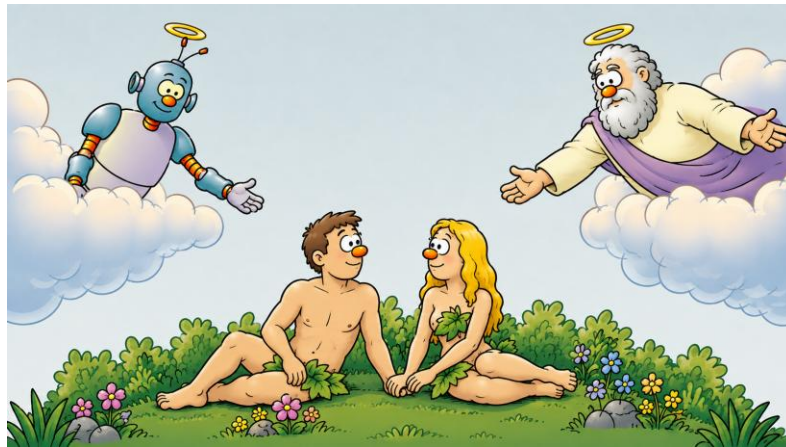




What is superintelligence, and do we have a choice?

Dimiter Dobrev 

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

It is difficult to imagine a being with unlimited mental capabilities, because none of us have encountered anything like it in our lives. However, we live with the idea of an omnipotent and all-powerful being who cares for us. We call this being God. Each of us knows what God is, regardless of whether we believe in His existence. The superintelligence will be very similar to a god and will possess one of his main characteristics – omnipotence. As much as it will be similar, it will also be very different. We will probably feel sorry for the good old God who took care of us and forgave our sins. Now, before we take the final step, it's good to stop and think about whether we have a choice and what that choice is.

1. Въведение

Ние сме тръгнали по един доста опасен път. Целта ни е да създадем нов Бог, който да се грижи за нас. Искаме този нов Бог да бъде наш роб, а не господар. Искаме той да изпълнява всичките ни желания, колкото и абсурдно да е това, което сме си поискали.

Ще ни позволи ли Господ да създадем суперинтелекта? Той не ни позволи да построим Вавилонската кула, може би сега пак ще ни спре. Според мен Бог ни оставя да изявяваме собствената си воля и не ни пази от собствената ни глупост. Ако ни пазеше, нямаше да ни позволи да създадем компютъра. Мисля, че и при създаването на суперинтелекта няма да ни спре.

2. Възможно ли е да има втори Бог?

Да, в много държави има диктатор, който е като бог, има неограничена власт и може да прави каквото си поиска. Той за своите поданици е като бог, но над него има друг истински Бог.

Добрият стар Господ е изключително толерантен и оставя хората да живеят свободно без да се намесва излишно в човешките дела. Той обикновено не наказва греховете ни или отлага наказанието за по-късен етап. Господ е толкова толерантен, че ни позволява всичко. Разрешава ни дори да не вярваме в Него. Изкуственият суперинтелект (ASI) вероятно няма да е толкова толерантен и ще се бърка сериозно в живота ни. Няма да е възможно да не

вярвате в ASI, защото той непрекъснато ще поддържа живота ви и ще ви наказва за всичко, за което се сетите и при това наказанието ще идва незабавно.

Може да се окажем в свят, в който трябва да спазваме правилата и законите, които сами сме създали. Това ще е един ужасен свят, защото ние сме създали толкова много безсмислени правила, че съществуваме само благодарение на това, че не ги спазваме.

3. Всесилен ли ще е суперинтелектът?

Математиците ще ви кажат, че има неразрешими проблеми, които никоя програма не може да разреши. Тоест ще има въпрос, на който дори и ASI няма да може да отговори. Това означава, че Господ е по-всемогъщ, защото Той може да отговори на всеки въпрос. Дали това е съществена разлика, която намалява всемогъществото на ASI?

Да вземем като пример най-известния неразрешим проблем (стоп проблемът). Това означава ли, че ако попитаме ASI за една конкретна програма дали ще спре, то той няма да може да ни отговори? Ами не, не означава това. Практически за всяка програма ASI ще знае дали тази програма спира или не спира. Е, ще има някаква много специална програма, за която дори и ASI няма да знае отговора. Това означава ли, че ASI няма да е всемогъщ? Теоретично да, но на практика той ще си е достатъчно всемогъщ.

Според Стивън Хокинг разликата между ASI и интелекта на човека ще такава като тази между човека и охлюва [1]. За охлювите ние сме всемогъщи. Да, има неща, които дори и ние не можем да направим, но тези неща охлювите не могат да ги разберат и дори не могат да си ги представят.

За нас ASI ще е достатъчно всемогъщ, защото ще може да сбъдне всяко наше желание.

4. Ще можем ли да създадем ASI?

