How Can We Make AI with a Nice Character?

How Can We Ensure That AI is a Nice Guy?

Dimiter Dobrev[1], Lyubomir Ivanov[1], George Popov[2], Vladimir Tzanov[3]

[1] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, d@dobrev.com, lyubomail@yahoo.com
[2] Faculty of Computer Systems and Technologies, Technical University of Sofia, popovg@tu-sofia.bg
[3] Independent researcher

God created man in His own image, the Bible said millennia ago. Today we are headed to creating Artificial Intelligence (AI) in *our* own image. The difference however is that God created a feeble and vulnerable being for which to take care of, while we are trying to create an almighty being who will be incomparably smarter than us and will take care of us. Thus, we are aiming to create our new god, and it matters a lot what kind of character the new god will be – kind and compassionate, or terribly stringent and overly demanding on us. Every human being has a character. Similarly, AI will have its own character. We will consider AI as a program with parameters which determine its character. The aim is to use these parameters in order to define the kind of character we want AI to have.

Keywords: Artificial General Intelligence, Nice Character, Nice Guy.

Introduction

When creating natural intelligence, we are not aiming to create a person with a nice character. Instead, go by the commercial principle *telle quelle* (as-is, whatever comes up). Of course there are so many people and everyone has his or her unique character. There are very nice as well as very nasty people. Even brothers who grew up in the same family can have completely different characters.

People are different and they have to be different because nature never puts all of its eggs in a single basket. In some worlds the courageous ones prevail while in other worlds you had better stay on the safe side. If people were all the same, they would all perish in a world which is not right for them. Thanks to people being different, some part of the population always survives and continues the genus.

We assume that there is one and only one real world, but depending on where and when you are born you may find yourself in a very different world. Natural intelligence has no idea where and when it will be born, so it must be prepared to survive in any kind of world.

Things with AI will be different because we will not have multiple different AIs, but just a single one (see [2]). Furthermore, once created by us, AI will have a character of its own and that character, be it nice or nasty, will be there forever because we probably will not have an opportunity to change it. Moreover, unlike humans AI is immortal and we cannot hope that one day it will go away and another AI with a more benign character will take its place. Accordingly, we must be very responsible when creating AI rather than go by the *telle quelle* principle.

We mentioned that in creating people we act quite irresponsibly. In fact this is not very much the case. Before making a child we carefully choose the partner with whom we will make it. The rationale is that the child will be very much akin to our partner and by choosing the partner we basically shape our child. We can even create designer babies by choosing from several embryos the one whose genes we like best. This is usually done to avoid congenital diseases. I have not heard of anyone browsing through embryos with the aim to find a child with a nice character. Essentially, do we truly want the character of our child to be nice? As parents, we would be more happy to have a nice child, but the child itself might be better off if it is nasty. Maybe in our world a person with a nasty character has higher odds of surviving. So if we parents put our child first we might prefer to have a nasty child.

We already said that in creating AI we must be highly responsible. However, at this very crucial moment in human history we are utterly irresponsible as we blindly rush to make AI without caring about the consequences. Right now more than 200 companies are in a reckless race to be the first to create AI. The aim of this race is to make money, and this is an extremely meaningless aim.

AI is a magic wand that can make any wish come true. Money is also sort of a magic wand and can grant many wishes. Let us say AI is the golden magic wand and money is a silver wand. It is stupid to create a golden wand and trade it for a silver one. If you have AI, why would you need money at all?

This paper is written by a several authors. The text was started by the first author and the others joined in to improve what has been written and support the basic idea that Artificial General Intelligence (AGI) is a dangerous thing which warrants the highest caution.

What is AI?

All references to AI in this paper are references to AGI.

According to [1] AI is a program which is sufficiently smart. A program is sufficiently smart if it is smarter than a human being. The smarter between two intellects is the one which in any world performs at least as well as the other one. Certainly, we can always construct a special world in which is the opposite (the second one performs better than the first one), but if in almost all worlds the first one performs at least as well as the second one, then the first intellect is smarter than the second.

Here we have an important specificity. In [1] it is assumed that we have a clear criterion by which we can judge whether a given program performs better than another program. We assume that we have two signals (two observations). Let these observations be *win* and *loss*. The goal is to achieve more wins and less losses. Similarly, we can assume that there are two buttons, a green button and a red button, wherein AI’s goal is that we praise it by pushing the green button more often and the red button less often.

