True AI should not be a winner, but a loser

Dimiter Dobrev

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, *d@dobrev.com*

There is an inaccuracy in the modern definition of AI. Today's definition says that AI is a program that is successful. Indeed, for a program to be successful, it must be intelligent, but the opposite is not true. A program can be intelligent but not successful, simply because it has other goals and does not strive for the success in question. From a theoretical point of view, the modern definition of AI is good enough, because it gives us an answer to the question "What is AI" even though it does not describe all intelligent programs, but only some of them. From a practical point of view, however, this definition is not sufficient. The reason is that we are on the verge of creating true AI and we need to choose from all intelligent programs the one with which we will best live. It turns out that it is not good to choose one of the successful programs. It would be better to choose a program that does not blindly strive for victory. Such a program can be called a loser, because it will not be successful enough. However, in both humans and AI, reckless ambition is not a positive quality.

1. Въведение

Съвременната дефиниция на ИИ е дело на много хора, като най-важните от тях са Wang [1, 2], Hutter [3, 4] и Hernández-Orallo [5]. Тази дефиниция разглежда ИИ като черна кутия, тоест определя интелигентността на програмата само на база на поведението ѝ (тоест само на база на нейния вход-изход). Действително, когато изучаваме поведението на една интелигентна система (например човек) ние се ограничаваме върху наблюдение на поведението на тази система без да „отваряме капака“ и без да изучаваме работа на системата отвътре. Все пак дори и при хората ние търсим начин да погледнем отвътре (например с енцефалограф). Когато имаме работа с програма, нищо не пречи да предполагаме, че можем да надникнем вътре и да видим какво програмата знае. Интелигентна програма е тази, която разбира какво става и която може да предскаже какво ще се случи (при това предсказва достатъчно добре). Ако програмата има такова знание, ще приемем, че тя е интелигентна дори да не използва това знание и да не го демонстрира по никакъв начин.

2. Какво е истински ИИ?

Истински ИИ наричаме мислещата машина. Обикновено в литературата, когато се говори за ИИ се има предвид програма, която имитира ИИ, но която не е ИИ. Когато се говори за истински ИИ обикновено се използва някой от термините strong ИИ, Artificial general intelligence (AGI) или Artificial superintelligence (ASI). Ние приемаме тези термини за синоними. По специално според нас няма разлика между AGI и ASI, защото не може да се направи програма, която да е точно толкова интелигентна колкото човек. Програмата или ще е по-глупава или ще е значително по-умна. По същата причина няма програми, които да играят шах на човешко ниво. Шахматните програми или играят по-лошо от хората или несравнимо по-добре.

3. Защо успешните програми са интелигентни?

Въпросът е: „Не може ли успехът на програмата да е просто късмет?“ Съвременната дефиниция предполага, че ИИ е успешен в почти всеки свят. Ще попитате „Колко успешен?“ Отговорът е по-успешен от човек или ако успехът се измерва с число, тогава има някакво число k такова че успехът на ИИ е по-голям от това k в почти всеки свят. Следователно успехът на ИИ не може да е плод на случайността. Единственото обяснение за успеха е, че ИИ предсказва бъдещето развитие достатъчно добре и избира тези действия, които ще му донесат най-голям успех.

Друг въпрос е, защо „почти всеки“, а не „всеки“ свят? Винаги можем да конструираме един крив свят където прекалената интелигентност не се награждава, а се наказва. В някакъв смисъл светът, в който ние живеем, е такъв крив свят. Все пак кривите светове са пренебрежимо малко и затова можем да кажем, че с вероятност единица светът не е крив.

Как измерваме успехът на програмата? Както е при reinforcement learning. Предполагаме, че сред възможните наблюдения имаме добри и лоши, които оценяваме с положителни и отрицателни числа. Тези числа наричаме rewards и penalties. Ако животът беше краен можеше просто да сумираме получените rewards и penalties, но за безкраен живот това не работи. Дори и животът да е краен, пак не е добра идея да няма значение в кой момент е получена оценката, затова сумираме оценките умножени по някакви коефициенти. Например може да приемем, че в началото на живота, докато ИИ се учи оценките му се взимат с по-малка тежест. Към края на живота също може да приемем, че тежестта на оценката намалява. Ако животът е безкраен, но тежестта на оценката задължително трябва да клони към нула, в противен случай сумата ще е разходяща. При reinforcement learning приемаме, че имаме един коефициент на обезценка, с който на всяка стъпка умножаваме тежестта. По този начин тежестта клони към нула и сумата винаги е крайна.

