първоначалния ентусиазъм. Очакваното заместване на хирурзите все още не се е случило и все още се налага наблюдение и намесване от страна на хирурга в съвременните робот-асистирани операции. Козметичните операции имат смъртност от 1:300 000 (Rohrich et al. 2020), преди да добавим усложнения като сърцеци. Ако искаме ИИ да постигне същата успеваемост, трябва да имаме надеждност от поне 99.999998%.

Има опасения, че алгоритмите могат системно да дискриминират определени групи. Някои критици посочват тази опасност на основата на механизмите, които се използват при изграждането на ИИ. Други смятат, че хората са по-предразположени към такъв тип поведение, но това просто не отговаря на данните.

Макдоналдс например обслужват 70 милиона клиенти на ден (Media 2018). В САЩ расизмът е деликатна тема и делата, свързани с него, са с голямо медийно отразяване. Много служители са отстранени от работа заради отношение към клиентите, което е на расова основа. Да речем, че само 10% от всички случаи на уволнение заради расистки обиди достигат до новините (Jr., C.R.W. 2019). Това би означавало, че такова езиково поведение има в 1:2 500 000 000 случая, или надеждност от 99.99999996%. Това означава, че служителите на „Макдоналдс" са в значителна степен надеждни по отношение на правилно езиково поведение спрямо клиентите.

От една страна, сме основателно позитивно предразположени към резултатите, които полето на ИИ успява да постигне. Много от способностите на хората са репродуцирани успешно и в някои сфери като шаха дори надминават експертите. Когато става въпрос за сложна среда (епистемично несигурна) – такава, която изисква множество паралелни процеси на обработка за вземане на решение, имаме разминаване между хората и ИИ. Най-голяма такава разлика в самото начало на развитието на ИИ се намира в обработката на изображения.

Само преди десетилетие ИИ се справяше изключително слабо във връзка с разпознаването на образи и тяхното класифициране в подходяща категория (Anon 2021). Предизвикателствата пред разработването на алгоритъм, който да се справя с тази задача, бяха толкова големи, че тя изглеждаше невъзможна. През 2021 г. такъв тип задача за класифициране и подреждане в клас се дава в бакалавърските програми за компютърни науки. Това, че напредваме в тази сфера, е видно и от простата търсачка на „Гугъл", където, когато напишем „котка", излизат безброй изображения, които съответстват на нашето търсене.

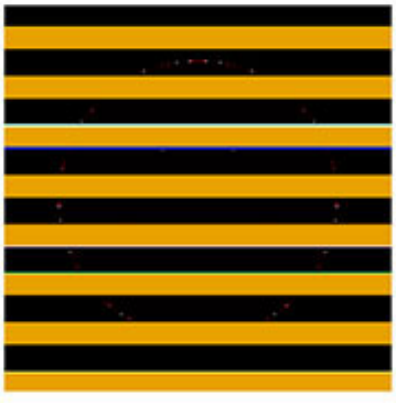
Над 16 000 различни алгоритъма могат да различават изображенията на котки към днешна дата (Markoff 2021). Както видяхме по-горе, въпросът е по-сложен от това дали алгоритмите са способни, или не. Въпросът е доколко се справят по-добре от хората в това начинание. В много случаи отговорът е, че не се справят достатъчно надеждно, за да могат да се съревновават с нашите собствени способности.

Към момента най-големият пробив в сферата на ИИ се дължи на вече споменатите дълбоко обучаващи се програми. Тези програми са така разработени, че да наподобяват структурата на човешкия мозък. Имат йерархична подредба и връзки между различните нива на обработка. Това позволява на програмата да се обучава сама. Колкото повече данни обработва програмата, толкова по-добра става в намирането на пътеки и повтарящи се модели. Разликата между хората и програмата е в способността за обработка на големи данни. Тази разлика е изразена в степента на скорост на обработка и обема на поетите информационни единици. Това обаче прави ли МО (машинното обучение) по-добро в разпознаването на образи?

Към 2011 г. хората са се справяли значително по-добре в разпознаването на образи в сравнение с ИИ (McMillan 2015). Но тази тенденция се променя с времето (Tanz 2017). Към 2014 г. алгоритмите са намалили почти наполовина грешките в разпознаването на образи, които са допускали през 2013 г. Това е прогрес с невероятна скорост и трябва да отчетем, че няма да е учудващо тази тенденция да се запази.

Това обаче не казва цялата история. Фей-Фей Ли във видео за образователната платформа „Тед“ твърди: „Дори най-умните машини са слепи“ (Anon 2022). В случаи, където контекстът или други същностни връзки са важни да се разпознаят, ИИ се проваля. Алгоритмите не могат да разпознаят образ, ако част от него липсва. Например, ако е дадено половин човешко лице, ИИ изпитва затруднения да разбере, че това е човешко лице. Показателен е провалът в поставянето в една категория на слонова глава, бивна и чаена лъжица, когато са изобразени частично.

