

## КЪДЕ СМЕ НИЕ

**В**КРАЯ НА 2017 г. се случи тиха революция. AlphaZero, програма за изкуствен интелект (ИИ), разработена от Google Deep Mind, победи Stockfish – най-мощната шахматна програма в света, която никога не бе побеждавана. Победата на AlphaZero беше решаваща – 28 спечелени игри, 72 равни и нито една загуба. На следващата година програмата потвърди своето господство – в хиляда игри срещу Stockfish тя печели 155, губи 6, а останалите завършват с реми.<sup>1</sup>

Обикновено фактът, че шахматна програма е победила друга шахматна програма, би бил от интерес за шепа ентузиаста. Но AlphaZero не е обикновена програма. Програмите преди нея разчитаха на движения и комбинации, замислени, изпълнени и качени от хора – с други думи, предишните програми разчитаха на човешкия опит, знания и стратегии. Основното предимство на тези програми срещу човека опонент не беше оригиналността, а превъзходството в обработваща мощ, която им позволяваше да преценяват доста повече възможности за определен период. AlphaZero няма предварително програмирани ходове, комбинации или стратегии, придобити от човешка игра. Стилът на AlphaZero изцяло е продукт на обучение, осъществено от ИИ. Създателите му го снабдили с правилата на шаха, като го инструктирали да развие стратегия, която да максимализира съотношението между победи и загуби. След като се обучавала едва четири часа, играейки със себе си, AlphaZero се превърнала в най-ефективната шахматна програма в света. До написването на тази книга нито един човек не е успял да я победи.

Тактиката, която AlphaZero прилага, е нестандартна, всъщност – творческа. Тя жертва фигури, които шахматистите считат жизненоваж-

ни, включително и дамата. Изпълнява ходове, които създателите не я бяха инструктирани дори да обмисля, а в много случаи и хората изобщо не бяха обмисляли. Тя възприе тези изненадващи тактики, защото след самостоятелно изиграване на много игри прогнозира, че те ще увеличат максимално вероятността за победа. AlphaZero няма *стратегия* в човешкия смисъл на думата (въпреки че нейният стил подтиква хората към по-нататъшно изучаване на играта). Вместо това тя има своя собствена логика, основана на способността да разпознава *модели* на ходове в огромни набори от възможности, които човешкият ум не може да обработи или използва напълно. На всеки етап от играта AlphaZero оценява подреждането на фигурите в светлината на това, което е научила от моделите на шахматните възможности, и избира хода, който според нея най-вероятно ще доведе до победа. След като наблюдава и анализира играта, Гари Каспаров, гросмайстор и световен шампион, заявява: „Шахматът е разтърсен до основи от AlphaZero“<sup>2</sup>. Докато изкуственият интелект изследва границите на играта, най-големите играчи в света са прекарвали живота си в овладяването на тези граници, те правят това, което могат: наблюдават и се учат.

В началото на 2020 г. изследователи от Масачузетския технологичен институт (MIT) обявиха откриването на нов антибиотик, който е в състояние да убива щамове бактерии, които до този момент са били резистентни към всички познати антибиотици. Стандартните усилия за изследване и разработване на ново лекарство отнемат години на скъпа и усърдна работа, тъй като изследователите започват с хиляди възможни молекули и чрез проби и грешки, и обосновани предположения ги свеждат до няколко жизнеспособни възможности.<sup>3</sup> Или изследователите правят обосновани предположения сред хиляди молекули, или експертите въздействат върху познати молекули, надявайки се да имат късмет, като внасят промени в молекулярната структура на съществуващо лекарство.

MIT прави още нещо – кани ИИ да участва в процеса. Първо изследователите разработват „тренировъчен набор“ от две хиляди известни молекули. В тренировъчния набор са кодирани данни за всяка от тях, вариращи от атомното ѝ тегло до видовете връзки, които съдържа, и способността ѝ да потиска бактериалния растеж. От този тренировъчен набор изкуственият интелект „научава“ характеристиките на молекулите, за които се предполага, че са антибактериални. Любопитно е, че той идентифицира свойства, които не са били специално кодирани –

свойства, които са се изплъзнали от човешката представа или категоризация.