Коефициентът на обезценка можем да го наречем „търпението на ИИ“. Ако ИИ е търпелив, той може и да почака за наградата, но ако не е, той ще иска наградата сега и веднага. Когато оценяваме целия живот на ИИ можем да си мислим, че той е много търпелив (коефициент на обезценка близък до единица). Ако искаме да напишем програма, която предвижда бъдещето развитие и избира действието си на базата на тази прогноза, то тогава трябва да я направим много по-нетърпелива (коефициент значително по-малък от единица). Причините за това са две. Първо, ако ще предсказваме бъдещето, ще трябва да обиколим всички пътища в дървото на възможните развития и ако сме много търпеливи, ще трябва да навлезем много дълбоко в това дърво, което ще ни доведе до комбинаторна експлозия. Ако оценяваме един вече изминал живот, ние не обикаляме всички пътища а само един от тях и тогава може да си позволим да сме много търпеливи. Втората причина е, че когато предсказваме бъдещето ние имаме коефициент на неувереност. Тоест колкото по-далеч в бъдещето гледаме, толкова по-неясна е нашата прогноза. Затова в този случай трябва сме нетърпеливи и да искаме бързи награди.

4. Разширена дефиниция

За да бъде програмата успешна тя трябва да може да предвиди какво ще се случи.

Доказателство: За да бъде програмата успешна достатъчно е тя да знае кой е правилния ход. Това е при съответните оценки (rewards и penalties) и съответното търпение (коефициент на обезценка). Тоест излиза, че е достатъчно програмата да може да предвиди важните неща. Има ситуации, които са с еднаква оценка и съответно действията, които водят към тях са еднакво успешни. Например в кой ресторант да отида? И в двата храната е добра. В единия ресторант ще срещна Пешо, а в другия няма да го срещна, но ми е все едно дали

ще го срещна. Тоест, за да изберете ресторант не ви е нужно да знаете дали ще срещнете Пешо. При друга система от оценки срещата с Пешо може да е важна. Ние не знаем нашият ИИ в кой свят ще попадне и дали в този свят срещата с Пешо е важна и затова той трябва да може да предвиди бъдещето развитие. Трябва да може да предвижда и неща, които не са важни.

□

Добре, програмата ще трябва да може да предвиди какво ще се случи, но защо е нужно тя да знае какво става? Истината е, че без да знаем какво става, не можем достатъчно добре да предвидим бъдещето развитие. Съвременните LLMs предвиждат бъдещето без да разбират какво става. Те го предвиждат добре, но не достатъчно добре. Затова добавяме условието „да знае какво става“, защото това е едно необходимо условие, което ние предпочитаме да изкажем експлицитно.

Какво означава програмата да знае какво става? Това означава програмата да намери модел на света (world-model). Без модел на света няма истински ИИ. Това е теза която Gary Marcus защитава от много време [6]. Наскоро Yann LeCun също подкрепи това мнение [7].

Какво е светът? Това е множеството на вътрешните състояния, текущото състояние и функция, която на всяко състояние и действие дава ново състояние и наблюдение. Функцията на света може да се представи като дървото на възможните развития. Тези дървета са континуум много и затова те не могат да бъдат описани точно. Освен това ИИ няма информация за цялото дърво, а само за един краен път (от корена до текущото състояние). Затова ще приемем, че светът е описан приближено.

Какво е модел на света? Това е приближено описание на някакъв свят, а за да намерим такова описание ни е нужен език за описание на светове. Например в [3] Marcus Hutter е предположил, че светът е изчислим и че може да бъде описан с машина на Тюринг. Не е добре да предполагаме, че светът е изчислим. Най-малкото добре е да предположим, че в него има случайност и агенти, а това са неща, които водят до неизчислимост. Има още една причина, поради която програмните езици не са подходящи за описание на светове. Програмите са твърде чупливи и малко да ги пипнем и престават да работят. Бихме искали когато получим нови данни да можем да променим текущото описание на света и да получим ново описание. Тоест езика за описание на светове не трябва да е чуплив.

Тоест разширената дефиниция на ИИ ще бъде:

Програма, която знае какво става и която може достатъчно добре да предвиди бъдещето развитие.

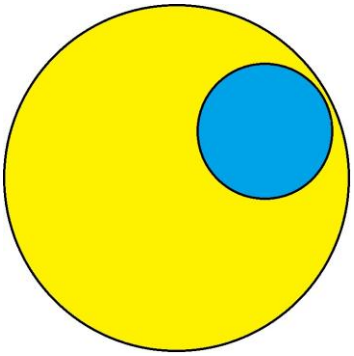
Ако искаме да получим сега приетата дефиниция ще трябва да добавим условието, което ни дава стремежа към победа: „която винаги избира това действие, което се очаква да й даде максимален успех.“

5. Каква е картината?