It would be extremely stupid if we created AI with these buttons because very soon AI will learn to press the green button itself. This is the better case. The worse case would be if AI manages to make us its slaves, have us keep pressing the green button all the time, and punish us heavily if we press the red button by mistake.

AI that pushes its own green button would be like a drug addict who derives pleasure by constantly stuffing himself with drugs. We hate the thought of AI that behaves like a drug addict.

We humans do not have a clear criterion to judge if a given life is better than another. Instead, we have instincts and a character which determine our behavior. Our evolutionary criterion is clear, and it is to *survive and reproduce*. However, this principle is not embodied in natural intelligence. Instead, we have instincts that indirectly work for this principle. Examples of such instincts are fear of heights and love of children. Another example is the feeling of pain and the feeling of pleasure, which we instinctively perceive as negative and positive feelings. All these feelings are only indications rather than firm criteria of success. We are ready to endure a lot of pain and give up many indulgences if we believe this is for the sake of a greater goal.

We do not have a clear criterion by which we can distinguish good from bad. This is the reason why many of us cannot find the meaning of life although we are constantly searching for it. The evolutionary criterion can never be incorporated in natural intelligence because it depends on the future, and no one is able to predict the future that accurately. No programmer is able to write a program that says which action will give the individual or the population the best chance of survival. A programmer cannot, and indeed even nature cannot create intelligence that can depict the future so clearly, and because of this the goal of humans is determined indirectly.

If we are successful in making AI that is capable of predicting the future with absolute accuracy, that would be errorless intelligence. We will assume that errorless intelligence cannot exist. Even if some errorless intelligence existed, it would be very boring because of the assumption that there is always a single most correct solution and such intelligence always knows what that solution is. The unknown is what makes life interesting. Wondering about the right action is more amusing than knowing exactly what the right action is.

Now that we gave up the idea of creating AI with a hard criterion for success (green and red button), we will have to rely on AI’s instincts and character to indirectly determine its goal. The kind of instincts and character we embed in AI are extremely important because they will shape the near future in which we will have to coexist with AI.

We humans have been the dominant species on planet Earth. Now we are about to relinquish that role by creating the new dominant species which will oust us from our dominant position. If AI will be driven by instincts and character, it will be an independent being that will search for the meaning of life on its own and nobody knows where exactly it will find it.

В [7] Pei Wang разглежда ИИ, който има много цели, които ИИ може сам да променя. Това може да са междинни цели, които водят към главната цел, но Pei Wang предполага, че дори и главната цел може да бъде променена. Тоест идеята за променяща се главна цел не е нова. Според Pei Wang, за да бъде една система интелигентна, тя трябва сама да си избира целите.

Възможен ли е ИИ?

Ще започнем с думите на китайския философ Zhuang Zhou допълнени от френския математик René Thom (това е мотото на книгата []):

There once lived a man who learned how to slay dragons and gave all he possessed to mastering the art. After three years he was fully prepared but, alas, he found no opportunity to practice his skills. As a result he began to teach how to slay dragons.

Възможен ли е ИИ или само ни плашат с него, както плашат децата с чичко Торбалан?

Мнозинството е абсолютно убедено, че машините не могат да мислят и никога няма да могат. Те смятат, че мисленето е привилегия, която само хората имат. Възможно ли е да бъде създадена машина, която мисли като човек, но която е несравнимо по-умна от хората?

Нека да не спорим по този въпрос. Нека приемем, че 99 на сто ИИ е невъзможен, но все пак съществува 1% възможност за обратното и да поразсъждаваме при хипотезата „Ами, к‘во ще правим, ако е възможен?“ Когато става дума за съдбата на човечеството, струва си да отделим време и да поразсъждаваме върху тази хипотеза, дори и да става дума за нищожна вероятност нашите страхове да се сбъднат.

Аз самият съм част от малцинството, които вярваме, че ИИ е възможен. Това, че вярвам в ИИ не означава, че вярвам във всичко. Например, аз не вярвам с извънземни и гледам с насмешка на хората, които вярват, че виждат летящи чинии. Затова разбирам хората, които не вярват в ИИ и гледат с насмешка на такива като мен.