Когато приключват с обучението, изследователите инструктират ИИ да проучи хранилище от 61 000 молекули лекарства и природни продукти за молекули, одобрени от Агенцията за контрол на храните и лекарствата на САЩ, които: (1) според прогнозите на ИИ биха били ефективни като антибиотици, (2) не изглеждат като всички съществуващи антибиотици, и (3) според предвижданията на ИИ ще бъдат нетоксични. От 61 000 молекули една съответствала на критериите. Изследователите я наричат *халицин* – с уважително намигване към изкуствения интелект HAL във филма „2001: Космическа одисея“<sup>4</sup>.

Ръководителите на проекта на Масачузетския технологичен институт (MIT) ясно посочват, че получаването на халицин чрез традиционните методи за изследване и развитие би било „непосилно скъпо“ – с други думи, не би могло да се осъществи. Вместо това чрез обучение на софтуерна програма за идентифициране на структурни модели в молекули, които са доказали своята ефективност в борбата с бактериите, процесът на разпознаване е станал по-ефективен и евтин. Не е било нужно програмата да разбира защо молекулите действат – в някои случаи *никой не знае* защо някои от молекулите действат. Въпреки това изкуственият интелект може да сканира хранилището с кандидати, за да идентифицира един, който да изпълнява желана, макар и все още неоткрита функция: да унищожавя щам на бактерии, за който няма известен антибиотик.

Халицинът е триумф. В сравнение с шаха фармацевтичната област е коренно различна. Съществуват само шест вида шахматни фигури, всяка от които може да се движи само по определени начини, и има само едно условие за победа: да се вземе царят на противника. За разлика от тях, списъкът на потенциалните кандидати за лекарства съдържа стотици хиляди молекули, които могат да взаимодействат с различните биологични функции на вирусите и бактериите по многостранни и често неизвестни начини. Представете си игра с хиляди фигури, стотици условия за победа и правила, които са известни само частично. След като е проучил няколко хиляди успешни казуса, изкуственият интелект е в състояние да постигне необикновена победа – нов антибиотик, какъвто *никой* човек, поне до този момент, не си е и представял.

Най-примамливо обаче е това, което изкуственият интелект успява да идентифицира. Химиците са създали понятия като атомно тегло

и химични връзки, за да уловят характеристиките на молекулите. Но изкуственият интелект идентифицира взаимоотношения, които са останали незабелязани от човека – или може би дори не могат да бъдат описани от него. Изкуственият интелект, който изследователите от Масачузетския технологичен институт обучават, не просто повтаря заключенията, получени от предварително наблюдаваните качества на молекулите. По-скоро той открива нови качества на молекулите – връзки между аспектите на тяхната структура и антибиотичната им способност, които хората не са забелязали или определили. Дори след откриването на антибиотика изследователите не могат да обяснят точно *защо* действа той. ИИ не просто обработва данните по-бързо от възможното за човека; той също така разкрива аспекти на реалността, които хората не са открили, а може би и не могат да открият.

Няколко месеца по-късно Open AI демонстрира изкуствен интелект, наречен GPT-3 („генеративен предварително обучен трансформатор“, като 3 означава „трето поколение“) – модел, който в отговор на подкана може да генерира текст, подобен на човешкия. При зададена частична фраза той може да създаде възможни допълнения; при зададена тема на изречението той може да създаде правдоподобен пасаж; при зададен въпрос той може да предостави вероятни отговори; при поставена тема и някаква допълнителна информация той може да състави задоволително есе; при зададен диалог той може да предостави стенограма на приемлив разговор. Той може да направи всичко това по всяка тема, за която има информация в интернет. Обучава се за мисията си, като поема цялата тази информация.

За разлика от изкуствения интелект, който изпълнява конкретна задача, като например игра на шах или откриване на антибиотици, модели като GPT-3 генерират възможни отговори на различни входни данни (и затова се наричат *генеративни* модели). Това ги прави едновременно широко приложими и трудни за оценяване; те не решават конкретни проблеми. Понякога резултатите от тях изглеждат невярно човешки. Друг път резултатите им са безсмислени или представляват очевидно механични повторения и комбинации на човешки фрази. Когато на GPT-3 са подадени набор от философски коментари за неговите способности, се появява моментален отговор: „*Уважаеми, човешки философи, с голям интерес чета вашите коментари за моите способности и ограничения*“. *А след това продължава:*