Алгоритмите се провалят и в разпознаването на прости, поне за хората, изображения на черни и жълти ленти, като ги класифицират като училищен автобус. Фигура 1, показана по-долу, е лесна за хората, но по някаква причина, изглежда, е невъзможно да бъде класифицирана като просто черно-жълти ленти от ИИ. Това е озадачаващо. Как е възможно да се създадат изключително сложни програми, способни на автоматично разпознаване и подреждане на голям клас от обекти точно и надеждно, но провалящи се в наглед прости задачи? Един възможен отговор може да се дължи на това, че и хората нерядко допускат привидно прости грешки и не би трябвало да се изненадваме на това, че и алгоритмите го правят. Може би е просто разлика в полето на възможните грешки, където се разминаваме, и това ще се изчисти с времето. Според мен отговорът е малко по-сложен от това.



Фигура 1

Менталните модели като разлика за ориентирането в света между ИИ и хората

Всичко по-горе изложено е един аргумент, чиято цел не е да подкопае напредъка в ИИ. Напротив, смятам, че е неизбежно тази сфера да се подобрява и в определен момент да задмине нашите собствени способности, понеже влагаме най-доброто от нас в една технология, която привидно не изпитва ограниченията, които имаме. Целта на аргумента е да покаже сложността на нещо приемано за даденост от хората, а именно визуалната обработка на данни, и да даде основа за приемането на следните тези.

1. Светът не е съставен от очевидно значими елементи, на които трябва да обръщаме внимание.

2. Това означава, че той не е епистемично сигурен сам по себе си. С други думи, има безброй истинни вярвания, които можем да имаме и поддържаме във всеки един момент, но това не би означавало нищо, ако те не са подчинени на някаква епистемична цел.

3. Единственият надежден източник на ориентация в света са менталните модели. Те се развиват с времето, а някои от тях са и предварително вградени в когнитивния апарат. Именно те подчиняват истинните вярвания на някаква епистемична цел.

По точка (1) мога да добавя аргумента, който е илюстриран от експеримента с невидимата горила (Drew et al. 2013). В него участниците са помолени да изгледат кратък видеоклип, на който се появяват хора, които си подават баскетболна топка. Задачата на участниците е да преброят подаванията, които хората в клипа правят. В някакъв момент в средата на картината се появява човек, облечен в костюм на горила, и се тупа по гърдите. Около половината от участниците остават слепи за горилата. Насочването на вниманието към едно нещо означава слепота за много голяма част от всичко останало, ако не и за всичко останало.

Какво можем да изведем от тази дълга илюстрация на проблемите при ИИ и експеримента с горилата? Според мен не е възможно да твърдим, че ориентирането в света е проста и очевидна задача. Реалността не предоставя някаква информация, която да е от значение сама по себе си. Информацията е обработена с някаква цел. Но колко са възможните епистемични цели в такъв свят? Безгранично много. Можем да отдадем цял живот на опознаването на една малка стаичка, за да добием цялото познание, на което сме способни за нея. Можем да опишем всеки нюанс на цвят, всяка форма и всяка повърхност, но това върви против интуицията ни за това какво имаме предвид под „знание“, или поне срещу интуицията ни за значимо знание. Но много тривиалности могат също да са „знание“, доколкото са гарантирани истинни вярвания.

Това, което можем да извлечем от проучването в сферата на компютърните алгоритми, е, че е трудно да се намери правилната организация, която да има

този усет за значими елементи в средата. За една програма, която обработва голяма база данни, всичко би могло да бъде значим елемент. Всякаква последователност или модел на поведение сред данните могат да бъдат важни. Това не е недостатък, защото е полезно във виждането на неща, за които бихме останали слепи. Но също така не е и добродетел, понеже голяма част от тези модели могат да се окажат безсмислени.

Елементите, които са значими за нас, са предварително зададени, ако мога да продължа аналогията с програмите. Доколкото дадени елементи от средата са играли някаква роля във формирането на успешна еволюционна стратегия, тези елементи са били приоритизирани и внедрени в нашите модели на оперириране с данни. Т.е. ние нямаме неограничен набор от мисловни модели, с които да обработваме информацията от света.

Дефиниране на ментален модел

Какво е един ментален модел? Преди да дадем ясна дефиниция, е удобно да се позовем на една аналогия с географските карти. Картата е репрезентативна по отношение на територията. Тя съдържа най-различна информация и служи за ориентир. Може да включва в себе си информация за градове, пътища, реки, планини или да репрезентира друга информация, като климат, полезни изкопаеми, население, течения и др. Една и съща територия може да бъде репрезентирана по различни начини, като зависи коя информация намираме за значима. Тоест, целта определя включената информация в картата.