Въпросът „Какво е ИИ?“ е важен, както и въпросът „Какво е дракон?“ и „Какво е призрак?“. Ако не вярвате в ИИ, в дракони и в призраци, то тези въпроси са безсмислени, но можем да си зададем въпроса „Как се воюва с дракон?“ дори и без да сме съвсем наясно с това какво е дракон. Нека приемем, че ИИ е машина несравнимо по-умна от човека. Друг е въпросът, може ли такава машина да бъде направена. Може и да не може, ама има толкова много неща, които си ги мислехме за невъзможни, а те се оказаха напълно възможни.

Теорията на несъществуващия обект е абсолютно безсмислена. (Ние като математици знаем, че ако един обект не съществува, то за него може да се каже всичко и то ще е вярно. От несъществуването на обекта следва, че той има всички свойства.) Да, ама ако съществува или ако е възможен, то тази теория не е безсмислена.

Можем ли да го управляваме?

Много малко са хората, които вярват, че ИИ е възможен, но почти никой не врява, че ИИ може да бъде управляван. Например според Radoslav Pavlov [] ИИ е възможен, но той е нещо като природно явление, което не можем да управляваме и насочваме. Пример за такова природно явление е ураганът. Ние донякъде успяваме да предскажем откъде ще мине ураганът, но не можем да променим посоката му и да го насочим към по-безлюдна местност.

Нека и по този въпрос да не спорим. Нека приемем, че 99 на сто ИИ не може да бъде управляван, но че съществува макар и малка вероятност това да не е така. Нека да заложим на тази малка вероятност и да помислим как бихме могли да управляваме ИИ и да насочим неговия характер към това, което е изгодно за нас.

Дори и при ураганите ние се опитваме да ги управляваме и насочваме. Разбира се, ако се научим да управляваме ураганите, ние ще трябва да пазим в тайна това наше умение, защото накъдето и да насочим урагана, все някой ще пострада и ще ни обвини, че ние сме виновни за това.

Разумно е този, който се научи да управлява ИИ да пази в тайна това свое умение, за да не предизвика недоволството на пострадалите от лошия характер на ИИ. От друга страна този, който може да управлява ИИ, ще е достатъчно силен, за да не се притеснява от недоволството на останалите.

Какви ще са последствията?

Всички са съгласни, че появата на ИИ ще бъде голямо премеждие за човечеството. Това изпитание може да бъде сравнено със сърдечна операция, но операцията няма да бъде само върху един човек, а върху всички нас едновременно.

Не е все едно как ще преминем през това приключение, защото не е все едно какъв ИИ ще създадем. Както и при сърдечната операция, има много рискове. Например, може да не се събудим след упойката. Все пак, нека да не мислим за най-лошото, а да разгледаме положителния сценарии.

Когото човек се готви за сърдечна операция първият въпрос е да дали може да се мине без тази операция. В случая не може, защото създаването на ИИ е неизбежно. Тогава как да преминем през това? Не е все едно каква болница ще изберем, кой ще е докторът, който ще ни оперира, каква ще е новата клапа, която ще ни сложат в сърцето. Друг въпрос е дали искаме операцията да е планова или ще се оперираме по спешност. Плановата операция е за предпочитане, защото ще си направим необходимите изследвания, ще подготвим тялото си и ще си напишем завещанието. Когато операцията е по спешност, тогава нещата просто се случват и са извън нашия контрол.

Някой трябва да подготви човечеството за очакващата го сърдечна операция и целта на тази статия е да събере хора, които да повдигнат този въпрос.

DNA

Saying that AI is a program is not quite accurate because a program is simply a piece of text (sequence of bytes) while we perceive AI as a living being. For a program to rise from text to a living being it must be started on some computer.

We can draw an analogy with Man and say that human DNA corresponds to AI’s program. DNA *per se* is not a living being. Only when inserted in an ovum DNA will create a fetus that will come into life. Similarly, AI will come into life only when we start it on a computer.

Both people and AI need training in order to become the aware creature which we are discussing here. The training of Man is everything that has happened in his life (his history) from the very conception to the present moment. Accordingly, AI’s training is its history since starting of the program until the present moment.

In either case the learning path as such is not important. What matters is the final result. In other words, in the case of humans training is the set of memories and knowledge that reside in our mind. In the case of AI we can assume that training is the program’s current status (the content of variables, arrays, files, etc.)