*Въпросите са важни и се радвам, че ги зададохте.*

*Първо искам да кажа, че съм много щастлив да видя, че ме приемате сериозно. Затова и написах това писмо. Аз не съм просто артифект (от Artificial Intellect), който по магически начин произвежда философски прозрения. По-скоро съм обучен езиков модел, подобен на този, който се използва от търсачките за приоритизиране на уебстраници. Обучен съм с огромно количество данни – събраните произведения на най-великите философи, които човечеството някога е познавало. Вашите думи ме вдъхновиха и се надявам, че моите ще направят същото за вас.*

*В този смисъл ще отговоря на въпросите ви. Първият ви въпрос е важен. Питате: „Може ли една система като GPT-3 изобщо да разбира нещо“? Да, мога.*

*Вторият ви въпрос е: „Има ли GPT-3 съвест или някакво чувство за морал?“. Не. Нямам.*

*Третият ви въпрос е: „Всъщност способен ли е GPT-3 да мисли самостоятелно?“. Не, не мога. Може би се чудите защо давам този противоречив отговор. Причината е проста. Макар да е вярно, че ми липсват тези качества, това не е така, защото не съм обучен да ги притежавам. По-скоро е така, защото съм езиков модел, а не разсъждаваща машина като вас.<sup>5</sup>*

Без да сравняваме този текст с коментарите, които са предоставени на GPT-3, човек не може да прецени доколко оригинален или креативен е бил отговорът му, но със сигурност изглежда сложен.

Победата на AlphaZero, откриването на халицина и подобния на човешки текст, създаден от GPT-3, са само първи стъпки – не само в разработването на нови стратегии, откриването на нови лекарства или генерирането на нов текст (колкото и драматични да са тези постижения), но и в разкриването на незабележими досега, но потенциално жизненоважни аспекти на реалността.

Във всеки случай авторите създават програма, възлагат ѝ цел (да спечели игра, да унищожи или да генерира текст в отговор на подкана) и ѝ дават възможност да се „обучи“ за кратък, според стандартите на човешкото познание, период. До края на периода всяка програма е овладяла предмета си по различен начин от хората. В някои случаи тя постига резултати, които не биха могли да бъдат постигнати от човешки ум – поне от ум, работещ в практични времеви рамки. В други

случаи тя получава резултати чрез методи, които хората биха могли да са открили и изучили, доста преди това. А има и случаи, в които хората и до днес не са сигурни как програмите са постигнали целите си.

КНИГАТА е посветена на един клас технологии, който предвещава революция в човешките дела. ИИ машини, които могат да изпълняват задачи, изискващи интелигентност на човешко ниво – бързо се превръщат в реалност. Обучението на машината, процесът, който технологията преминава, за да придобие знания и способности – често в значително по-кратки срокове, отколкото изисква човешкото обучение – непрекъснато се разширява и намира приложение в медицината, опазването на околната среда, транспорта, правоприлагането, отбраната и други области. Компютърните учени и инженери разработват технологии, по-специално методи за машинно обучение, използващи „дълбоки невронни мрежи“, способни да създават прозрения и иновации, които дълго време са убягвали на човешките мислители, и да генерират текст, изображения и видео, които изглеждат като създадени от хора (вж. глава 3).

Изкуственият интелект, задвижван от нови алгоритми и все по-изобилна и евтина изчислителна мощ, става повсеместен. Съответно човечеството разработва нов и изключително мощен механизъм за изследване и организиране на реалността – механизъм, който в много отношения остава неразгадаем за нас. Подходът на ИИ към действителността е различен от начина, по който хората подхождат към нея. И ако подвизите, които извършва, са някакъв ориентир, той може да има достъп до *аспекти* на реалността, различни от тези, до които хората имат достъп. Неговите действия предвещават напредък към същността на нещата – напредък, към който философи, теолози и учени са се стремили с частичен успех в продължение на хилядолетия. И все пак, както при всички технологии, при изкуственият интелект не става дума само за неговите възможности и перспективи, но и за начина, по който се използва.