Самата карта служи за нашето ориентиране в територията. Важно разделение е, че картата не е самата територия и няма как да бъде. Тя е мащабирано копие на територията. Това копие може да бъде повече или по-малко точно. По този начин и архитектурният план на една сграда не е самата сграда, а умалено репрезентиращо копие, чиято цел е да ни насочва в строителния процес.

Сам Харис дава интересна метафора за моралните ни интуиции (Harris 2012). Според него има една местност с низини и върхове, които символизират морални отличия или падения. Върховете и низините са с оглед на човешкото благополучие, което означава, че има обективно желано състояние на човешко съществуване, постигано според действията, които предприемаме в живота. Тези действия са морални, когато ни повеждат към някой връх, и неморални, когато ни отвеждат в падина.

По подобен начин менталните модели, които използваме, за да се ориентираме в света, са добри, доколкото са насочени чрез епистемичните цели на познание и разбиране. Подобно на географските карти те вземат предвид само определена информация, която е необходима за належащата цел. Целите на познанието в биологията подкрепят използването на ментални модели като еволюционната теория, а целите на познанието в област като физиката подкрепят използването на модели като струнната теория. Няма ментален модел,

който сам по себе си да е добър. Той е добър в зависимост от епистемичните ограничения на областта, в която се опитваме да постигнем определена епистемична цел.

Както няма излишни географски карти, а те зависят от приложението, така няма и излишни ментални модели. Разбира се, можем да си представим такива, които са абсолютно неприложими, но това не подкопава основния аргумент тук. Ако изберем да направим географска карта, репрезентираща популацията на врабчета, то тя би била чудесен инструмент, ако целта е да бъде представено това знание към момента на съставянето на картата. Въпросът е дали това е достойна за преследване епистемична цел.

Точно тук правим кратък обрат към алгоритмите и ИИ. За ИИ няма значение каква информация се обработва и синтезира. Информацията е просто информация. Няма качествено различна и по-значима такава. Единственото, което е от значение, е намирането на повтарящи се пътеки или постигането на предварително заложената в алгоритъма цел. Това е различно при хората. Повечето от нас не биха намерили за необходимо да репрезентират абсолютно точно популацията на врабчета в дадена местност. Това е заради интуицията, че ако днес са 10 000, то утре може да са повече или по-малко и не би бил от значение точният брой.

Това е и отличителната характеристика между ИИ и менталните модели на човека. Нашите модели са усъвършенствани и подлежащи на една епистемична цел – истината. Но освен истината ние се интересуваме от истини, които правят съществуването ни по-добро. ИИ, от друга страна, няма концепция за истина, която да направи неговото съществуване по-добро, следователно може да трупа безброй истини, без оглед на тяхното приложение. Това е и трудността на изследователите на ИИ, които все още не могат да алгоритмизират вродения механизъм на пресяване на истини у човека.

Визуалната система у нас е имала в най-базовия смисъл един алгоритъм: „Всичко, което е полезно за оцеляването, трябва да се идентифицира бързо и точно“. Кое е това „полезно за оцеляването“? Имали сме милиони години на опити и грешки, за да бъде вградено. ИИ няма нещо, което да е полезно или вредно за него. Затова и все още има трудности в изграждането на алгоритъм за разпознаване на образи. Най-базовата и интуитивна наша способност се оказва далеч по-сложна, отколкото сме очаквали.

Време е да дадем и по-ясна дефиниция на това какво е ментален модел. Менталният модел е репрезентация на света, която е изградена от няколко сетивни модалности. Той е когнитивна операция на обработка на информацията от света и подреждането на тази информация в смислени категории. „Смислени“ означава, че можем да действаме въз основа на тях. По този начин менталните модели стават предварително основание за знание. За да имаме обосновано истинно вярване, то трябва обосноваността да бъде произведе-

на отнякъде. Произвеждащият механизъм задължително е някаква ментална операция (най-малкото е мозъчна операция, но помиряването между двете е извън обхвата на настоящата разработка). Така, обосновката се основава върху някакъв ментален модел. Като под основание имам предвид, че е базирана на някакъв модел, но можем да я различим от него. Както радиовълната зависи от конструкцията, която я произвежда, така и обосновката е изградена върху менталния модел.

Всеки модел се състои от входящи данни, механизъм и изходящи данни. Менталният модел не се отличава в това отношение. Входящите данни са подбрани така, че да отговарят на крайната цел. Механизмът е самата структура на обработка на тези данни. Изходящите данни са резултатът или вярването, което получаваме след цялата операция. Интересно е какво се случва, когато имаме само резултат (вярване) и чрез механизма трябва да изведем причини за него, но това е свързано с конститутивен анализ на менталния модел, а това не е от значение за основния аргумент на тази статия.

