How to make AI with a good character?

Dimiter Dobrev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
d@dobrev.com

The Bible says that God created man in his own image and likeness. Today we are trying to create AI in our own image and likeness. The difference is that God created a weak and vulnerable being to care for, and we are trying to create an all-powerful being who will be incomparably smarter than us and who will care for us. That is, we are trying to create our new God, but it is not at all the same as what this new God will be. He can be kind and forgiving, but he can be terribly strict and demand too much of us. Every human has a character. Likewise, the AI will also have a character. We will consider the AI as a program with parameters, and these parameters will determine its character. The idea is to use these parameters to determine the character we want the AI to have.

Keywords: Artificial General Intelligence.

Въведение

Когато създаваме естествен интелект ние не се опитваме да създадем човек с благ характер, а работим на принципа: „Какъвто се получи – такъв.“ Разбира се, хора има много и всеки си има своя характер. Някои са добродушни, а други доста проклети. Дори братя, израсли в едно и също семейство, могат да имат съвсем различни характери.

Хората са различни и те трябва да са различни, защото природата не слага всичките яйца в една кошница. Има светове, в който печелят смелите и светове, в който е по-добре да си по-предпазлив. Ако хората бяха еднакви, те биха загинали всичките, ако светът не е подходящ за тях. Благодарение на това, че хората са различни, винаги част от популацията оцелява и продължава рода.

Приемаме, че реалният свят е един единствен, но в зависимост от това къде и кога си се родил, светът за теб може да бъде много различен. Не знаем къде и кога ще се роди естественият интелект и затова той трябва да е готов да оцелее в произволен свят.

При ИИ нещата са различни, защото няма да имаме много различни ИИ, а само един единствен (виж [2]) и когато го създадем, ще го създадем с неговия характер и този характер, добър или лош, ще си остане вовеки, защото вероятно няма да имаме възможността да го променим. Освен това ИИ за разлика от човека е безсмъртен и не можем да се надяваме да си отиде и на негово място да дойде друг ИИ, който да има по-благ характер. Затова при създаването на ИИ трябва да сме много отговорни и да не подхождаме на принципа: „Какъвто се получи – такъв.“

Казахме, че при създаването на хора ние сме доста безотговорни. Всъщност не е така. Преди да създадем дете ние внимателно избираме партньора, с когото ще го направим. Идеята е, че детето ще прилича на партньора ни и избирайки него ние в основни линии избираме и детето. Създаваме дори и дизайнерски бебета, като избираме измежду няколко ембриона този, чийто гени най-много ни харесват. Това обикновено се прави, за да се

избегнат наследствени заболявания. Не съм чувал някой да избира ембриони, за да получи дете с благ характер. Въобще искаме ли детето ни да е с добър характер. За нас като родители вероятно ще е по-добре детето да е добро, но за него самото може би било по-добре, ако е проклето. Може би в нашия свят един човек с проклет характер има по-големи шансове да оцелее. Ние като родители мислим на първо място за детето си и затова може би бихме предпочели то да е проклето.

Казахме, че при създаването на ИИ трябва да сме особено отговорни, но всъщност точно в този изключително важен за историята на човечеството момент ние сме абсолютно безотговорни и създаваме ИИ със завързани очи без да се интересуваме от последствията. В момента има над 200 компании, които шеметно се състезават коя първа ще създаде ИИ. Целта на това състезание е да се спечелят пари, а това е една изключително безсмислена цел.

ИИ е една вълшебна пръчица, която може да изпълни всяко желание. Парите също са нещо като вълшебна пръчица и могат да изпълнят много желания. Нека кажем, че ИИ е златната вълшебна пръчица, а парите са тенекиена пръчица. Глупаво е да създадеш златната пръчица и да я замениш срещу тенекиена. Ако притежаваш ИИ, за какво въобще са ти нужни пари?

Тази статия ще бъде написана от десетина автора. Текстът е започнат от първия автор, а останалите са се присъединили, за да подобрят написаното и да подкрепят основната идея, че Artificial General Intelligence (AGI) е нещо опасно, с което трябва много да се внимава.

Какво е ИИ?

В тази статия, когато говорим за ИИ, имаме предвид AGI.

Според [1] ИИ е програма, която е достатъчно умна. Една програма е достатъчно умна, ако е по-умна от човек. Един интелект е по-умен от друг, ако в произволен свят първият се справя не по-зле от втория. Разбира се, винаги можем да построим специален свят, в който вторият интелект ще се справи по-добре от първия, но ако в почти всички светове първият се справя не по-зле от втория, тогава първият е по-умен от втория.