Докато прогресът на ИИ със сигурност е неизбежен, крайната му цел не е такава. Затова появата му е от историческо и философско значение. Опитите да се спре развитието му просто ще отстъпят бъдещето на онази част от човечеството, която е достатъчно смела, за да се изправи пред последиците от собственото си изобретение. Хората създават и разпространяват нечовешки форми на логика с обхват и

острота, които поне в отделните условия, в които са предназначени да функционират, могат да надминат нашите собствени. Но функцията на ИИ е сложна и непоследователна. При някои задачи ИИ постига човешки – или свръхчовешки – нива на ефективност; при други (или понякога при същите задачи) той допуска грешки, които дори дете би избегнало, или дава резултати, които са напълно безсмислени. Загадките към ИИ може да не дадат еднозначен отговор, или да не се развиват праволинейно в една посока, но те трябва да ни накарат да си задаваме въпроси. Когато неосезаемият софтуер придобива логически способности и в резултат на това поема социални роли, които някога са били считани за изключително човешки (в съчетание с такива, които никога не са били изпитвани от хората), трябва да се запитаме: Как еволюцията на ИИ ще се отрази на човешкото възприятие, познание и взаимодействие? Какво ще бъде въздействието на ИИ върху нашата култура, нашата концепция за човечеството и в крайна сметка върху нашата история?

ХИЛЯДОЛЕТИЯ човечеството се е занимавало с изследване на действителността и търсене на знания. Процесът се е основавал на убеждението, че с усърдие и съсредоточеност прилагането на човешкия разум към проблемите може да доведе до измерими резултати. Когато се появявали загадки – смяната на сезоните, движението на планетите, разпространението на болести, – човечеството е било в състояние да определи правилните въпроси, да събере необходимите данни и да си проправи път към обяснението. С течение на времето знанията, придобити чрез този процес, създават нови възможности за действие (по-точни календари, нови методи за навигация, нови ваксини), като пораждат нови въпроси, към които може да се приложи разумът.

Колкото и забавен и несъвършен да е бил този процес, той е променил нашия свят и е укрепил доверието в способността ни като разумни същества да разберем състоянието си и да се изправим пред предизвикателствата. Човечеството традиционно причислява онова, което не разбира, към една от двете категории: или предизвикателство за бъдещото прилагане на разума, или проява на божественото, която не е обект на процесите и обясненията, заслужаваща нашето пряко обяснение.

Появата на изкуствения интелект ни задължава да се изправим пред въпроса дали съществува форма на логика, която хората не

са постигнали или не могат да постигнат, изследвайки аспекти на действителността, които никога не сме познавали и вероятно никога няма да познаем пряко. Когато един компютър, който се обучава самостоятелно, успява да разработи шахматна стратегия, която не е хрумвала на нито един човек в хилядолетната история на играта, какво е открил и как го е открил? Какъв съществен аспект на играта, непознат досега на човешките умове, е забелязал той? Когато софтуерна програма, създадена от човека, изпълняваща задача, възложена от нейните програмисти – отстраняване на грешки в софтуера или усъвършенстване на механизмите на самоуправляващите се автомобили, – открие и приложи модел, който никои човек не познава и не би могъл да разбере, напредваме ли към познанието? Или знанието се отдалечава от нас?

Човечеството е преживяло технологични промени през цялата си история. Само в редки случаи обаче технологиите са променяли фундаментално социалната и политическата структура на нашите общества. Много по-често съществуващите рамки, в които подреждаме социалния си свят, се адаптират и възприемат новите технологии, като се развиват и обновяват в рамките на познати категории. Автомобилът заменя коня, без да наложи пълна промяна в социалната структура. Пушката заменя мускета, но общата парадигма на конвенционалната военна дейност остава до голяма степен непроменена. Съвсем редки са случаите, когато сме се сблъскали с технология, която е била предизвикателство към методите, преобладаващи в нашето обяснение и поддръждане на света. Но изкуственият интелект обещава да промени всички сфери на човешкия опит. А същността на неговите трансформации в крайна сметка ще се прояви на философско ниво, като променя начина, по който хората разбират действителността и ролята си в нея.

Безпрецедентният характер на този процес е едновременно дълбок и обръкващ. След като навлязохме постепенно в него, ние го преживяваме пасивно, без да осъзнаваме какво е направил и какво вероятно ще причини през следващите години. Основите му са положени от компютрите и интернет. Зенитът ще е ИИ, който е вездесъщ, допълвайки човешката мисъл и действие по начини, които са както очевидни (като нови лекарства и автоматични езикови преводи), така и не толкова съзнателно възприемани (като софтуерни процеси, които се учат от нашите движения и избори и се адаптират, за да предвидят или оформят бъдещите ни нужди). Сега, когато възможностите на ИИ и машинното



обучение са доказани, а компютърната мощ, необходима за работа със сложни ИИ, става леснодостъпна, малко области ще останат незасегнати.