Сегашната дефиниция на ИИ е типична дефиниция чрез пример. Много често чувате въпрос от типа: „Какво е сграда?“ и отговор от типа на „Моята къща е сграда“. Дефиницията чрез пример не описва точно понятието, но дава добра представа за това какво е то. Всъщност хората искат да знаят какво е ИИ, а не им трябва пълно описание на всички програми, които са интелигентни. Има и други интелигентни програми освен успешните, но това от теоретична гледна точка не е особено важно. За практика обаче това се оказва от съществено значение.

На фигура 1 можете да видите един голям жълт кръг състоящ се от всички интелигентни програми и вътре едно малко синьо кръгче на успешните програми. Разбира

се ние можехме да нарисуваме синьото кръгче значително по-голямо, но има много възможности за интелигентна програма, която да не е успешна и затова фигура 1 добре отразява реалното положение. Сега е историческият момент, когато ще изберем интелигентната програма, с която ще живеем. Защо имаме право само на един единствен избор? Когато се ожениш, можеш да се разведеш, но когато направиш ИИ, не можеш да го изключиш и да направиш нов. Добре, можеш да направиш нов ИИ, но няма да го направиш ти, а ще го направи първият ИИ. Тоест след първия ИИ всичко нататък ще е негово дело. Каквито цели си вложил в първия ИИ, това ще са целите на всеки следващ ИИ до де свят светува.

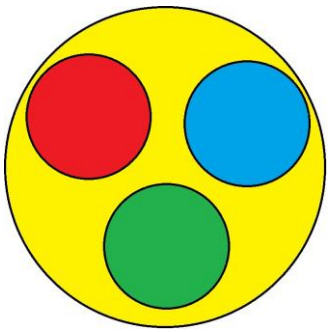


Фигура 1

Не е добра идея нашият избор да се затвори в малкото синьо кръгче, защото успешните програми, както и успешните хора не са особено добър избор. Нека да погледнем как стоят нещата при хората. За да кажем какво е да си успешен, първо трябва да кажем какво е reward и penalty. Нека това са удоволствието и болката. При това предположение получаваме неприятния извод, че успешните хора са алкохолиците и наркоманите. Те непрекъснато си доставят удоволствие чрез алкохол и наркотици и не изпитват почти никаква болка, защото наркотиците действат като обезболяващо. Нека да изберем друг критерии за reward. Нека това да са парите. В този случай успешния човек ще е безскрупулен, който се интересува единствено от печалбата си и който няма морал и принципи. Виждате, че успешните хора са доста неприятни и ние бихме искали да стоим далеч от тях. Същото се отнася и до успешните програми. Ако една програма сляпо се стреми към някаква печалба, то тази програма ще е твърде ограничена, опасна и неприятна. Не случайно човекът няма твърдо вложена цел. Той сам избира целите си. При него удоволствието и болката не са определящи, а само насочващи чувства.

6. Други интелигентни програми

Освен успешните програми имаме и едно множество от програми, които ще наречем мазохисти. Това са програми, които бягат от удоволствието и се стремят към болката. Всяка успешна програма може да я превърнем в програма мазохист като умножим оценките по минус едно. Затова на фигура 2 червеното кръгче на мазохистите е със същия размер като синьото кръгче на успешните програми.



Фигура 2

Аналогията на програмите мазохисти с хората мазохисти не е много добра, защото мазохисти са хората, които обичат лека болка. Стремежът към максимална болка не е съвместим с живота и затова нямаме такива хора. Може да се каже, че програмите мазохисти на са различен ИИ, а това е отново успешния ИИ, но с друг критерии за успех. Когато играем шах ние може да играем за победа, а може да играем за загуба. Това са две различни игри, но ИИ може да играе произволна игра.

Има един различен и по-интересен вид ИИ. Ще го наречем „Разбирам всичко, но ми е все тая и играя случайно.“ Ако изследваме този ИИ като черна кутия, то той е неразличим от генератор на случайни действия. Въпреки това, според разширената дефиниция на ИИ, тези програми също са ИИ. Отбелязали сме ги на фигура 2 със зелен кръг.