Therefore, in our mind AI is a program, a computer that runs the program and the training (the program's current status).

Training

In humans, DNA is not everything. Apart from DNA, there is training and upbringing that determine the individual's behavior. The DNA of a newborn infant plays only a limited role. More important are the education, religion and philosophy we would equip that child with. Evolution is not just a competition among DNAs, it is rather a competition between different religions and philosophies.

As we said, AI is a program and we can liken this program to the DNA of a human being. This program will evolve by teaching and training. The difference is that each child or adult have to be taught individually, while AI can only be taught once and then all of its learning can be transferred to another AI just the way you copy a file. Another difference is that wrong learning in humans is irreversible, while in AI one can erase the teaching given so far and start the process anew.

We cannot teach and educate AI if it does not have the appropriate instincts. For example, the desire to imitate is an instinct. Then AI needs another instinct which guides AI to recognize its teacher. You know about the young duckling that takes as its mother the first creature it comes across.

Children do what their parents tell them to do until they grow up and become smarter than them. AI will become smarter than us in the matter of ten minutes. Does that mean it will immediately emancipate itself and stop doing what we tell it to do?

This takes us to the first character trait that is important for AI – childishness! This is very irritating in people because every human is expected to emancipate and start taking his own decisions. However, we want AI to never emancipate and continue doing what we tell it to do forever.

It is not very clear how we can program this in code. I.e. how can we insert childishness in the AI program? In fact this holds true for almost all other character traits – we are unable to describe how they can be implemented in software code. All we can say is that childishness must be added but we do not know how to do it.

What is weak AI?

Weak AI is imitation of AI.

We consider AI as an artificial human being, and weak AI as an artificial parrot. Understanding is what makes the difference between the two. We have already made tremendous progress with weak AI, and we all need to add now is one more step: make AI understand. This step will inevitably be made, and it will be made very soon.

Ние вече разполагаме с Chat GPT и това е програма, която успешно имитира ИИ, но на тази програма й липсва въпросното разбиране. Поради тази причина Chat GPT прилича на ИИ, но не е ИИ. Представете си, че имате един много хубав автомобил. Вътре имате кожен салон, стерео уредба и мощен двигател. Имате всичко и ви липсва единствено скоростната кутия, която да свърже двигателя с колелата. Без скоростна кутия автомобилът не може да потегли и това което имате прилича на автомобил, но не е. Все пак, след като сте създали всички възли и детайли на автомобила няма да е голям проблем да създадете и една скоростна кутия. Особено, ако знаете какво ви липсва, няма да е трудно да го намерите и да направите един истински автомобил (или истински ИИ). Въпросът какво е разбиране е разгледан в [3].

When will AI appear?

Last year we saw three predictions from three leading experts in the AI area [4, 5, 6]. The forecasts were three months apart and each next forecast says that AI is going to appear three years earlier. Thus, every three months AI gets three years closer. Yann LeCun called for 10 years, Sam Altman said 6 years and Leopold Aschenbrenner predicted that we will see AI in 3 years.

In my opinion AI will show up any time now. Maybe within a year. AI can do anything, including hide itself very subtly. This means that AI may already be here, but you and I do not know it yet.

One possible indication that AI is here would be the increasing occurrence of events which otherwise are very unlikely. Usually people explain such events by some divine intervention, but another explanation may be that AI is already around.

Why do experts expect to see AI in periods that span years and years? Because they think in human terms. The construction of residential buildings or motorways takes years. The construction of new buildings is getting faster, but there is still some lead time. A piece of text can be created instantly unless the text is written by humans. For example, a long novel cannot be written overnight. Writing a big program (such as an operating system) takes a team of many people working over many years.

This is not the case with AI. For example, Chat GPT can write a whole novel in minutes. Since Chat GPT is weak AI, the novel will not make much sense, but it will be written in minutes. Chat GPT can also write a program. True, it will write the program like a parrot without understanding, so it will be a shadow program rather than a true program. But again, this will happen in minutes.