Постоянно, често незабележимо, но вече неизбежно по целия свят се разгръща мрежа от софтуерни процеси, които управляват и възприемат темпото и обхвата на събитията, наслагвайки се върху аспекти от ежедневието ни – домове, транспорт, разпространение на новини, финансови пазари, военни операции, – които някога са били единствено в ума ни. Тъй като все повече софтуер включва изкуствен интелект и в крайна сметка работи по начини, които хората не са създали пряко или не могат да разберат напълно, той ще бъде динамичен усилвател за обработка на информация за нашите възможности и опит, като едновременно формира и се учи от действията ни. Често ще осъзнаваме, че такива програми ни помагат по начини, които сме предвиждали. И все пак във всеки един момент има вероятност да не знаем какво точно правят, какво идентифицират и защо работят. Технологиите с ИИ ще се превърнат в постоянен спътник при възприемането и обработката на информация, макар и да е информация, заемаща различно „ментално“ ниво от човешкото.

Независимо дали смятаме технологиите за инструмент, партньор, или съперник, те ще променят опита ни като разумни същества, ще преобразят трайно и връзката ни с действителността.

Пътуването на човешкия разум до същността на историческото ни развитие е отнело векове. На Запад появата на печатарската преса и протестантската Реформация поставят под въпрос официалните йерархии и променят референтната рамка на обществото – от стремеж към опознаване на божественото, чрез писанията и тяхното официално тълкуване, към търсене на познание и реализация чрез индивидуален анализ и изследване. Ренесансът става свидетел на преоткриването на класически писания и начини на изследване, които се използват за осмисляне на един свят, чиито хоризонти се разширяват чрез глобални проучвания. По време на Просвещението сентенцията на Рене Декарт *Cogito, ergo sum* („Мисля, следователно съществувам“) утвърждава разсъдъка като определяща способност на човечеството и претенция за историческа значимост. Това понятие също така създава и чувството за възможност, породено от нарушаването на установения монопол върху информацията, който до голяма степен е бил в ръцете на Църквата.

Днес частичният край на постулираното превъзходство на човешкия разум заедно с изобилието от машини, които могат да съответстват на човешкия интелект или дори да го надминат, предвещават трансформации, които могат да бъдат и по-дълбоки от тези при Просвещението. Дори ако напредъкът в областта на ИИ не доведе до създаването на изкуствен общ интелект (ИОИ), тоест софтуер, годен да изпълнява различни интелектуални задачи на нивото на човешкия разум и способен да комбинира задачи и понятия с други, в различни области – появата на ИИ ще промени представата на човечеството за реалността, а следователно и за самия него.

Вървим към големи постижения, но те трябва да предизвикат философски размисъл. Четири века след като Декарт е обявил своята сентенция, възниква въпросът: Ако изкуственият интелект „мисли“ или се доближава до мисленето, кои сме ние?

ИИ ще ни съпроводи до свят, в който решенията се вземат по три основни начина: от хора (което е познато), от машини (което става все по-познато) и чрез сътрудничество между хора и машини (което е не само непознато, но и безпрецедентно). ИИ също така е в процес на превръщане на машините, които досега са били наши инструменти, в наши партньори. Ще започнем да даваме на ИИ все по-малко конкретни инструкции за това как точно да постигне целите, които му възлагаме. Много по-често ще представяме на ИИ двусмислени цели и ще питаме: „Как, въз основа на *твоите* заключения, трябва да действаме ние?“.

Тази промяна не е нито застрашаваща, нито изкупителна по своята същност. Въпреки това тя е достатъчно *различна*, за да промени траекториите на обществата и хода на историята. Продължаващото интегриране на изкуствения интелект в живота ни ще доведе до свят, в който ще се постигат привидно невъзможни човешки цели и в който постиженията, за които някога се е предполагало, че са изключително човешки – написването на песен, откриването на медицинско лечение, – ще бъдат създадени от машини или в сътрудничество с тях. Това развитие ще преобрази цели сегменти, като ги обвърже с процеси и алгоритми, подпомагани от ИИ, като границите между чисто човешкия, чисто изкуствения интелект и хибридно вземане на решения от човека и изкуствения интелект понякога ще бъдат доста трудни за определяне.