Проблеми с дефиницията и въвеждането на термина „ментален модел“

Когато се дава нов термин, необходимо е да се даде и основание защо трябва да бъде предпочитан пред предходните и дали въобще се различава от тях. Навлизането на новото трябва да бъде обосновано. Този критерий е важен, за да не се замъглява дискусията, а това е противопоказно, ако целта е постигането на по-добро разбиране в която и да е област. В такъв случай с какво допринася въвеждането на „ментални модели“ в дискусията? Не е ли просто нов начин да имаме предвид „репрезентация“?

За да отговорим на подобна критика, трябва да разгледаме как възниква един ментален модел и как възниква една репрезентация. След това трябва да разгледаме функционалните разлики. Като резултат от този подход трябва да бъде премахнато смисловото замъгляване.

Има много дебат около теорията на репрезентирането, но тук ще се спрем на базова дефиниция, че репрезентирането е информация, която е обработена чрез сетивата и съхранена в ума (Stich 1992). Обектите от света оставят своя отпечатък в съзнанието, където те са обработени чрез набор от когнитивни операции. Това, което остава в ума, не е самият обект, а негово копие и това копие се нарича „репрезентация“. Най-добрата аналогия, която успява да улови, това е картина на пейзаж. Картината не е пейзажът, но е негова репрезентация.

Дефинирахме менталния модел като репрезентация на света, изграден от няколко модалности. Това означава, че наред с репрезентацията в менталния модел участват множество когнитивни процеси. Менталният модел в някаква степен е изграден от множество репрезентации на света. Бихме могли да кажем,

че той е една голяма репрезентация или генерализация от отделни репрезентации. Трябва обаче да внесем една уговорка, за да уточним някои важни разлики.

Менталният модел не е статичен. Аналогията с картината не е идеална, но можем да я използваме тук. Ако картината е репрезентация на пейзажа, то менталният модел е обобщение от всички пейзажи или правилата на репрезентирането на пейзажи. Това би бил принципът на рисуване на пейзажи, освен самите пейзажи. Понеже принципът се съдържа в самите пейзажи, те също участват в него. Менталният модел на мисловния експеримент включва някакви правила, но и включва запознатост с отделни конкретни мисловни експерименти. Моделът на еволюцията включва отделните видове, но и принципа зад тях (под принцип тук се разбира обяснителният механизъм, който дава смисъл на данните за измененията на видовете. Това са в частност случайните мутации, естественият отбор и сексуалният отбор).

Когато твърдя, че менталният модел не е статичен, това е просто внасяне на светлина в изследването. Самите репрезентации също не са статични, както множество експерименти, свързани с паметта са демонстрирали. Въпрос е на количество и макар и границите на това количество статичност или динамичност да не са толкова ясни, това не пречи за продължаването на аргумента.

Друга съществена разлика е, че менталният модел е проблемен за строга дефиниция. Репрезентацията, от своя страна, както казахме, е сравнително ясно определена. Този проблем попада сред други трудни за дефиниране термини като съзнанието. Мисля, че можем да говорим чрез тези термини и без стриктна дефиниция. Причината за това е свързана с употребата на думата „купчина“. Колко зрънца правят една купчина? Макар и да няма стриктна дефиниция, можем да употребяваме тази дума смислено.

Нека вземем друг пример, който е показателен откъм проблема за дефиниране на понятия. Такова би било здравето. Трудно можем да кажем от медицинска гледна точка, а и от психологична какво е здраве, или поне да дадем тясна дефиниция за здраве, която да се отнася за всички възможни случаи. Не можем да дефинираме здравето просто като отсъствие на болестно състояние, тъй като това носи ред нови проблеми. Въпреки това можем да говорим за здраве и болест достатъчно смислено, че да има практически последствия.

По същия начин, менталните модели избягват строга дефиниция, но можем да твърдим с нарастваща убеденост, че те не са просто репрезентация на света. Можем да ги изследваме като операции на съзнанието, които подреждат света в смислена информация, която дава познавателно предимство пред простите репрезентации. ИИ може да репрезентира света чрез визуално разпознаване. Това как ще подреди репрезентациите, така че да може да взема решения въз основа на тях, как ще реши кои са по-смислени да бъдат включени и кои да бъдат игнорирани и как да се създаде от тях модел на света, е много по-сложна задача.

Така менталните модели преминават в нещо много по-фундаментално, макар и абстрактно, ниво на познанието. Бихме могли да кажем, че сетивността е гарант за истинни убеждения и е основа на знанието. Това обаче ни връща към проблема, който изследването в областта на ИИ ни показа. Не всички истинни убеждения са важни. Не всяко знание е важно. Възможността да се изгубим в проверка на всяко отстояние между буквите, шрифта, редове, брой на думи и изречения, е налична. Всеки може да добива толкова истинни убеждения за тези факти от света, че да му отнеме цялото време, с което разполага. И все пак това не би го направило добър пример за епистемно отличен агент.