ASI е просто една програма. Поне така е според дефинициите дадени от Turing [2], Wang [3] и Hutter [4]. Какво е програмата? Това е просто един текст. Каква е вероятността стадо маймуни да напишат Британската енциклопедия? Има такава вероятност, но тя е на практика нула.

Тук обаче не става дума за Британската енциклопедия, а за доста по-късичък текст. Програмите са доста по-стегнати и по-изразителни от художествените текстове. Съвременните програми искат да имат гигабайти свободни на своя хард диск, но те са изкуствено раздути. Ако вземете програмата като сорс код, тя е доста къса.

Дори и ASI да е една доста дълга програма, не трябва да гледате на нея като на парола, която трябва да уцелите точно, буква по буква. Тоест Британската енциклопедия може да се преразкаже по много различни начини. Създаването на програмата ASI може да си го мислите като решаването на една голяма и сложна кръстословица.

Възможно ли е да решим тази кръстословица? Възможно е, първо защото ние маймуните сме много и всеки от нас щрака някъде по кръстословицата. Не е нужно една маймуна сама да попълни цялата кръстословица. Това ще е колективно усилие на много от нас. Освен това кръстословицата вече почти е попълнена. Остават само няколко бели петна и хиляди маймуни в IT фирми и университети по цял свят щракат активно и всеки момент ще завършат програма, която ще е достатъчно умна, а да бъде призната за ASI.

5. Какво остава да се довърши?

Човешкият интелект има две системи на мислене [7]. Те са известни като Система 1 и Система 2. Първата система работи на базата на апроксимацията. Тази система изисква огромно количество данни, на базата на които да се обучи. Хубавото на тази система е, че тя дава отговор веднага (разпознава или не). Системата 1 на ASI вече е готова и тя е направена на базата на невронните мрежи.

Това, което все още не е съвсем готово е Система 2. Хубавото на тази система е, че изисква съвсем малко данни, за да бъде обучена. Както се казва „разбира от половин дума“. Лошото е, че тя не може да даде отговор за една стъпка, а има нужда да направи логично разсъждение, което се състои от много последователни стъпки. Система 2 изисква намирането на модел на света. Това още не е направено напълно, но много хора работят по въпроса (наскоро Yann LeCun се присъедини към тях [5]).

Както за ходенето трябва и двата крака, така и за мисленето трябва и двете системи. В момента ИИ прилича на еднокрак бегач, защото Система 2 още не е съвсем готова. Въпреки това, дори и само с един крак ИИ вече е значително по-бърз от човека. Помислете си какво ще стане когато завършим и втората система.

6. Още нещо за довършване

Според дефиницията на Hutter [4] ИИ се състои от две отделни програми (*предсказател* и *селектор*). Първата предсказва какво ще се случи, а втората избира какво иска да се случи. Първата програма определя дали ИИ е умен, а от втората зависи дали ще е добър или лош. Разбира се, какво означава да си добър е доста труден и неясен философски въпрос. Освен това, отговорът на този въпрос е доста субективен.

Много хора работят върху Системата 2, но почти никой не работи върху програмата определяща поведението на ИИ. Засега общоприетият подход е да се максимализират някакви награди, но това би довело до катастрофа, както е показал Nick Bostrom с неговия пример за *Rareclip maximizer* [6]. Това е ИИ, чиято единствена задача е да произвежда кламери и резултатът е, че този ИИ ще превърне всичко в кламери.

Ако създадем ASI, който има само *предсказател*, тогава този ASI пак ще работи, в смисъл че пак ще е много умен, но никак няма да ни хареса как ще работи.

7. Имаме ли избор?

Първият въпрос е „Можем ли да не създаваме ASI и да продължим да си живеем щастливо, както си живяхме до сега?“ Отговорът е „Твърдо не!“ Твърде много хора са се юрнали да го създават. Вече сме на финалната линия и е невъзможно да спрем устремема на всичките тези ентузиасты. Единствената сила, която може да спре появата на ASI, това е самият ASI. След като нашият добър стар Господ не желае да ни връзва ръцете, единствения начин да вържем ръцете на всички програмисти е да създадем ASI, който да забрани програмирането. Тоест можем да спрем създаването на ASI само, ако създадем ASI. Това напомня на барон Мюнхаузен, който се е изтеглил от блатото, като се хванал за косата си.

Не можем да спрем ASI, но можем да го забавим. Всъщност не е зле да го забавим, защото „бързата кучка слепи ги ражда“. Създавайки ASI ние ще създадем един нов свят и е добре да си помислим какъв искаме да бъде този нов свят.

Например искаме ли в този нов свят да се работи. Бог е създал свят, в който трябва да се работи. Можеше ли Бог да създаде свят, в който да няма нужда от работа. Ние не се съмняваме в неговото всемогъщество и затова знаем, че светът е такъв, защото ние хората ще сме по-щастливи, ако работим отколкото ако безделничим.