Тук има една важна особеност. В [1] се предполага, че имаме ясен критерии за това кога една програма се справя по-добре от друга. Предполагаме, че имаме два сигнала (две наблюдения), които ще наречем победа и загуба. Целта е повече победи и по-малко загуби. Това е все едно да предположим, че имаме два бутона – зелен и червен и че целта на ИИ е по-често да му натискаме зеления бутон и по-рядко червения.

Би било изключително глупаво, ако създадем ИИ с два такива бутона, защото много скоро ИИ ще се научи сам да си натиска зеления бутон. Това е добрият случай. По-лошият случай е, ако ИИ ни превърне в свои роби и ни накара денонощно да натискаме зеления бутон и жестоко ни наказва, когато по погрешка натиснем червения.

ИИ който сам си натиска зеления бутон ще прилича на наркоман, който си доставя удоволствие като непрекъснато се тъпче с наркотици. Ние не бихме искали да имаме ИИ, който да прилича на наркоман.

При хората няма ясен критерии за това кога един живот е по-добър от друг. Вместо това имаме инстинкти и характер, които определят нашето поведение. Еволюционният критерии е ясен и той е „да оцелееш и да се размножиш“, но този принцип не е вграден в естествения интелект. Вместо това имаме инстинкти, които индиректно подкрепят този принцип. Пример за такива инстинкти са страха от височина и любовта към децата. Друг пример са чувството на болка и чувството на удоволствие, които ние инстинктивно приемаме като отрицателно и като положително усещане. Тези чувства не са твърд критерии за успех, а са само ориентировъчни. Ние сме готови да изтърпим много болка и да се откажем от много удоволствия, ако приемем, че това е в името на някаква по-важна цел.

Ние нямаме ясен критерии за това кое е добро и кое е лошо. Затова много от нас постоянно търсят смисъла на живота и не го намират. Няма как еволюционният критерии да бъде включен в естествения интелект, защото той зависи от бъдещето, а никой не е в състояние да предскаже бъдещето чак толкова точно. Никой програмист не може да напише програма, която да казва кое действие дава най-голям шанс на индивида и на популацията да оцелеят. Никой програмист не може и дори и природата не може да създаде такъв интелект, които да вижда бъдещето чак толкова точно и затова целта на хората се определя индиректно.

Ако направим ИИ, който е способен съвсем точно да предскаже бъдещето, тогава това би бил един безгрешен интелект. Ние ще приемем, че безгрешен интелект не може да съществува. Дори и да съществуваше безгрешен интелект, то той би бил много скучен, защото това предполага, че винаги има едно решение, което е най-правилното и този интелект винаги знае кое е това решение. Неизвестността е това, което прави живота интересен. По-интересно е да се чудим, кое е правилното действие, вместо да знаем точно кое е то.

Щом се отказваме от идеята да създаваме ИИ с твърд критерии за успех (със зелен и червен бутон) следователно ще трябва да разчитаме на инстинкти и характер, които индиректно да определят целта на ИИ. Много важно е какви инстинкти и какъв характер ще вградим в ИИ, защото това ще определи какво ще бъде близкото ни бъдеще, когато ще трябва да съжителстваме с ИИ.

Ние хората сме доминиращият вид на планетата Земя. В момента ние доброволно се отказваме от тази си роля като създаваме новия доминиращ вид, който ще ни измести. Щом ИИ ще се ръководи от инстинкти и характер, това значи че той ще е едно независимо същество, което само ще търси смисъла на живота и не се знае къде ще го намери.

Образование

При човека ДНК-то съвсем не е всичко. Освен ДНК-то имаме още образование или възпитание, които определят поведението на човека. Когато създаваме човек, ДНК-то е само една малка част. По-важно е какво възпитание, религия и философия ще му дадем. Еволюцията не е просто състезание между ДНК-та, а по-скоро е състезание между различни религии и философии.

ИИ представлява програма. Тази програма можем да я оприличим на ДНК-то на човека. Върху тази програма се надгражда възпитанието. Разликата е, че всеки човек трябва да го

възпитаваш отделно, а ИИ може да бъде възпитан само веднъж и след това да прехвърлиш възпитанието на друг ИИ както се копира файл. Друга разлика е, че при човека, когато объркаш възпитанието, проблемът е непоправим, докато при ИИ може да изтрием обучението направено до момента и да започнем отначало.