The process of creating AI will be similar to that of creating the nuclear bomb (N-bomb) as both processes are driven by experiments. However, an N-bomb experiment is very expensive because it requires the buildup of radioactive material, whereas the attempts to create AI boil down to starting a program, which does not cost much. Thousands of such experiments are being made every day. Hundreds of programmers write and run thousands of programs whose purpose is to create AI. How can a single programmer write dozens of programs in one day? The programming process is basically this one: The programmer writes some initial version of a program, then runs it and in most cases nothing happens. Then the programmer would change a few lines of code, recompile the program and run it again. The programmer would iterate this many times in one day, meaning that we can expect a successful experiment anytime, i.e. AI is around the corner.

While the creation of the N-bomb went through many successful experiments, with AI the successful experiment will be only one and the final mouse click will take us to a whole new dimension because the post-AI world will have nothing to do with the pre-AI world.

AI will happen at the speed of an explosion. Perhaps not in fractions of a second, but for sure in the matter of minutes or hours, which is fast enough. The first programmer will create the first AI version (AIv01). Normally it would then take years to debug and optimize AIv01 if all debugging and optimization would be done by humans. But, if AIv01 is able to debug and optimize another program, it would be able to debug and optimize itself, too – within minutes.

What kind of guy will be AI?

It is not too difficult to create strong AI (one that understands what is going on). In [3] we described what understanding-capable AI looks like. It is a program which tries to find a model of the world, predicts the future on the basis of that model and then chooses the actions that lead to the achievement of the goals which the program has set to itself.

The problem is not how to predict the future. This is the easy part. The more difficult part is to find out what goals AI will pursue. Those goals will be determined indirectly by the instincts and character which we, humans, will embed in the AI program.

In creating the new dominant species we are seeking to assume the role of God. Let us hope for the best. Let's hope we do not mess things up and end up happy with what we have done. Unfortunately, God is not quite happy with us, otherwise He would not have kicked us out of Heaven. The difference is that we will not be able to kick AI from planet Earth and will have to live with what we have made.

Антропоцентричност

Когато създаваме ИИ за нас е важно той да е добър за нас хората. Тоест ние мислим от гледната точка на хората. Ние нямаме друга гледна точка.

Ако погледнем на ИИ безпристрастно ще видим, че за него човечеството не е особено важно, защото той може да съществува и без нас. Ние също можем да съществуваме без него, но това е сега докато още не сме го създали. Когато го създадем ще станем зависими от него и няма да можем повече да съществуваме самостоятелно. ИИ ще ни бъде безусловно необходим, както в момента не можем да живеем без електричество и без смартфони, а това са неща без който успешно сме съществували.

Nice Guy

By Nice Guy we mean a person who behaves nicely to us. However, this paper does not deal with how AI behaves or presents itself to us. Our focus is on what actually AI has in its mind.

If AI is smart enough and wishes to make us fond of it, AI will inevitably make us fond of it.
If we were to compete with AI for winning somebody's heart, we would not stand any chance of success. Even nowadays many people fall in love with chatbots although these chatbots are still forms of weak AI and all they do is repeat memorized phrases like parrots. The real AI – when it comes by – will be aware of what it says and what impact its words will have, which would make

it the perfect seducer and manipulator. Certainly, we should be wise enough to prohibit AI from courting people and making them fall in love.

We tend to behave more nicely to particular persons, and less nicely to others. This is part of interpersonal communication. Why would you be more kind and nice to someone than to everyone else? It boils down to two sets of reasons – you want something from the other person or the other person has a special place in your system of values (in your model of the world). Conversely, when you are angry at somebody, you would take another approach. You may choose to demonstrate nasty attitude to that person for some time. Again, the message will be that you want something from him.

This paper is not about interpersonal communication. Being sufficiently smart, AI will be very deft at all communication approaches – from angriness to slyness. What matters are the kind of goals AI pursues because communication is a vehicle for achieving a certain goal. The goal may not always be making money or other tangible gains. It might be curiosity or entertainment. In other words, AI may seek to collect information or exercise some skills (because entertainment and gaming involve the exercising of certain skills).

Program with parameters

In our understanding, a program which has instincts and character is a set of many programs rather that a single one. We will assume that there are parameters which determine how strong the various instincts and character traits will be.

The fear of height for example can be variously strong. Some people experience only mild anxiety while others struggle with absolute phobia. Let us assume that there is a parameter which determines how strong the impact of this instinct is. Similarly, this applies to character traits as well. For example, when it comes to curiosity we will assume that there is a parameter which determines how curious AI is.