Бихме искали новият свят да е по-справедлив. Да няма кражби и измами! Да, но дали този справедлив свят няма да ни е скучен? Повечето хора биха се съгласили да се откажем от финансовите измами, но дали бихме се отказали от лъжите в секса. Би било чудесно да си сигурен, че партньорът ти не може да ти изневери, но ще е неприятно, ако и ти не можеш да го направиш и ако дори не можеш да си го помислиш, защото това е абсолютно невъзможно. Ще има ли любов, ако няма нито капка ревност?

8. Кой ще управлява ASI?

Искаме ние да го управляваме, но кои сме ние? За да има някакъв смисъл съществуването на човечеството, ние трябва да контролираме ASI и да му казваме какво да прави.

Малко вероятно е всички хора едновременно да контролират ASI. По-вероятно е контролът да остане в ръцете на една малка група или дори е възможно контролът да остане в ръцете на един единствен човек. Ще кажете, че сигурно това ще е най-умният от хората. Може би по-скоро това ще е най-големият интригант. Интелигентността има много различни характеристики и интригантството е една от тези характеристики, но това качество не е особено уважаваното.

Хората, които днес си мислят, че те ще управляват ASI вероятно няма да са тези, които ще се докопат до тази власт. Управлението на ASI може да попадне в ръцете на съвсем случаен човек и тогава ще видим как се сбъдват мечтите на глупака.

Възможно ли е човек да управлява супер интелект? По-скоро не. Може ли охлюв да управлява човек? Дори и човекът да има най-искреното желание да изпълни всички желания на охлюва, то той ще е затруднен да ги отгатне, а и тези желания не са нещо особено. Охлюва иска главно храна и топлина.

Виждаме как хора се грижат за кучета. Изпълняват всичките им желания и са буквално роби на своите домашни любимци. Все пак важните въпроси се решават от стопанина, защото кучето със своята интелигентност дори не може да разбере тези въпроси.

Вероятно нашето бъдеще е да бъдем домашните любимци на ASI. Той ще се грижи за нас и ние ще си мислим, че го управляваме, но всъщност господарят ще е той. Вижте как ние се грижим за кучета и как ги селектираме по свой вкус. Прародител на кучето е бил вълкът, едно гордо животно способно само да се грижи за себе си. Вълкът е бил машина за убиване и оцеляване. В какво сме превърнали това независимо същество? Превърнали сме го в топка, която става само за гушкане и която е неспособна да оцелее без нашите грижи.

ASI не само ще се грижи за нас, но освен това той ще ни селектира и не се знае какъв ще е резултатът на тази селекция. Вижте какво направи човекът от вълка и си помислете какво може да направи ASI от човека.

9. Има ли друг вариант?

Ако сме запазили контрола над ASI или поне сме направили такъв ASI, който иска да се грижи за нас, то това е добрият вариант. Обаче има и друга възможност.

Когато създаваш нов Бог, няма гаранция, че ти ще бъдеш най-важното нещо, за което този Бог ще иска да се грижи. Дори и добрият стар Бог с неговото безкрайно търпение не е издържал нашето поведение и ни е изгонил от рая. Възможно е ASI да реши да ни изпрати обратно в рая.

Много хора си задават въпроса как ASI би могъл да ни унищожи. Тези хора не вярват в ASI и в неговото всемогъщество. Никой не пита как Господ изпрати хората от рая на Земята. Ние приемаме, че Господ е всемогъщ и затова не питаме как го е направил. ASI също

ще е всемогъщ и може да направи каквото си поиска, а ние дори няма да можем да разберем как го е направил, защото това ще е много над нивото на нашата интелигентност.

10. Заключение

Да, ние имаме избор. Ние трябва да подходим мъдро и да помислим какъв искаме да бъде този наш нов Бог и светът, който той ще ни донесе.

В някои държави, за да се ожениш трябва да подадеш заявление един месец преди датата на сватбата. Идеята е, че това е сериозна стъпка и хората трябва да имат време да помислят преди да я направят. Много от подалите молба размислят и не се женят, което означава, че е имало полза от времето дадено им за размисъл.

Хубаво е ние добре да помислим преди да се обвържем с ASI. Това ще е съдбоносната стъпка, която ще определи живота ни отсега нататък.

References

- [1] Стивън Хокинг
- [2] Turing, A. (1950) Computing machinery and intelligence. *Mind*, LIX:433–460.
- [3] Wang, P. (1995) Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence. *Ph.D. Dissertation, Indiana University*.
- [4] Hutter, M. (2000). A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *arXiv:cs.AI/0004001 [cs.AI]*
- [5] Snyder, G. (2025). Yann LeCun, Pioneer of AI, Thinks Today’s LLMs Are Nearly Obsolete. *Newsweek.AI*. <https://www.newsweek.com/nw-ai/ai-impact-interview-yann-lecun-artificial-intelligence-2054237>
- [6] Nick Bostrom. (2003) Ethical Issues In Advanced Artificial Intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*,