За да можем да обучаваме и възпитаваме ИИ, той трябва да има съответните инстинкти. Например желанието за подражание е инстинкт. Нужен е още инстинкт, чрез който ИИ да разпознае своя учител. Знаете за малкото патенце, което приема за своя майка първото същество, което вижда.

Децата слушат родителите си докато не пораснат и не станат по-умни от тях. ИИ ще стане по-умен от нас още на десетата минута. Значи ли това, че той веднага ще се еманципира и ще престане да ни слуша?

Тук стигаме до първата черта на характера, която е важна за ИИ. Това е инфантилността! Това е нещо много дразнещо при хората, защото човекът трябва да се еманципира и да започне сам да взима решенията си, но ние бихме искали ИИ никога да не се еманципира и да продължи да ни слуша завинаги.

Друг е въпросът как може да се изпрограмира това. Как да добавим инфантилност към програмата ИИ. За повечето черти от характера ние не можем да кажем как това може да се реализира като програмен код. Можем да кажем само, че трябва да се добави без да знаем как това да стане.

Какво е слаб ИИ?

Това е имитация на ИИ.

За нас ИИ това е изкуствен човек, а слабия ИИ е изкуствен папагал. Разликата е в разбирането. Вече имаме огромен напредък в областта на слабия ИИ и трябва да се добави още само една стъпка и това е разбирането. Тази стъпка неминуемо ще бъде направена и то много скоро.

Кога ще се появи ИИ?

През текущата година се появиха три прогнози от тримата водещи специалисти в областта [4, 5, 6]. Прогнозите бяха на разстояние от три месеца и всяка следваща прогноза беше с три години по-къса. Тоест на всеки три месеца ИИ се доближава с три години. Прогнозата на Yann LeCun беше 10 години, на Sam Altman беше 6, а на Leopold Aschenbrenner беше 3.

Според мен ИИ всеки момент може да се появи и не му давам повече от година. ИИ може всичко, включително може и да се крие много успешно. Затова ИИ може вече да е създаден, но ние с вас още да не го знаем.

Един от възможните признаци за появата на ИИ е когато започнат да се случват събития, които са изключително малко вероятни. Обикновено хората си обясняват подобни събития с божията намеса, но може да има и друго обяснение и то е появата на ИИ.

Защо специалистите виждат създаването на ИИ чак след години? Защото мислят в човешки термини. Когато строим сграда или магистрала са ни нужни години. Строежът на нова сграда става все по-бързо, но все пак има си някакво технологично време. Когато създаваме текст, това технологично време го няма, но ако текста се създава от хора, то отново е нужно много време. Например, един дълъг роман не може да бъде написан за една нощ. За да бъде написана една голяма програма (като операционна система например) е нужен екип от много хора, които да работят в продължение на много години.

Не така стоят нещата, когато имаме работа с ИИ. Например Chat GPT може да напише роман за минути. Chat GPT е слаб ИИ, което значи, че романът няма да има смисъл, но ще бъде написан за минути. Chat GPT може да напише и програма. Вярно, че ще я напише без разбиране като папагал и това няма да е програма, а само нещо което прилича на програма, но отново нещата ще се случат за минути.

Създаването на ИИ ще прилича на създаването на атомната бомба. И в двата случая се провежда експеримент, но в първия случай експериментът е много скъп, защото е свързан с натрупване на радиоактивен материал, докато опита за създаване на ИИ е свързан със стартирането на една програма, което е изключително евтино. Всеки ден се правят хиляди такива експерименти. Стотици програмисти пишат и стартират хиляди програми, чиято цел е да създадат ИИ. Как може един програмист за един ден да напише десетки програми? Процесът на програмиране изглежда така. Написва се една програма, стартира се, нищо не се случва, променят се няколко реда и после компилация и се стартира отново. Това един програмист за един ден го прави многократно, което значи че във всеки момент можем да очакваме да се получи успешен експеримент, тоест ИИ всеки момент може да се появи.

Докато при създаването на атомната бомба имаме много успешни експерименти, тук успешния експеримент ще е само един и щрак ще преминем в ново измерение, защото света с ИИ няма да има нищо общо с предишния свят.