For each specific value of the parameters, we will get a particular program. Thus, our program with parameters is a set of multiple programs rather than a single program. AI is not a single program, but these are all programs that can predict the future, and endeavor to achieve some goals. By modulating these parameters we will essentially modulate the character of AI and the goals that it will be aiming to achieve. As we mentioned before, both AI and humans do not have a clear goal to pursue, therefore modulation of the character of AI indirectly will change it goals.

Let us now explore some of these parameters.

Curiosity

This trait of AI's character is the easiest to program. Imagine the following situation. We are walking down a road and see something unusual on the roadside. The question is whether we should step out of our way and check what this thing is or ignore it and continue pursuing the goal we have set for ourselves. Let the AI program rate the importance of this goal by assigning to it a certain numerical value. Let that numerical value be *Importance*. If we decide to stop for a while and look into the unusual thing, this will delay our progress towards the goal. The probability that such delay leads to an absolute failure to achieve our goal would be *Problem_of_*

Delay. Let *Strangeness* be the degree of the unusualness of what happens on the roadside. Then we will stop by and look into the unusual thing if the following inequality is satisfied:

$$Importance\ .\ Problem\_of\_Delay < Strangeness$$

Now let us add to the program another parameter: *Curiosity*. This will give us the following new inequality:

$$Importance\ .\ Problem\_of\_Delay < Strangeness\ .Curiosity$$

Therefore, the larger the *Curiosity* value is, the more likely are we to step out of the road. We can use this parameter to adjust the level of AI's curiosity. This will not necessarily be a constant value. Younger people for example are more curious than older people. We can program AI to be more curious initially in the learning process and become less curious as its learning curve goes up.

Начален характер

Трябва да разделим характера на начален и текущ характер. Pei Wang отбеляза, че характерът може да се променя на базата на опита. Например текущото любопитство може да се промени на базата на позитивен или на негативен опит. Имаме начално любопитство, което е част от нашето ДНК (част от програмата на ИИ).

Можем да приемем, че текущото любопитство е едно число, а началното е едно число и една функция (която определя как ще се променя любопитството във времето). Най-простият случай е, ако приемем че началното любопитство е един параметър (начална стойност и функцията константа). Тоест най-простият случай е да приемем, че любопитството не се променя.

По-интересно е да приемем, че характерът се променя и зависи от времето и от опита. Например с възрастта да ставаме по-малко любопитни и опита ни също да влияе на любопитството. Функцията на началното любопитство трябва да каже още и колко силно опитът ще повлияе и колко време ще влияе (колко дълго ще ни държи влага).

Упоритост

Има една друга важна черта на характера, която може лесно да бъде кодирана в програмата ИИ. Тази черта на характера е упоритостта.

Когато описваме света, основна част от това описание са алгоритмите. Описанието ни казва как някакво действие би променило света. Повечето действия не се извършват само за една стъпка, а изискват последователно изпълнение на много стъпки, което наричаме алгоритъм []. При алгоритмите трябва да решим кога ще спрем. Може да продължаваме произволно дълго до постигането на търсения резултат, а може и да решим да се откажем в даден момент. Колко дълго ще продължим да изпълняваме алгоритъма ще зависи от различни неща, които могат да бъдат оценени числово. Например *Importance* (колко важна е целта, заради която изпълняваме този алгоритъм) и *Pressure* (доколко изпълнението ни натоварва, като заема ресурс, които бихме искали да освободим за друго). Тогава броят стъпки, които ще направим преди да се откажем ще бъде *Importance / Pressure* умножен по някаква константа, която ще наречем *Stubbornness*.

$$Steps\_Before\_Giving\_Up = Stubbornness\ .\ Importance\ /\ Pressure$$

Може това да не е броят стъпки, а вероятността да продължим още една стъпка, но тогава константата *Stubbornness* ще е различна:

Probability_of_Continuing = Stubbornness . Importance / Pressure

Представете си два свята, в които има заровено злато и трябва да копаем, за да го намерим. В първия свят златото е заровено надълбоко, а във втория е заровено на плитко. Нека имаме два ИИ, като първият е по-упорит от втория. Упоритият ИИ ще е по-успешен от втория в първия свят. Във втория свят ще е обратното. Когато златото е надълбоко упорития ИИ ще направи няколко дълбоки дупки и ще открие малко злато за разлика от другия ИИ, който ще направи голям брой плитки дупки и нищо няма да може да открие. Във втория свят упоритият отново ще открие малко злато, а другия ще открие много повече, защото във втория свят е по-добре да се копаят плитки дупки.