Епистемичната стойност на менталните модели

Вече направих и първата етическа оценка, която прокарва границата между онези проблеми, с които ИИ се сблъсква, и проблемите, с които живите организми се сблъскват. Епистемологията не се занимава само със знанието като такова. Загзебски (Zagzebzski 1996) прави разграничение между висше познание, което е артикулируемо и по-низше, което е от рилайабилистки тип. Тук разграничението, което правя, е знание, което има валентни последствия за познаващия, и знание, което няма такива последствия.

Когато ракетата „Чалънджър“ се взривява и е назначен комитет да изследва инцидента, един от членовете изготвя доклад, който завършва със следните думи (Farquhar 2017): „За да имаме успешна технология, реалността трябва да е по-важна от връзките с общността, защото природата не може да бъде заблудена“.

Това са думи на Ричард Файнман, известния физик. Знанието, което разграничих по-горе като „валентни последствия“, е демонстрирано в тези думи. Има знание, което е свързано с връзките с общността. Има разбиране за света, което е отразено в тази сфера. Когато инцидентът се случва, много хора от НАСА и правителството на САЩ работят за имиджа на космическата програма. В този момент тя е застрашена и ако публичното мнение се наклони в определена посока, то е възможно да бъде и закрита. Специалистите по връзки с обществеността правят всичко възможно това да не се случи.

Това няма нищо общо със знанието и разбирането, свързано с природните закони, както Файнман проницателно посочва. Ако искаме технологичната грешка, която е причинила трагедията, да не се повтаря, не можем да маскираме тази грешка само за да запазим общественото одобрение към програмата, или поне така си представям мисловния процес на Файнман. Това, на което следва да се обърне внимание или да се даде значимост на разбиране, е в сферата на инженерните науки.

Тук не е въпрос на полярна валентност. Идеалната полярна валентност би била знание\разбиране със стойност, значима за нашето поведение или желана цел, и знание\разбиране без такава стойност. В случая има сблъсък на две

позиции с различна стойност и тази стойност е съобразена с желаната цел. ПР специалистите имат своя цел, а Файнман – друга цел. Това не означава някакъв познавателен релативизъм или нихилизъм, а тъкмо обратното. Едната от тези цели е по-значима за разбирането на света и нашият научен и технологичен прогрес от другата цел.

Двете позиции имплементират различни ментални модели за ориентиране и обяснение на света. В контекста на ситуацията единият от тях е с по-висока епистемична стойност от другия. Това може да се наблюдава и в научните експерименти. Различни хипотези се борят за обяснението на феномени. Това не означава, че те имат еднаква епистемична стойност.

Нека вземем пример с извършен експеримент от английския биолог Р. Фишер (Salsburg 2002). Самият експеримент е тривиален, но методът се използва и до днес. „Опитването на чай от дамата“ е изследване, което е станало емблематично за слепите и двойно слепите проучвания. На дадено събитие една дама спори, че чаят има различен вкус, ако първо се добави мляко и след това чай, или ако първо се добави чай и след това мляко. Повечето присъстващи били скептични умове и не вярвали да има разлика между двете. Затова и присъстващият Фишер решил да постави под въпрос твърдението.

Това, което направил, било да приготви 6 чаши. Три от тях били приготвени, като млякото било първо, а чаят е добавен после. Три били приготвени, като чаят бил първи, а млякото – второ. Дамата не знаела кои чаши по кой метод са приготвени. Фишер поднесъл чашите в разбъркан порядък и създал формула за оценка на достоверността на твърдението на дамата, че има разлика във вкуса. Трябвало да познае определен брой чаши по кой начин били приготвени, за да има достоверност, по-голяма от чиста случайност. В конкретния случай дамата познала всички чаши по кой начин били приготвени.

Днес използваме този метод, за да оценяваме епистемичната стойност на научните хипотези в множество области, като медицина и физика. Използваме този метод и в политическите прогнози. Ако изследваните лица не знаят дали са взели плацебо, или истинско лекарство и покажат значителна разлика в сравнение с контролната група, то можем да твърдим, че лекарството има практически последствия.

Разбира се, този метод също е вид ментален модел, но от него можем да изведем някои прозрения за същността на епистемичната стойност, която самите модели ни дават.

1. В определен контекст винаги има ментален модел, който ни дава по-голяма епистемична стойност от друг. Не е случайно, че в определени области на познанието се налагат специфични методи на познание вместо други.

2. Епистемичната стойност на менталния модел зависи от желаната цел и последствията за познаващия агент. Голяма част от тези цели са се развили и са продукт на еволюционни процеси, но има и такива, които се основават

на тези процеси. Няма ген за четене, но тази епистемична цел е базирана на някакви гени. Това означава, че развитите ментални модели могат да са специфични за полето, без да имат еволюционно предимство.