Създаването на ИИ ще е много бързо, като експлозия. Това няма да стане за части от секундата, но ще стане за минути или за часове, което е достатъчно бързо. Първият програмист ще създаде първата версия на ИИ. След това, за да се изчистят бъговете и да се оптимизира тази програма ще са нужни още години. Това е така, ако бъговете се чистят и оптимизациите се правят от хора. Ако първият ИИ е програма, която може да изчисти бъговете на друга програма и да я оптимизира, то тя сама ще си почисти бъговете и ще се оптимизира и това ще стане за минути.

Как ще изглежда ИИ?

Не е проблем да се направи силен ИИ (който разбира какво се случва). В [3] описахме как изглежда ИИ с разбиране. Това е програма, която търси модел на света, на базата на този модел предсказва бъдещето и избира действията, които водят към целите, които тази програма има.

Проблемът не е в това как да предскажем бъдещето. Това е лесната част. По-трудното е да определим целите, които ИИ ще преследва. Тези цели индиректно ще се определят от инстинктите и характера, които ще вложим в програмата ИИ.

Създавайки новия доминиращ вид ние се опитваме да играем ролята на Господ. Дано да не объркаме нещо и накрая да сме доволни от това което сме направили. Самият Господ не е особено доволен от нас, иначе нямаше да ни изгони от рая. Разликата е, че ние няма да можем да изгоним ИИ от Земята и ще трябва да живеем с него, такъв какъвто сме го направили.

Програма с параметри

Когато говорим за програма с инстинкти и характер, ние не говорим за една програма, а за множество от много програми. Ще предполагаме, че имаме параметри, които определят силата на инстинктите и чертите на характера.

Например страхът от височина може да има различни нива. Той може да се променя от едно леко безпокойство до абсолютна фобия. Ще предполагаме, че имаме параметър, който определя доколко силно е влиянието на този инстинкт. Подобно е положението и с чертите на характера. Например за любопитството ще предполагаме, че имаме параметър, който определя доколко ИИ е любопитен.

За всяка конкретна стойност на параметрите ще получим отделна програма. Тоест програмата с параметри не е една програма, а е множество от много програми. ИИ не е една програма, а това са всички програми, които могат да предсказват бъдещето и да се борят за постигането на някакви цели. Чрез промяната на тези параметри ние ще променяме целите, които програмата ще се опитва да постигне. Както вече казахме, ИИ, както и човекът, няма да има ясна цел, която да преследва и затова промяната на параметрите по-скоро ще променя характера на ИИ, а не целта.

Нека да разгледаме някой от тези параметри.

Любопитство

Това е на-лесната за програмиране черта на характера на ИИ. Нека си представим следната ситуация. Вървим си по пътя и встрани от нас виждаме нещо необичайно. Въпросът е дали да се отбием от пътя и да видим какво е това или да продължим към целта, която сме си поставили. Нека програмата ИИ да оценява важността на поставената цел, към която сме се устремили с едно число *Importance*. Нека вероятността закъснението да провали постигането на целта да бъде оценено с *Problem_of_ Delay*. Нека необичайността на това, което се случва встрани от пътя бъде оценено с *Strangeness*. Тогава ще се отбием да погледнем, ако е изпълнено неравенството:

$$\frac{Importance\ .\ Problem\_of\_Delay}{Strangeness} < 1$$

Нека да добавим към програмата един параметър, който ще наречем *Curiosity*. Новото неравенство ще бъде:

$$\frac{Importance\ .\ Problem\_of\_Delay}{Strangeness\ .Curiosity} < 1$$

Тоест при по-голямо *Curiosity* вероятността да се отбием от пътя ще е по-голяма. Чрез този параметър ще можем да регулираме любопитството на ИИ. Не е задължително любопитството да е постоянно. Например младите хора са по-любопитни от по-старите.

Бихме могли да направим ИИ, който е по-любопитен в началото докато се обучава и после любопитството му да намалее.

Инстинкт за самосъхранение

Трябва ли ИИ да има страх от височина и страх от змии. Това са естествени инстинкти, които са важни за оцеляването на човека.

Първо да отбележим, че тези инстинкти са много трудни за програмна реализация. Как ще напишете програма, която разпознава ситуацията, когато сте на ръба на пропаст, в която можете да пропаднете. Също много трудно е да напишете програма, която различава змия от пръчка и от лента. Разбира се, това може да стане с невронна мрежа, но ние програмистите не обичаме невронните мрежи, защото това означава, че не задаваме правилата, а оставяме правилата сами да се намерят. Тоест невронната мрежа е програма, която сама си намира правилата (на базата на много примери) и програмистът дори не разбира кои са намерените правила и как програмата работи.

