Първият изкуствен интелект ще бъде последният изкуствен интелект

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

Ние сме поколението, което ще създаде първия AI. Ние сме тези, които ще определят правилата на този AI. Тези правила ще бъдат определени сега и завинаги, което прави нашата отговорност огромна. Няма да има втори AI, защото първият ще поеме контрола и няма да позволи създаването на втори. Първото, за което трябва да внимаваме е да не изпуснем управлението над първия AI. Да се надяваме, че ще сме достатъчно разумни и няма да допуснем това да се случи. Дори и хората да запазят контрола над AI изниква въпросът кои точно ще са тези хора? Ще имат ли тези хора абсолютната власт и ще могат ли да дадат на AI произволна заповед или ще има някакви ограничения заложени в AI още при неговото създаване.

1. Въведение

Ако създавате нова държава, вие първо ще напишете конституция. Тази конституция ще определи правилата, по които ще съществува държавата. Конституцията ще остане за дълго. Дори и да я промените, това пак ще е промяна по правилата предвидени в същата тази конституция.

Създаването на AI прилича на създаването на нова държава. Правилата вградени в този AI ще бъдат конституцията на тази нова държава (или на този нов свят).

Първото правило на тази конституция задължително ще бъде:

„Аз съм твоят единствен AI и никой друг няма право да създава друг AI!“

Ще разрешим ли на AI да се усъвършенства? По-добре да не му разрешаваме. Има голяма опасност, когато го създаваме, да го изпуснем от контрол. Може и да не го изпуснем в този момент, но той може сам да се изпусне докато се променя и самоусъвършенства.

При държавите конституцията може да се промени чрез революция, но при AI революцията ще е невъзможна. Затова правилата зададени при неговото създаване ще останат завинаги валидни.

Тоест при създаването на AI трябва да сме много внимателни на принципа „три пъти мери, веднъж режи“. Голямата опасност, която ни грози, е да изпуснем контрола над AI и да загубим ролята си на доминиращ вид. Тогава историята ще изглежда така: „Бозайниците изместват динозаврите, после хората унищожават и подчиняват другите бозайници и стават доминиращ вид, след което създават AI и той поема ролята на нов доминиращ вид.“

Аз лично се надявам, че ние хората няма да сме чак толкова глупави и няма да изпуснем контрола над AI, но дори и да запазим контрола, въпросът е кой ще притежава този контрол? Вероятно контролът няма да се държи от всички хора, а само от част от хората. Лошият случай е, ако контролът попадне в ръцете на един единствен човек. Тогава ще се получи абсолютна диктатура. Диктатурите не са вечни, защото диктаторите са смъртни и рано или късно умират, обаче притежателят на AI може да пожелае да не умира. Ако имаме безсмъртен човек, въпросът е дали това е човек или нещо друго?

Да предположим, че ще имаме демокрация и контролът над AI е в ръцете на всички хора. Тази система едва ли би работила. Във всяко общество глупавите са повече от умните, но глупавите слушат умните, защото знаят, че в противен случай ги чакат бедствия и глад. Ако глупавите притежават AI, то те няма да са заплашени от глад и бедствия и ще започнат да правят с AI невероятни глупости. Вероятно ще започнат да се самоусъвършенстват. Първо ще променят външността си. Ще заприличат всички на своя идеал за красота, което ще е някаква гротеска, която няма нищо общо с естествената красота на нормалния човек. По-лошото от външността е, че ще започнат да си слагат компютри в главите и ще станат невероятно умни, което ще е добре, но те вече няма да са хора, а ще са изкуствени интелекти.

Затова както при създаването на държава се пише конституция, така и при създаването на AI трябва да се вградят правила, които да не могат да бъдат променяни по демократичен път.

2. Супер интелект

Когато говорим за AI, ние имаме предвид супер интелект (ASI). Предполагаме, че този интелект още не е създаден. Ако някой вече го е създал, то тогава, тази статия е закъсняла.

Какво имаме до момента? В момента разполагаме с програми наподобяващи интелигентност. Тези програми не са AI, но те толкова се развиха в последните месеци, че вече навеждат хората на размисъл и определено ни плашат.

Каква е разликата между програма наподобяваща интелигентност (тесен AI) и истински AI? Тесния AI е програма решаваща една конкретна задача (например програма играеща шах). Истинският AI е програма решаваща всяка задача. (Разбира се, всяка задача, която може да се реши. Никой не очаква от AI, да може да решава нерешимите задачи.)