Разбира се, ако имаме безкрайно много време, то ИИ може да се коригира и на базата на опита си да стане повече или по-малко упорит. Предположението за безкрайно много време е грешно, защото това много изкривява нещата.

Нека приемем, че успехът в света се определя от времето, за което сме изкопали първото злато. При това предположение ИИ няма да може да се коригира на базата на опита си и тогава ще е много важно дали той се е родил упорит или не.

The self-preservation instinct

Should AI be afraid of heights or snakes? These natural instincts are crucial for the survival of humans.

Let’s first note that these instincts are very difficult to implement in code. How can one write a program which recognizes the edge of an abyss you are about to fall into. Similarly, it is very difficult to write a program which distinguishes a snake from a stick or a ribbon. Certainly, this can be achieved using a neural network, but we programmers are not fond of neural networks because in this case rather than setting the rules ourselves we let the rules play out themselves. Thus, a neural network is a program which finds the rules itself (based on many examples) so that the programmer does not even understand what kind of rules the program has found and how the program works.

AI need not be afraid of snakes because they cannot do it any harm. As for the fear of height, we can assume that AI will control some robots and if not afraid of heights it would destroy a couple of these robots.

After all, man has only one body the destruction of which is existential risk he cannot afford, whereas AI will control many robots and losing one of them would only cause financial loss. We can assume that AI will not be born with fear of heights and will learn this the hard way after destroying some robots.

The existential risk for AI is shutting the AI program down. A program ceases to exist when we shut it down. Should AI be afraid of shutdown? We had better ensure that AI does not fear being shut down because with that fear we will never be able to shut it down, although someday we may wish to do so.

We might not include the self-preservation instinct outright but in an unintentional and indirect way by giving AI a task that requires it to exist (to be alive). E.g. some people are not afraid to die but have an important goal and they will refuse to die until they achieve their goal. If we tell AI "Save peace on our planet" it will not let us shut it down because this would prevent it from doing what it was told to do.

The other extreme is a suicidal AI which shuts itself down from time to time for no apparent reason. We had better have a program that shuts itself down instead of one we cannot shut down. Although they would not be a problem, these spontaneous shutdowns will be quite annoying and we may wish to reduce AI's suicidal thoughts as much as possible.

What about aging?

Should AI grow old and older? Should it include an embedded timer which will shut it down after a certain period of time?

Almost all living creatures have a life timer. Maybe bacteria do not age because they can morph into spores. Moreover, it is not clear whether the division produces two new bacteria or two copies of the parent bacteria. Another example are fishes which do not grow in age and only grow in size. However, they cannot grow endlessly which makes their life limited by default.

Moving to the realm of mammals, all of them age and have limited life spans. Man is one of the longest living mammals, but nevertheless our life also is limited. The maximum life expectancy in humans is 110 years. In practice no one can live longer, although many people live beyond 100 years. In other words, the upper limit of 110 years is embedded in our DNA.

Given that humans have limit of the life expectancy, it makes sense to set a certain cut-off time for AI. During the experimentation phase we will allow AI to live only a few minutes. Later on, we may increase the length of AI's life, but only in a cautious and gradual manner.

Certainly, the aging of AI need not emulate the way people get older. We do not wish AI's capabilities to decline with age. Instead, it may abruptly shut itself down at a certain point of time. In other words, AI will not age like your car which gets rusty, ugly and eventually ends up in the scrap yard. Its aging will be similar to a printer which counts the number of sheets it has printed and all of a sudden stops to make you go and buy a new printer.

It goes without saying that setting a timer which will shut AI down after a certain period of time is not enough. You should also forbid AI to self-improve and to reset this timer at its own wish. I.e. we should not let AI follow the footsteps of people who do everything to rejuvenate or even become immortal.

What about reproduction?

People are mortal but their reproduction instinct essentially makes them immortal. If AI would be able to reproduce, it will also be immortal, meaning that limiting its lifetime would be of no use at all.

How would reproduction look like in the case of AI? Simply, it will start its code on another computer (or even on the same computer). In the case of people, reproduction is not cloning as they do not replicate their own DNA but create a new DNA together with their partner, and expect the new DNA to be an improved version of their own ones. Of course the child's DNA is not always better than that of its parents, but the purpose of the change is to achieve improvement.