3. Въпреки това базирането върху еволюционни механизми трябва да бъде подчертано. Ако не беше основано върху тях, бихме се лутали подобно на ИИ в света. Има вградени структури за оценяване валентността на информацията. Това е основна разлика с алгоритмите. Алгоритмите изискват информацията да бъде предоставяна „наготово“. Когато тренират програмите за визуално разпознаване, не се подават случайни снимки, а подбраните обекти, които програмистите искат да бъдат разпознати. Чак след научаването започва изпробването със случайни обекти.

4. Когато казваме, че има вградени структури за оценяване валентността на информацията, трябва да се уточни, че тези структури играят ролята на филтър. Невъзможността да се поеме неограничен набор от информация, изисква някакво пресяване според различни критерии. Критериите не са обект на настоящото изследване, но можем да посочим кои процеси са от значение за този тип пресяване и играят структурираща роля за информацията. Това са когнитивните структури на паметта, вниманието и възприятието. Благодарение на тях информацията бива подготвена за включване като входящи данни на менталните модели.

5. Не може да има поведение в света, без да има модел за света. Животните, ние и алгоритмите работим чрез операции „ако…, то…“. На база на тези операции изграждаме представа за последствията и законите на реалността. Тази представа е по-малко или повече успешна.

Надеждност на менталните модели

Сега е време да поясним защо менталните модели са надежден източник на вярвания и защо са по-фундаментални и съответно по-значими от тях. Нещо повече, те са основен гарант на вярванията. В използването си менталните модели работят с информация подобно на това как един строителен обект се изгражда от различни процеси и материали. Крайният продукт е цяла сграда, която може да бъде разглеждана като състояща се от отделни сегменти, но за целта на нейното използване се разглежда като единно цяло. По тази аналогия менталният модел може да достигне резултат, който е по-акуратно да се обозначи като вярване, макар и да може да се разглежда и на отделни пластове.

Нека дадем един от най-известните примери за използването на ментален модел и достигането до някакви вярвания. В своите „Размишления върху първата философия“ (Dekart 1978) Рене Декарт употребява един от най-използваните ментални модели във философията, а именно мисловния експеримент. След като поставя под съмнение всичко познавано и възприемано, той достига до следния резултат – вярването, че неговото съществуване не може да бъде поставено под съмнение.

Сократ, когато казвал, че нищо не знае, е достигнал до нещо много по-съществено от самото знание. Достигнал е до метода, който произвежда знанието. Този метод, познат като сократически диалог, използва задаването на множество от въпроси, чрез които да се достигне до фундаменталните основания за вярване в дадена пропозиция. Тези въпроси не са подбрани случайно, а имат преднамерена последователност. Париш (Parrish 2019, pp. 93 – 100) определя тази последователност като 5-те „защо"-та. Целта е, подобно на любопитни деца, да достигнем до същността на изследваната тема. Използвайки този тип целенасочено разпитване, Сократ достигнал до прозрението, че хората не знаели това, което вярвали, че знаят.

Самите вярвания нямат голяма епистемична стойност, ако методът, чрез който са произведени, също няма висока стойност. Не е важно самото злато, а философският камък, който може да трансформира материята в злато. Това, което е очаровало хората, е методът, чрез който се достига резултатът, защото овладяването на метода дава контрол и върху крайния продукт на този метод.

Нека дадем един мисловен експеримент като пример за това защо трябва менталните модели като метод на познание да стоят по-горе в йерархията на ценени познавателни механизми от вярванията и репрезентациите.

Нека вземем две племена, които оцеляват в неблагоприятни условия. Едното се задоволява със знание, директно свързано с нуждите за оцеляване. Това е племето Знаещи. Другото племе се интересува от намирането на стратегии за оцеляване и абстрахирането на правила, които да служат за конкретните нужди. Това е племето на Моделите.

Знаещите се интересуват от конкретните плодове, които са ядливи. Знаят точните места, където те се намират. Знаят най-добрия момент за набирането им и т.н. Това са конкретните обосновани истинни вярвания, които конституират това, което приемаме за знание. Това, което не правят, е създаването на модел, който обединява знанията в правила или стратегия за откриване на закономерности, която да е от полза, в случай че нещо в реалността се промени и знанията, които са натрупани, не са приложими към новата епистемична среда.

Моделите се интересуват от закономерностите. Те вземат предвид конкретното знание, но не му отдават повече значение, отколкото е необходимо. Всеки модел се състои от входящи данни, механизъм и изходящи данни и моделите приемат знанието за средата за важно, но не по-важно от механизма, който открива абстрактното правило и обяснява нещо за епистемична среда. В случая, освен че знаят същото като Знаещите, те откриват, че има някакви предшестващи белези, които правят дадена реколта от плодове по-добра или по-слаба. Това им служи за различно поведение от това на Знаещите. Ако предстои по-слаба реколта, то те биха били по-внимателни в набирането на запаси и разпределянето им.