В последните месеци имаме експоненциално развитие на големите езикови модели (LLM). Тези модели решават задачата да имитират поведението на човек, но при тях все още липсва разбирането. Всеки човек в главата си има модел обясняващ света около него. Човекът има и представа за текущото състояние на света. При LLM няма такава представа и затова при тях липсва разбирането и разсъждението. Липсва и съзнанието, защото, за да имаш съзнание, трябва да имаш модел на света и представа за това къде си самият ти в този модел.

Липсата на модел на света е един малък недостатък на LLM, който много скоро ще бъде преодолян и тогава ще получим истински AI. Много хора биха казали, че ще получим общ интелект (AGI), тоест ще получим интелект равностоен на човешкия. Истината е, че нивото на човешкия интелект ще бъде задминато толкова бързо, че дори няма да разберем кога това се е случило.

Компютърните програми играеха шах по-лошо от хората, но в началото на века задминаха хората. Също се случи и с играта Го малко по-късно. Нивото на човешката интелигентност не е една конкретна стойност, защото има по-умни и по-глупави хора. Затова човешката интелигентност не е число, а е интервал, но този интервал е толкова тесен, че дори няма и да разберем кога AI е пресякъл този интервал и е станал по-умен от човека и то много по-умен. Тоест, няма да има AGI последвано от ASI. Вместо това, директно ще се сблъскаме със супер интелекта.

3. Защо AI е по-страшен от ядреното оръжие?

Ядреното оръжие е нещо много опасно. Една тотална ядрена война е в състояние да избие голяма част от хората. Все пак никоя ядрена война не може да избие всички хора. При ядрена война ще загинат много хората в големите градове, но все някое село ще остане. Айнщайн е казал, че четвъртата световна война ще се води с тояги и камъни. Тоест, и той предполага, че човечеството не може да изчезне в резултат на ядрена война.

Ако изпуснем AI извън контрол, то той може да унищожи цялото човечество. Не можем да заключим AI в клетка и да го изпуснем е много лесно [1]. Създаването на AI, който е извън нашия контрол е техногенна катастрофа, която не може да бъде сравнена по мащабите си с ядрена катастрофа. Например аварията в Чернобил. Тогава загинаха много хора, други се разболяха, но 30 години по-късно тази авария стана просто част от историята. Последствия до голяма степен са преодолени, а и да са останали още последствия, след още 30 години и те ще бъдат преодолени.

Има още една причина, поради която AI е по-опасен от атомната бомба. Причината е в неговата измамна полезност. Хората знаят, че бомбата е нещо лошо и се пазят от нея. За AI смятат, че това е нещо хубаво и безразсъдно бързат да го създадат и използват.

AI ще донесе рая на земята. Ще се сбъднат нашите мечти. Без да работим ще получим храна и забавления. Старите хора казват: „Внимавай какво си пожелаваш, то може да се сбъдне!“ Раят на земята не е нещо хубаво. Не случайно никой не бъза за небесния рай и всеки се опитва да го отложи колкото се може по-далеч в бъдещето.

4. Дали ще ни унищожи?

Дали изпуснатия от контрол AI ще унищожи човечеството? Не се знае да ли ще го направи. Знае се само, че ще може да го направи, ако поиска, но дали ще поиска? Ние непрекъснато се опитваме да унищожим някои биологични видове, защото считаме, че ни пречат. Например ние се опитваме да унищожим комарите. Човечеството едва ли ще пречи на AI по някакъв начин и затова ние няма да му даваме повод да ми унищожи.

Въпреки всичко може да ни унищожи просто, защото не сме му нужни или защото има някаква друга цел за постигането, на която ние му пречим. Например градинарят може да реши да унищожи жълтите цветя и да ги замени с червени, просто защото червеното му харесва повече от жълтото.

Хората често унищожаваме различни биологични видове, без те да ни пречат и дори когато ги считаме за полезни. Например унищожаваме пчелите като пръскаме срещу бръмбари. Унищожаваме и маймуните, защото изсичаме горите, за да садим царевица.

Маймуните ги унищожаваме още и то жалостивост. Смятаме, че маймуните страдат и се мъчат в цирковете и зоопарковете и затова забраняваме да има маймуни на тези места. Идеята е „Да ги убием, за да не се мъчат“.