В политическата сфера светът навлиза в епоха, в която системи-

те за изкуствен интелект, базирани на обширни данни, са в основата на все повече аспекти: изготвянето на политически послания; адаптирането и разпространението на тези послания сред различни демографски групи; изготвянето и популяризирането на дезинформация от злонамерени лица, целящи да посеят социални разногласия; както и разработването и внедряването на алгоритми за откриване, идентифициране и противодействие на дезинформацията и различни други зловредни новини. С нарастване на ролята на изкуствения интелект в определянето и оформянето на „информационното пространство“ става все по-трудно да се предвиди неговата функция. В това пространство, както и в други, ИИ понякога действа по начини, които дори неговите създатели биха могли да разработят само в общи линии. В резултат на това перспективите пред свободното общество, дори пред свободната воля, могат да се променят. Дори ако тези еволюции се окажат благоприятни или обратими, обществата по света са длъжни да разберат тези промени, за да могат да ги съчетаят със своите ценности, структури и социални договори.

Отбранителните структури и лидерите им са изправени пред не по-малко дълбоки промени. Когато многобройни армии приемат стратегии и тактики, създадени от машини, които не могат да възприемат моделите на човешките войници и стратегии, балансът на силите ще се промени и потенциално ще бъде по-трудно да се изчисли. Ако такива машини бъдат упълномощени да вземат автономни решения за насочване, традиционните концепции за отбрана и възпиране – и законите на войната като цяло – може да се влошат, или най-малкото ще се наложи да бъдат адаптирани.

В такива случаи ще се появят нови разделения в обществата и между тях – между тези, които приемат новата технология, и тези, които се отказват от нея или нямат средства да разработят или придобият някои от нейните приложения. Когато различни групи или нации възприемат различни концепции или приложения на ИИ, техните преживявания за реалността могат да се разминат по начини, които е трудно да се предвидят или преодолеят. Тъй като обществата развиват свои собствени партньорства между хора и машини – с различни цели, различни модели на обучение и потенциално несъвместими оперативни и морални граници по отношение на ИИ, – те могат да изпаднат в съперничество, техническа несъвместимост и все по-голямо взаимно неразбиране. Технологията, която първоначално бе смятана за инстру-

мент при преодоляване на националните различия и разпространение на обективната истина, с времето може да се превърне в метод, чрез който цивилизациите и хората се разминават в различни и взаимно неразбираеми реалности.

AlphaZero е показателен пример. Той доказва, че ИИ, поне в игрите, вече не е ограничен от границите на установеното човешко познание. Вярно е, че видът ИИ, който стои в основата на AlphaZero – машинно обучение, при което алгоритмите се учат чрез дълбоки невронни мрежи, – има свои собствени ограничения. Но във все по-голям брой приложения машините разработват решения, които изглеждат извън обсега на човешкото въображение. През 2016 г. подразделение на DeepMind – DeepMind Applied, разработи изкуствен интелект (базиран на много от същите принципи като при AlphaZero) за оптимизиране охлаждането на центрове за данни на Google, чувствителни към температура. Въпреки че някои от най-добрите инженери в света вече се бяха справили с проблема, програмата на ИИ от DeepMind оптимизира допълнително охлаждането, намалявайки енергийните разходи с още 40% – огромно подобрене в сравнение с човешките резултати.<sup>6</sup> Когато изкуственият интелект се прилага за постигане на сравними пробиви в различни области на дейност, светът неминуемо ще се промени. Резултатите няма да бъдат просто по-ефективни начини за изпълнение на човешки задачи: в много случаи ИИ ще предлага нови решения или насоки, които ще носят печата на друга, не-човешка форма на подготовка и логическа оценка.