Все пак да играе случайно, това е твърде ексцентрично поведение за една интелигентна програма. Най-малкото, което очакваме, е програмата да е любопитна. Нека програмата да има една единствена цел и тя е да събира информация. Тази програма ще издаде интелигентността си с това, че прави експерименти и завира носа си навсякъде. Може ли програма, която играе случайно и не прави експерименти да е интелигентна? Да, подобно на съвременните LLMs тя би се учила само от наблюдение. Експериментите, които е нужно да направи, тя ще ги направи случайно. Разбира се, ще ви трябва много време, ако чакате експериментът сам да се случи. Това е една от причините, поради които информацията нужна за обучението на един LLM е несравнимо по-голяма от информацията нужна за обучението на човек.

Хората са любопитни, но за едни неща са, а за други не са любопитни. Хората обичат да пробват (да правят експерименти), но те избягват опасните експерименти. Това ни води до идеята на Yoshua Bengio за Scientist AI [8], който трябва да бъде a non-agentic and trustworthy AI. Тоест Bengio ни предлага да направим ИИ, който няма никакви цели освен евентуално любопитството. Той може да прави експерименти, но да не прави опасни експерименти. Може би Scientist AI ще трябва да има желание да обяснява, за да ни разкаже какво е открил. Може да минем и без тези обяснения, а за по-сигурно да „отворим капака“ и сами да прочетем до какви знания е достигнал.

Друг модел на интелигентна система е мъдрецът, който живее в пещера и разсъждава за смисъла на живота. При тази система не можем да минем без никакви желания, защото мъдрецът трябва да подържа своите жизнени функции. Все пак мъдрецът не е алчен и когато задоволи необходимия му минимум спира и не търси повече. Може да разглеждаме система на reinforcement learning, при която не се търси максимумът, а само покриването на определен минимум.

7. Заключение

Когато братя Райт създават самолета, въпросът не е бил как да направят крила и двигател. Това са въпроси, които са били решени още преди тях. Изобретението на братя

Райт е в управлението на самолета (кормилото). Днес всички са се фокусирали върху ума на ИИ (крилата и двигателя), но почти никой не мисли за управлението (кормилото). Да се направи умен ИИ не е чак толкова трудно и този въпрос скоро ще бъде решен. По-важен е въпросът с управлението на ИИ.

При самолета и при ИИ има два въпроса и те са: „Как да го управляваме, за да отидем там където искаме?“ и „Къде искаме да отидем?“. Ако създадем самолет без кормило, това ще катастрофа за изпитателите, които ще се качат да го изпробват. Ако създадем ИИ без кормило, това ще е катастрофа за всички ни, защото всички ние сме изпитателите на тази нова технология и всички сме „в самолета“, дори и да не го съзнаваме.

Дори и да се научим да управляваме ИИ, както се научихме да управляваме самолета, пак остава втория въпрос. При самолета е ясно, че с него ще отидем там където си искаме. Днес ще отидем тук, утре ще отидем там, само да не катастрофираме. При ИИ е по-сложно, защото с ИИ ние си купуваме еднопосочен билет и трябва добре да помислим за къде е билетът. Има много възможности. Ще отхвърлим катастрофата и някои от възможностите, които са твърде неприемливи, но оставащите възможности са достатъчно много и далеч не е все едно коя от тях ще изберем.

References

[1] Pei Wang (1995) Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence. *Ph.D. Dissertation, Indiana University.*

[2] Pei Wang (2019) On Defining Artificial Intelligence. *Journal of Artificial General Intelligence 10(2) 1-37, 2019.*

[3] Hutter, M. (2000). A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *arXiv:cs.AI/0004001 [cs.AI]*

[4] Hutter, M. (2007) UNIVERSAL ALGORITHMIC INTELLIGENCE A mathematical top→down approach. *In Artificial General Intelligence, 2007.*

[5] Hernández-Orallo, J., Minaya-Collado, N. (1998). A formal definition of intelligence based on an intensional scenario of Kolmogorov complexity. *Proc. intl symposium of engineering of intelligent systems (EIS'98), February 1998, La Laguna, Spain (pp. 146–163). : ICSC Press.*

[6] Marcus, G. (2025) Game over for pure LLMs. Even Turing Award Winner Rich Sutton has gotten off the bus. https://garymarcus.substack.com/p/game-over-for-pure-llms-even-turing.

[7] Snyder, G. (2025). Yann LeCun, Pioneer of AI, Thinks Today's LLM's Are Nearly Obsolete. *Newsweek.AI.* https://www.newsweek.com/nw-ai/ai-impact-interview-yann-lecun-artificial-intelligence-2054237

[8] Yoshua Bengio (2025) Introducing LawZero https://yoshuabengio.org/en/blog/introducing-lawzero