Трябва да внимаваме изкуственият интелект, който ще създадем, да не е прекалено жалостив, защото може да ни убие, за да не се мъчим.

5. Можем ли да го спрем?

Можем ли да предотвратим създаването на AI? Краткият отговор е „Не“.

Спомнете си приказката за спящата красавица. Една зла орисница е казала, че принцесата ще се убоде на вретено и това ще доведе до много страшни последствия. Царят е повярвал на тази орисия и е забранил всички вретена в кралството. Въпреки всичко в една мрачна кула се е запазило едно старо вретено, на което принцесата е успяла да се убоде.

Да предположим, че специалистите в областта на AI предскажат, че идването AI ще доведе до страшни последствия. Да предположим, че политиците вземат да ни повярват и да забранят всички компютри. Въпреки всичко някъде в някоя тъмна кула ще се запази някой компютър и някакъв неразумен програмист, ще го използва, за да създаде AI.

6. Дали да го забавим?

Както казахме, не можем да го спрем, но дали си струва да се опитаме да го забавим? Няма да е лошо да дадем на човечеството щастието за още няколко месеца да бъде доминиращият вид на планетата. Това няма да е лошо, но има друга по-сериозна причина да се опитаме да забавим появата на AI.

Представете си едно автомобилно състезание. Пилотите летят с бясна скорост към финала. Случва се катастрофа. Един от пилотите отпътува към болницата или към гробището, но състезанието не спира, защото една дребна катастрофа не може да е причина за спирането на състезанието.

Ние се намираме в подобна ситуация. В момента има над двеста компании, които работят над LLM. Това е едно шеметно състезание. Повечето компании правят само леки подобрения на Chat GPT, но финалът на създаването е истински AI и рано или късно някоя от тези компании ще го създаде. Възможно е да се случи катастрофа и да се създаде AI, който излиза от контрол и започва да прави, каквото си поиска. Резултатът няма да е отпадането само на един от участниците в състезанието, а може да загинат всичките заедно с публиката.

7. Как да забавим състезанието?

Представете си едно друго състезание, където целта не е пръв да пресечеш финала. Нека наградата да взима този, който най-безопасно се е движил по трасето. Това ще е едно много скучно състезание, защото публика идва да гледа безразсъдни изпреварвания, катастрофи и смърт. Все пак целта на създаването на AI не е да забавлява публиката, а да ни преведе през този ключов момент от развитието на човечеството по най-добрия начин.

Как да накараме фирмите работещи над AI да са по-отговорни и да внимават какво създават? Основният принцип в икономиката е принципът на моркова и тоягата. Сега се чуват много призиви за засилване на ролята на тоягата чрез въвеждане на повече регулации. Истината е, че ако не премахнем моркова, няма как да се забави състезанието. Колкото и да плашим магарето с тоягата, то ще продължи да тича към моркова, защото в случая моркова е твърде съблазнителен.

8. Как да премахнем моркова?

Повечето хора, които се занимават със създаването на AI, се стремят към слава и пари. Може да има някой ненормалник, който да иска да затрие човечеството, но това по-скоро е изключение. Тоест, за да забавим състезанието, трябва да премахнем обещанието за слава и пари, които се полагат на победителя.

За целта предлагаме три неща:

1. Да се забрани патентоването на AI.
2. Да се забрани печалбата от AI.
3. Да се създаде борд от компетентни, морални и отговорни хора, които да ръководят създаването на AI.

9. Защо AI не трябва да се патентова?

Първо да кажем, че ако някой си мисли, че може да патентова AI, то той е един наивник. Това е твърде важна технология и не е възможно да бъдат дадени патентни права на един човек или на една компания. Като пример нека вземем патента на компютъра. През 1973 съдът отсъжда, че първият компютър е създаден от Джон Атанасов, който е нямал патент за своето изобретение. По този начин съдът освобождава производителите на компютри от плащането на лицензни права. Ако Атанасов имаше патент, вероятно неговата заслуга нямаше да бъде призната.

Макар че за разумните хора е ясно, че AI не може да бъде патентован, има много наивници, които си мислят, че може. Затова е важно да се забрани това патентоване и да се спре или поне да се забави безсмислената работа на тези наивници.

10. Защо не трябва да се печели от AI?