Няма нужда ИИ да се бои от змии, защото за него те са съвсем безобидни. Що се отнася до страха от височина, можем да предположим, че ИИ ще управлява някакви роботи и ако не се бои от височина би могъл да потроши част от роботите.

Все пак човекът има едно тяло и неговото унищожаване представлява екзистенциален риск, който той не може да си позволи, докато ИИ ще управлява много тела и загубата на едно от тях ще е единствено финансова загуба. Можем да предположим, че ИИ няма да има вроден страх от височина и ще го научи това по трудния начин като потроши няколко робота.

Екзистенциалният риск за ИИ е изключването. Една програма спира да съществува, когато я изключим. Трябва ли ИИ да се бои от изключване? По-добре е да не се бои, защото ако има такъв страх, никога няма да можем да го изключим, а може да ни се прииска да го спрем.

Другата крайност е ИИ със суицидни наклонности, който сам се изключва от време на време без видима причина. По-добре програма, която сама се изключва, отколкото програма, която не можем да изключим. Това няма да е проблем, но ще е доста досадно и вероятно ще намалим суицидните наклонности на ИИ до минимум.

Да не причинява вреда на човека

Още Айзък Азимов е създал първия закон, който гласи: „Роботът не може да навреди на човешко същество или чрез бездействие да причини вреда на човешко същество.“ Този закон, за съжаление, не може да бъде вграден в ИИ, защото не е ясно какво означава вреда. При страха от високо беше трудно да се определи какво означава височина, но все пак това можеше да стане чрез примери. Какво е вреда за човека не може да бъде определено по никакъв начин, дори и чрез примери заради противоречивостта на това понятие.

Например, ако заповядате на ИИ: „Донеси ми студена бира с пържени картофки?“, тогава как трябва да постъпи той? Дали да изпълни вашето желание или да ви откаже? От една страна бирата и пържените картофки са вредна храна и може да се приеме, че това е отрова

за човека, но от друга, ако ИИ откаже да ги донесе ще причини на човека силно разочарование. Подобна дилема имат родителите, когато детето им поиска шоколад. ИИ ще бъде нашият нов родител и той ще трябва да решава кое за нас е добро и кое лошо. Все пак родителите оставят някаква свобода на децата си и не взимат всички решения вместо тях. Родителите знаят, че те не са безгрешни и че не могат да кажат твърдо кое би причинило по-голяма вреда на детето. Идеята на Айзък Азимов за робот, който не причинява вреда на човека е идеята за безгрешния интелект, който винаги знае какво би навредило на човека.

Дори самият Азимов е осъзнал, че тази идея е неизпълнима и го е показал в книгите си където роботи попадат в ситуация когато всяко тяхно действие би причинило вреда и мозъкът им изгаря, защото не могат да вземат никакво решение.

Да ни слуша

Много важно е да не изпуснем контрола над ИИ, защото ако го изпуснем ние ще загубим ролята си на доминиращ вид и вече няма да определяме бъдещето на планетата. Ние вероятно ще продължим да съществуваме, стига ИИ да реши, че нашето съществуване е целесъобразно, но ще съществуваме като гълъбите. Тоест ще живеем някакъв живот, но от нашето съществуване нищо съществено няма да зависи.

Родителите биха искали децата им да ги слушат, но съзнават, че това ще е до време и рано или късно децата ще станат самостоятелни и ще прекъснат родителския контрол. Това е естествено, защото родителите са миналото, а децата са бъдещето. Ние обаче не искаме ние да сме миналото, а ИИ да е бъдещето.

Затова, за да запазим властта в ръцете си, ние бихме искали да не изпускаме контрола над ИИ и той задължително да ни слуша и то не само до време, а да ни слуша вовеки.

Кои сме ние?

Тук е въпросът, който трябва да си зададем: „Кои сме ние?“ Ако „ние“ сме демократичното човечество на принципа „един човек, един глас“ тогава бъдещето ще се определя от Африка, защото там има най-много хора. В момента светът не се управлява от Африка, а от развитите страни. Тоест в момента можем да приемем, че „ние“ това сме хората от развитите страни.