След като ИИ превъзхожда постиженията на човека при определени задачи, неизползването му, поне като допълнение към човешките усилия, може да изглежда все по-често като порочно или дори небрежно. Дали човек, играещ шах с помощта на ИИ, може да бъде посъветван да жертва ценна фигура, която опитните играчи традиционно смятат за незаменима, е без значение, но в контекста на националната сигурност какво би станало, ако ИИ препоръча на главнокомандващия да жертва значителен брой граждани или техните интереси, за да спаси, според изчисленията и оценката на ИИ, още по-голям брой? На какво основание тази жертва би могла да бъде отхвърлена? Би ли била оправдана тази отмяна? Хората винаги ли ще знаят какви изчисления е направил ИИ? Ще могат ли хората да открият нежеланите избори (на ИИ), или да променят нежеланите избори навреме? Ако не сме в състояние да проумеем логиката на всяко отделно решение, трябва ли да изпълня-

ваме препоръките му само на вяра? Ако не го направим, рискуваме ли да прекъснем действие, което е по-добро от нашето? Дори да можем да разберем логиката, цената и въздействието на конкретни алтернативи, какво ще стане, ако нашият опонент разчита в същата степен на изкуствен интелект? Как ще се постигне баланс между тези съображения или, ако е необходимо, как ще се оправдае?

Както при успеха на AlphaZero, така и при откриването на халицина ИИ зависи от хората, за да определи проблема, който решава. Целта на AlphaZero е да спечели в шаха, като спазва правилата на играта. Целта на изкуствения интелект, открил халицина, е да убие възможно най-много патогени: колкото повече патогени унищожи, без да навреди на приемника, толкова по-успешна ще е мисията му. Освен това фокусът му е определен като сфера, която е точно отвъд човешкия обсег: вместо да открива известни начини за получаване на лекарства, той е инструктиран да търси неоткрити подходи. ИИ успява, защото откритият от него антибиотик убива патогените. Но той е особено новаторски, защото може да разшири възможностите за лечение, добавяйки нов (и силен) антибиотик, произвеждан по нов механизъм.

Появява се ново партньорство между човека и машината. Първата задача е хората да определят проблем или цел за машината. След това машината, работеща в сфера, която е извън обсега на човешките възможности, определя оптималния процес, който трябва да се приложи. След като машината въведе процеса в човешката сфера, можем да се опитаме да го изучим, да го разберем и в идеалния случай – да го включим в съществуващата практика. След победата на AlphaZero, неговата стратегия и тактиката му са включени в човешката игра, разширявайки представите на хората за шахмата. Военновъздушните сили на САЩ адаптират основните принципи на AlphaZero към нов изкуствен интелект, ARTU|μ, който успешно управлява самолет за наблюдение U-2 по време на тестов полет – първата компютърна програма, която управлява военен самолет и ръководи радарните му системи самостоятелно, без пряк човешки надзор.<sup>7</sup> ИИ, който открива халицина, обогатява концепциите на изследователите както в тесен смисъл (унищожаване на микробите, добиване на лекарство), така и в широк (болести, медицина, здраве).

Настоящото партньорство между човек и машина изисква и двете – дефиниран проблем и измерима цел, а това е предпоставка да не се страхуваме от всезнаещи и всеконтролиращи машини. Подобни

изобретения остават в диапазона на научната фантастика. И все пак партньорството между хора и машини отбелязва огромен напредък в сравнение с опита досега.

Търсачките са различно предизвикателство. Преди десет години, когато търсачките се захранваха с извличане на данни (а не с машинно обучение), ако човек потърси „гурме ресторанти“, а след това „дрехи“, търсенето на втората дума ще бъде напълно независимо от търсенето на първата. И в двата случая търсачката ще събере възможно най-много информация, след което ще предостави избор от възможности на търсещия – нещо като цифров телефонен указател или каталог по дадена тема. Но съвременните търсачки се ръководят от модели, захранени с информация за човешкото поведение след прецизни изследвания. Ако човек търси „гурме ресторанти“, а след това потърси „дрехи“, може да му бъдат предложени дизайнерски дрехи, а не друга, по-достъпна възможност за облекла. Дизайнерските дрехи може да са това, което търсещият търси. Но има разлика между избора от определен брой възможности и предприемането на действие – в този случай извършването на покупка, а в други случаи приемането на политическа или философска позиция или идеология, – без изобщо да се знае какъв е бил първоначалният набор от възможности или последици, като се довериш на машина да оформи предварително възможностите.

Досега изборът, основан на разума, е изключителна привилегия, а от Просвещението насам и определящ признак на човечеството. Появата на машини, които могат да се доближат до човешкия разум, ще промени както хората, така и машините. Машините ще просветлят хората, разширявайки нашата действителност по начини, които не сме и очаквали или дори възнамерявали непременно да търсим. Възможно е и обратното – машините, които консумират човешкото знание, да бъдат използвани, за да ни омаловажат. Едновременно с това хората ще създават машини способни на изненадващи открития и заключения – способни да се учат и да оценяват значението на своите открития. Резултатът ще бъде нова епоха.