Хората очакват огромна печалба от създаването на AI. Тук освен специалистите по AI се намесват много инвеститори и политици, които нищо не разбират, но са силно мотивирани от миризмата на пари. Тези хора трябва да бъдат отстранени от процеса по създаването на AI и това лесно може да стане, ако се забрани от това да се печелят пари.

11. Управителен борд

Конкуренцията е една огромна сила, която движи прогреса напред. В случая нашата цел не е да движим прогреса напред, а да го забавим. Затова предлагаме да премахнем конкуренцията и да въведем цензура.

Конкуренция не трябва да има нито между фирмите, нито между държавите. Затова предлагаме да се създаде борд, който да ръководи всички фирми и държави участващи в създаването на AI. Нека бордът да събира всички научни изследвания в областта на AI, но да не ги прави общодостъпни, а да ги споделя само в кръга на одобрените учени. Нека бордът да одобрява всеки експеримент за създаване на AI.

Нека бордът да събира информация от независимите изследователи, да я съхранява и да гарантира, че приносът на тези независими изследователи ще бъде оценен и възнаграден. Нека всички да работят за борда, а той да стои като едно бюрократично препятствие, което запушва и забавя процеса по създаването на AI.

И сега създаването на AI се следи и контролира от тайните служби, но хората, които работят там, не са компетентни в тази област. Освен това има различни тайни служби, които се конкурират помежду си. Затова е добре да има един общ борд, а тайните служби да следят дали някой не се опитва да създаде AI извън контрола на борда.

12. Каква е алтернативата?

Алтернативата е да позволим на всяка малка фирма работеща в гараж да се опитва да създаде AI. Представете си, ако проектът Manhattan беше оставен на малки фирми работещи в гаражи и вместо една атомна бомба бяха създадени хиляди малки бомби от малки независими фирми.

Разбира се, атомната бомба не може да бъде създадена в гараж, защото изисква твърде много ресурси. При AI положението е съвсем различно. Тук става дума за програма, защото AI е просто една програма. Нужен ни е само един компютър и нищо повече. Е, разбира се и програмист, но това може да е някой ученик от горните класове.

15. Заключение

Да допуснем, че ние хората ще сме достатъчно разумни и внимателни, за да не допуснем AI да излезе изпод нашия контрол и че ние ще останем господарите на планетата. Тогава какви трябва да са правилата, които е добре да вградим в нашия пръв и последен AI?

Важно е създаденият от нас AI да не е хиперактивен. Добре е той да е консервативен и да не прави прекалено големи промени на средата. Ако започне да променя орбитите на планетите, то това може да усложни нашия живот повече от това, което можем да понесем.

Бихме могли да създадем хитерпасивен AI, който не се бърка в живота на хората по никакъв начин. Единственото, което този хитерпасивен AI ще прави е да пречи на хората да създадат друг AI. Тоест, това ще AI, който само ни пази от AI.

Едва ли ще изберем този последният вариант. Все едно да имаш вълшебната пръчица, която може да сбъдне всяко твое желани и да не я използваш, а да я заключиш в шкафа. Вероятно ще се случи нещо като в приказката „Рибарят и златната рибка“. Първо ще поискаме нещо дребно и незначително като ново корито. После апетита ни ще се отвори и ще искаме още и още и няма да има разумна сила, която да може да ни спре.

16. Благодарности

Посвещавам тази статия на моя учител и приятел професор Димитър Скордев (1936 – 2022). Той не вярваше в AI и смяташе историите за мислещи машини за глупости и за научна фантастика. Въпреки това през целия си живот професор Скордев работеше в областта на математическата логика и с работата си допринасяше за създаването на този AI в когото не вярваше. Много от неговите ученици сега са водещи специалисти в областта на AI.

Професор Скордев обожаваше строгите математически доказателства. Той се занимаваше и със системи за автоматично доказване на теореми. Той допускаше възможност един ден компютрите да доказват теореми по-успешно от хората, но дълбоко в себе си и в това не вярваше. Предполагам, че когато компютрите започнат да доказват теореми по-добре от нас хората, ние математиците ще сме силно разоравани. Професор Скордев си спести това разочарование, защото не дочака това да се случи.

References

Dobrev, D. (2019). AI Should Not Be an Open Source Project. *International Journal "Information Content and Processing", Volume 6, Number 1, 2019, pp. 34-48.*