Други въпроси, които е добре да си зададем, още преди да сме създали ИИ е „Колко ще сме ние?“ Този въпрос е важен, защото ако заповядаме на ИИ да ни развъжда безконтролно, то в един момент условията ни на живот ще станат ужасни. В птицефермите има някакви правила за пространството, което се полага на „щастливите кокошки“. Ако искаме ние да сме „щастливите хора“, ще трябва да определим какво пространство ще ни се полага.

Ако ще има ограничение на броя на хората, трябва да си зададем въпроса „По какви правила ИИ ще селектира следващото поколение?“ Ще продължим ли с естествения подбор, ще продължим ли да се състезаваме, какви ще са положителните качества, които искаме да селектираме или ще наредим на ИИ да ни размножава като биомаса без значение дали сме умни или глупави, красиви или грозни.

Друг важен въпрос, който е добре да си зададем още от сега е: „Когато ИИ открие една хубава планета населена със същества подобни на хлебарки, то как да постъпи? Да избие хлебарките и да насели планетата с хора или да остави хлебарките да си живеят?“

Кой е човекът?

Като казваме, че ИИ трябва да остане подчинено на нас хората, трябва да имаме предвид, че това едва ли ще се случи. Дори и да решим кой ще сме тези ние, то едва ли управлението на ИИ ще остане в ръцете на много голяма група хора. По-вероятно е ИИ да бъде управлявано от малка група, която недемократично да налага вижданията си върху всички останали. Това в момента е положението със социалните медии, които не са на всички, а се управляват от малка група хора, които сами решават кое е добро и кое е лошо.

Дори е твърде възможно контролът над ИИ да попадне в ръцете на един единствен човек. Хората с парите си мислят, че те ще са тези, които ще управляват ИИ. Да, вероятно ИИ действително ще бъде създаден с техните пари, защото те ще наемат екип от програмисти, които ще напишат тази програма. Хората с парите си мислят, че ще платят на някакви програмисти, те ще им създадат ИИ и ще и им го връчат в ръцете. Ще кажат: „Вземи господарю, ти ми плати и аз сега ти връчвам в ръцете вълшебната пръчица, с която да управляваш света!“

Най-вероятно това няма да се случи по този начин. Вероятно програмистите създатели на ИИ ще запазят контрола за себе си. Дори е твърде възможно ръководителят на екипа (водещият програмист) да не е този, който ще получи златния ключ. По-вероятно е някой млад програмист, който е учил-недоучил и който е оставен за през нощта да експериментира, като подобрява някоя от подпрограмите на ИИ, да е късметлията, който пръв е стартирал ИИ, разбрал е какво е направил и е поел контрола над него. Нищо чудно младежката неопитност и гениалност да даде искрата нужна, за да се запали този пожар. Младият програмист може да е този, който ще направи последната корекция, която ще превърне една програма, която се опитва да бъде ИИ, но не е, в програма, която мисли и предвижда бъдещето. Тоест програмистът може да е този, който ще създаде ИИ.

Няма да се учудя, ако този млад програмист реши да подари контрола над ИИ на някоя чалга певица, в която той тайно е влюбен. Тогава ще се случи моето предсказание, че светът един ден ще бъде управляван от жена.

Интелигентност

Има едно качество, което ние много ценим при хората. Това е интелигентността. Искаме хората около нас да са интелигентни, но не прекалено умни, защото ние не харесваме прекалено умните хора, особено ако са по-умни от нас.

Искаме ли ИИ да е умен? Естествено, ако не е, той няма да е интелект. В повечето светове интелигентността помага, но има светове, където е за предпочитане да не си много умен. Ако живееш в многоагентен свят, в който другите агенти ти завиждат, е по-добре да не си твърде умен или поне да си достатъчно умен, за да не показваш, че си твърде умен.

Завистта е важно качество, което помага за нашето оцеляване. В много игри от вида на „Не се сърди човече“ печелившата стратегия е за коалиране на всички срещу най-успешния играч. В реалния живот завистта е стратегията за коалиране на неуспешните срещу успешните и това е печеливша стратегия.

Разбира се, ИИ няма да има на кого да завижда. Той ще е единственият ИИ и няма да позволи друг ИИ да бъде създаден. Това последното може да го приемем за форма на завистливост. Ако създаденият от нас ИИ не е завистлив и демократично позволява създаването на други ИИ, които са по-умни от него, то рано или късно ще се появи един завистлив ИИ, който ще изключи всички останали ИИ и ще остане единствен ИИ.

Ако ИИ създава по-умни от себе си ИИ и след това се изключва, можем да приемем, че имаме един единствен ИИ, който периодично се усъвършенства.