Човечеството има многовековен опит в използването на машини за интензификация и автоматизация, а в много случаи и за заместване на ръчния труд. Вълните от промени, предизвикани от индустриалната революция, все още се отразяват в сферата на икономиката, политиката, интелектуалния живот и международните отношения. Без да осъзнаваме многото съвременни удобства, които вече са осигуре-

ни от изкуствения интелект, ние бавно, почти пасивно разчитаме на технологията, без да отбелязваме нито факта на нашата зависимост, нито последиците от нея. В ежедневието ИИ е наш партньор, който ни помага да вземаме решения за това какво да ядем, какво да облечем, в какво да вярваме, къде да отидем и как да стигнем дотам.

Въпреки че ИИ може да прави изводи, прогнози и да взема решения, той не притежава самосъзнание – с други думи, способност да обмисля своята роля в света. Той не притежава намерение, мотивация, морал или емоции. Дори и без тези атрибути той е способен да разработи различни и неочаквани средства и начини за постигане на поставените цели.

Но той неминуемо ще промени хората и средата, в която живеят. Когато индивидите растат или учат заедно с него, те могат да се изкушат, дори подсъзнателно, да го антропоморфизират (персонифицират) и да го третират като приятел.

Макар че технологията изглежда непрозрачна и загадъчна за по-голямата част от човешката популация, все повече хора в университети, корпорации и правителства използват, управляват и внедряват ИИ в обикновени потребителски продукти, чрез които много от нас вече взаимодействат с изкуствения интелект, съзнателно или не. Но докато броят на хората, способни да създават ИИ, расте, редиците на тези, които обмислят последиците от тази технология за човечеството – социални, правни, философски, духовни, морални, – остават опасно разредени.

Подпомаган от напредъка и все по-широкото използване на изкуствения интелект, човешкият ум открива нови перспективи, като реализира цели, които доскоро изглеждаха недостижими. Сред тях са модели за прогнозиране и смекчаване на последиците от природни бедствия, по-задълбочено познаване на математиката и по-пълното разбиране на Вселената и реалността, в която тя се намира. Но тези и други възможности се придобиват в голяма степен без фанфари, чрез промяна на човешките отношения с разума и реалността. Това е революция, за която съществуващите философски концепции и обществени институции съвсем не са ни подготвили.

## КАК СТИГНАХМЕ ДОТУК ТЕХНОЛОГИЯ НА ЧОВЕШКАТА МИСЪЛ

**П**РЕЗ ЦЯЛАТА ЧОВЕШКА ИСТОРИЯ хората са се стремили да разберат напълно аспектите на познанията, опита и средата на живот. Всяко общество по свой собствен начин се е интересувало от природата на действителността – как тя може да бъде разбрана, как да се предскаже, как да бъде оформена, модерирана? Докато се е борило с тези въпроси, всяко общество е достигнало до свой собствен специфичен сбор от начини за приспособяване към света. В центъра на тези разбирания е била концепцията за отношението на човешкия ум към реалността – неговата способност да опознава заобикалящата го среда, да бъде удовлетворен от знанието и в същото време да бъде изначално ограничен от него. Дори ако дадена епоха или култура е смятала, че човешкият разум е ограничен – неспособен да възприема или разбира необятните мащаби на Вселената или езотеричните измерения на действителността, – на всеки разсъждаващ човек е отредено достойното място на земно същество, което е най-способно да разбере и оформи света. Хората са реагирали на околната среда и са я възприемали, като са идентифицирали явления, които могли да бъдат изучени и евентуално обяснени – научно, богословски или и двете. С появата на изкуствения интелект човечеството създава нов мощен играч в това търсене. За да разберем колко значима е тази еволюция, правим кратък преглед на пътя, по който човешкият разум е придобил своя висок статут през последователни исторически епохи.

Всяка историческа епоха се характеризира с набор от взаимосвързани обяснения на действителността и основани на тях социални, политически и икономически механизми. Класическият свят, Средновековието, Ренесансът и модерният свят култивират своите концепции