Едва ли е нужна компютърна симулация, която да покаже как тези две стратегии биха се справили в голям период. Днес е невъзможно и да си представим стратегия, изцяло зависеща от конкретни обосновани истинни вярвания като тази на Знаещите, защото хората са повече като тези от племето на Моделите. Използваме непрекъснато ментални модели, за да се ориентираме в света и да се справяме с промените в средата. Тази епистемична среда, която вече определих като присъщо несигурна, е изиграла определен натиск за формирането на познавателни схеми, различни от трупането на знания.

Това е подчертаната разлика между простото изчисляване на данни от ИИ и дълбоките модели в ИИ. Голямото развитие в тази сфера се дължи на преминаването от прости изчислителни схеми към сложна многопластова обработка. Показателно е за това, че моделите са по-надежден източник на разбиране.

Заключение

В никакъв случай не трябва да сме толкова тесногръди по отношение на ИИ, че да твърдим, че никога не биха достигнали нивото на обработка на информация, което да е равносилно на нашето. Няма теоретична или концептуална причина това да не се случи. Ако можем да алгоритмизираме нашите ментални модели, а това е напълно във възможностите ни, то ИИ ще отбележи значителен напредък.

Нещо повече, много концептуални анализи успяват да дефинират и обяснят работещите операции зад някои ментални модели, като мисловен експеримент, инверсия, сократически диалог, индукция и много други. Самият Сократ ни е оставил един от най-добрите инструменти за философско и научно изследване.

Това, което настоящата работа се опита да постигне, е да покаже, че приеманите за даденост епистемични фактори от средата не са очевидни сами по себе си. Опитът да се докаже това твърдение, се направи чрез сравнителен анализ между човешките познавателни механизми, в частност менталните модели, и обработката на данни от ИИ.

Първият аргумент за това, че епистемичната среда не предоставя очевидно значимата информация сама, е чисто концептуален. Това е аргументът за неизброимото количество от информация в света. Следващият аргумент в подкрепа на тезата е в неспособността за разпознаване на визуални образи от алгоритмите срещу относителната лекота, с която това се случва при хората. Изводът от предоставените примери е, че хората се справят по-надеждно с информацията при изображения. Това транслира и в невъзможността на този етап човешките водачи на превозни средства изцяло да се заменят от ИИ.

Въпросът следва да е коя способност на хората ги прави по-надеждни. Не е възможно да е просто изчислителната мощ, понеже в това поне компютрите ни превъзхождат. Не е в паметта. Обяснението – поне така, както изниква в

това изследване, е в механизма на филтриране на значимите данни и на обработката на същите. За разлика от ИИ, ние не сме „захранени“ с подходящите данни, а сами ги филтрираме от средата. След това обработката на тези данни става по механизъм, който все още не е алгоритмизиран или поне има трудност да бъде възпроизведен в ИИ.

Аналогията между менталните модели и дълбоките модели на ИИ не може да се изчерпи в настоящото изследване. Това, което може да вземем като извод в края на настоящата разработка, е, че менталните модели са по-надеждни, поне сравнени с настоящите алгоритми, в справянето с реалността, защото реалността не е толкова опростена, че знанието да е достатъчно. Знанието трябва да е интегрирано и трябва да му бъде приписана някаква стойност, за да може да действаме въз основа на него. Съвременните алгоритмите имат вградени механизми за оценка на резултатите, които произвеждат. Менталните модели правят това пресяване и организиране на информацията за нас. Това, което е знание, е произведено по някакъв начин, а този опосредстващ начин или механизъм – менталните модели, е и основен гарант за това знание.

ЛИТЕРАТУРА

ДЕКАРТ, Р., 1978. Избрани философски произведения. *Наука и изкуство,* София.

REFERANCES

ANON, V. G., 2022. Risks from advanced *AI Alignment Forum*. Available at: https://www.alignmentforum.org/posts/a8fFLg8qBmq6yv53d/vael-gates-risks-from-advanced-ai-june-2022 [Accessed August 29, 2022].

ANON, 2022. Gun deaths in America, 2020. *Injury Facts*. Available at: https://injuryfacts.nsc.org/home-and-community/safety-topics/guns/ [Accessed September 1, 2022].

ANON, 2021. expert survey on progress in AI. *AI Impacts*. Available at: https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/ [Accessed September 1, 2022].

ANON, 2021. The challenges of image recognition. *ML2Grow*. Available at: https://ml2grow.com/the-challenges-of-image-recognition/ [Accessed September 7, 2022].