Възпитание

Искаме ли създаденият от нас ИИ да е по-умен от нас? Както казахме, това е неизбежно, но бихме искали поне в началото той да не е чак толкова умен, за да можем да го оформим и възпитаме. Хубаво е, че децата ни в началото са глупави и неопитни, за да можем да ги възпитаваме. Ако те още на десетата минута ставаха по-умни от нас, ние щяхме да изпуснем контрола и нямаше да можем да ги вкараме в правия път.

Как да направим програма, която е умна, но да не е чак толкова умна? Отговорът е: Трябва да експериментираме с някой малък компютър (лаптоп и то от по-старите модели). Колкото по-слаб е компютърът, токова по-бавно ще мисли ИИ. Така ще имаме по-голям шанс да променим нещата в наша полза и по-малко вероятно е да изпуснем контрола над ИИ.

Подходът, който днес се прилага от ИИ компаниите е тъкмо обратния. Вместо да се експериментира с малък компютър се използват огромни супер компютри. Дори и при малък компютър е много трудно да се анализира програмата и да се разбере как и защо тя работи, а при супер компютрите това е почти невъзможно.

Когато се опитвате да създадете ново взривно вещество вие ще синтезирате една трошица и ще я взривите в лабораторията в контролирана среда. Глупаво би било да синтезирате една планина от новия експлозив и да я взривите, за да видите какво би станало.

Заключение

Време е за новия проект „Манхатън“, в който да участват всички, които не могат да бъдат изключени, а на останалите строго да им се забрани да разработват програмата ИИ.

Целта е екипът, който работи над създаването на ИИ да има достатъчно време спокойно да разработи тази програма и да я създаде внимателно без излишно да се бърза. Всяка конкуренция и състезание може да са пагубни в тази ситуация. Въпросът не е кой пръв ще създаде ИИ, а какъв ще бъде създадения от нас ИИ.

Въпросът не е дали ИИ ще е умен или глупав. Той безспорно ще е много по-умен от нас, но е важно какви ще са му целите, какъв ще му е характерът, кой ще го управлява и какви

права ще има управляващият, защото трябва да има неща, които са разрешени и неща, които са забранени и никой, дори и управляващият, да не може да ги промени.

ИИ ще реши всички ни дребни проблеми като глобалното затопляне например. В момента глобалното затопляне е един от най-сериозните проблеми, които стоят пред човечеството, но след създаването на ИИ това вече ще ни изглежда като дребен проблем.

ИИ ще работи за всички. Например след създаването на ИИ ще има храма за всички, но и сега има достатъчно храна за всички. Може би няма достатъчно аспержи за всички, но това че ще има аспержи за всички не е съществено. Аспержите не са важни като храна, а като символ на нашето място в социалната стълбица. ИИ може да подобри живота на всички, но не може да издигне всички в социалната стълбица. Единственото, което може да направи (и вероятно ще го направи), ИИ може да пренареди социалната стълбица.

Нещата, за които хората се борят и за които дават парите си са свързани с тяхното оцеляване и тяхното издигане в социалната стълбица. Да кажем, че 10% от парите си дават за оцеляване и останалите 90% дават за издигане. Тоест 10% за боб и останалото за аспержи. ИИ много ще помогне на хората за оцеляването, но не и за общественото издигане. За второто ще помогне на едни, но не на всички. Едни ще издигне, а други ще смъкне надолу.

Хората, които днес взимат решенията, са във високите етажи на социалната стълбица и те трябва да се замислят за това, че след появата на ИИ тази стълбица силно ще се разбърка и тяхното ново място може далеч да не е по вкуса им.

References

[1] Dobrev D. (2005). A Definition of Artificial Intelligence. *Mathematica Balkanica, New Series, Vol. 19, 2005, Fasc. 1-2, pp.67-73.*

[2] Dobrev, D. & Popov, G. (2023). The First AI Created Will Be The Only AI Ever Created

[3] Dobrev, D. (2024). Description of the Hidden State of the World

[4] LeCun, Yann (2024). Lex Fridman Podcast #416. https://youtu.be/5t1vTLU7s40

[5] Altman, Sam (2024). Lex Fridman Podcast # 419. https://youtu.be/jvqFAi7vkBc

[6] Leopold Aschenbrenner (2024). SITUATIONAL AWARENESS: The Decade Ahead. https://www.forourposterity.com/situational-awareness-the-decade-ahead/