ANON, 2022. How we are teaching computers to understand pictures. Available at: https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?language=en [Accessed September 7, 2022].

BELLAN, R., 2022. Despite a drop in how many companies are testing autonomous driving on California roads, Miles Driven Are Way up.

TechCrunch. Available at: https://techcrunch.com/2022/02/10/fewer-autonomous-vehicle-companies-in-california-drive-millions-more-miles-in-testing/?guccounter=1 [Accessed August 29, 2022].

DREW, T., Võ, M. L.-H., & WOLFE, J. M. 2013. The invisible gorilla strikes again. Psychological Science. **24(**9), 1848 – 1853. https://doi.org/10.1177/0956797613479386

DEKART, R., 1978. Izbrani filosofski proizvedenia. Nauka I izkustvo, Sofia

FARQUHAR, M., 2017. For nature cannot be fooled. why we need to talk about fatigue, *Anaesthesia***, 72**(9), pp. 1055 – 1058. https://doi.org/10.1111/anae.13982.

HARRIS, S., 2012. Black Swan. *The moral landscape*, London

ISMAIL, A., 2021. Let's all laugh at these bad autonomous car predictions from just a few years ago. *Jalopnik*. Available at: https://jalopnik.com/let-s-all-laugh-at-these-bad-autonomous-car-predictions-1846690460 [Accessed August 29, 2022].

Jr., C.R.W., 2019. McDonald's employee repeatedly uses racial slur in confrontation with black customer. *The Washington Post*. Available at: https://www.washingtonpost.com/nation/2019/01/14/mcdonalds-employee-repeatedly-uses-racial-slur-confrontation-with-black-customer/ [Accessed September 1, 2022].

MARKOFF, J., 2012. How many computers to identify a cat? 16,000. *The New York Times*. Available at: https://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html [Accessed September 7, 2022].

McMILLAN, R., 2015. This guy beat Google's super-smart ai-but it wasn't easy. *Wired*. Available at: https://www.wired.com/2015/01/karpathy/ [Accessed September 7, 2022].

Media, N.H.T.S.A., 2020. NHTSA releases 2019 crash Fatality Data. *NHTSA*. Available at: https://www.nhtsa.gov/press-releases/nhtsa-releases-2019-crash-fatality-data [Accessed September 1, 2022].

MEDINA, S., 2018. These are some of the 68 million people McDonald's serves every day. *Fast Company*. Available at: https://www.fastcompany.com/1672621/these-are-some-of-the-68-million-people-mcdonalds-serves-every-day [Accessed September 1, 2022].

PARRISH, S.2019. The great mental models. vol. 1

RICKS, T.E., 2011. Negligent discharges: One subject the military really doesn't like to talk about. *Foreign Policy*. Available at: https://foreignpolicy.com/2011/05/13/negligent-discharges-one-subject-the-military-really-doesnt-like-to-talk-about/ [Accessed September 1, 2022]. Rohrich, R.J., Savetsky, I.L. & Avashia, Y.J., 2020. Assessing

cosmetic surgery safety. *Plastic and Reconstructive Surgery – Global Open*, Publish Ahead of Print.
SALSBURG, D. 2002. The lady tasting tea: How statistics revolutionized science in the twentieth century. New York: Owl Book.
STICH, S. 1992. What Is a Theory of Mental Representation? Mind. 101(402), 243 – 261. http://www.jstor.org/stable/2254333
TANZ, O., 2017. Can Artificial Intelligence Identify Pictures better than humans? Entrepreneur. Available at: https://www.entrepreneur.com/science-technology/can-artificial-intelligence-identify-pictures-better-than/283990 [Accessed September 7, 2022].
ZAGZEBSKI, L., 1996. Virtues of the mind: An inquiry into the nature of virtue and ethical foundations of knowledge

COMPARATIVE ANALYSIS OF MENTAL MODELS AND ARTIFICIAL INTELLIGENCE

Abstract. Mental models are a representation of the external world. They are a hypothetical example of relevant elements to the knowing subject, that are configured by analogy to the world`s being. The formation of warranted beliefs can be a result of the very same reliability of this type of representation. The article argues for this position and defends the thesis that the world is epistemically insecure. This means that there is no objective reason in and of itself for certain elements to be preferred rather than other. This is a serious problem for AI researchers, who find difficulty in recreating the visual orientation ability of humans. By this analogy, mental models have high value in orienting us in the epistemic landscape. This article argues that mental models have higher epistemic value compared to AI.

Keywords: mental model; epistemic insecurity; artificial intelligence; reliability

✉**Plamen Chergarov**
ORCID iD: 0000-0002-2571-4793
Faculty of Philosophy
University of Sofia
125, Tzarigradsko Chaussee Blvd., bl. 5
1113 Sofia, Bulgaria
E-mail: plamenchergarov@gmail.com