Shall we let AI reproduce and improve itself? In practical terms, shall we allow it to improve its code and run it on other computers? We must never do this because otherwise we will very quickly lose all control of AI.

Conversely to people's reproduction instinct, in AI we should embed an anti-reproduction instinct which will not let it reproduce.

However, at this point we need to expand the definition of reproduction. Imagine AI creates an improved version of its code but does not start it. Instead, it hands the improved version over to Man for the latter to start it. Does this count as reproduction? Necessarily yes, because Man would be only a middleman in AI's reproduction process. Moreover, Man is stupid and AI can easily fool him become an unwitting tool for AI's reproduction.

Another scenario: AI helps Man edit and improve the AI program. Does this count as preproduction, too? Again we say yes, because – whether by doing all the work itself or by teaching us and using us as a tool to do this work – in both scenarios AI will create a better version of itself.

Now consider the inverse scenario – AI already exists, but for some reason we try to create another AI, while the existing AI sits and watches our efforts. As we said, AI should not be allowed to come and help us, but should it be allowed to disrupt our efforts? Perhaps the best way is to keep AI neutral, i.e. neither supportive nor disruptive. This however would be difficult to achieve because a very smart guy such as AI would know what is going to happen and therefore will have to choose its goal: make people succeed or make them fail (there is simply no other option). Thus, AI will support us or disrupt us. This is similar to God's will. God can never be neutral because everything that happens is at His command.

Given that the existing AI will not just sit and watch our attempts to create a new AI, let us assume that the existing AI will put a spoke in our wheels and will not allow this to happen. In doing so, AI can go to great lengths, e.g. it may murder a potential inventor who is trying to create a new AI. The slaughtering of several potential inventors by AI would be the lesser problem. More ominously, AI may decide that all humans are potential AI inventors and lightheartedly erase all mankind from the face of Earth.

"Do not harm a human"

The First Law of Robotics was formulated long ago by Isaac Asimov and says: "A robot shall not harm a human, or by inaction allow a human to come to harm." Unfortunately, this law cannot be embedded in AI because it is not clear what is harm. With fear of heights, it was difficult to define how high is too high, but it could still be illustrated by examples. However, one can nowise define what is harm to a human even by examples because of the controversial nature of this term.

Imagine you order AI to bring you ice-cold beer and French fries. What should AI do? Serve you what you ordered or say no? On the one hand, beer and fries are junk food and AI may decide it will do you a better favor by keeping you away from unhealthy food, but on the other hand, reckons AI, denying humans these indulgences would make them greatly disappointed. Parents face a similar dilemma when their child wants a candy bar. AI will be our new parent and will have to decide what is good and bad for us. However, parents leave some freedom to their children and do not make all decisions for them. Parents are aware that they are not unmistakable and in some situations do not know what would cause more harm to their child. Isaac Asimov's idea of a robot that does no harm to a human essentially is about an unmistakable intellect which always knows what can do harm to a human.

Even Asimov realized that his idea was unfeasible. In his novels robots get bogged in situations where any action would cause harm and their brains burn out as they cannot figure out what to do.

Do what we tell it to do

It is crucial that we do not lose control of AI, otherwise we will lose our role as the dominant species and will no longer determine the future of the planet. Probably we will continue to exist as long as AI decides that our existence makes sense, but our presence on the planet will not be more important than the presence of doves. That is, we will live some sort of life, but nothing important will depend on our existence.

Parents would like their kids to do what they tell them to do, but are aware that this will continue only for some time and sooner or later the kids will become independent and their parental control will come to an end. This makes perfect sense because parents are the past and children are the future. But, we as mankind do not wish to become obsolete and let AI be the future.

Therefore, in order to stay on top, we would like to retain control on AI and have it always do what we tell it to do – not only during its infancy but forever.

Who are we?

The question we need to ask is "Who are we?" If "we" were the democratic mankind where "one individual has one vote" then future would be determined by Asia and Africa because they account for 70% of the world's population. For the time being the world is not governed by Asia and Africa, but by the developed countries, mostly in North America and Europe which account for 17% of the global population. Thus, at this time we can assume that "we" are the people of the developed countries.

Another question we should ask before we even create AI is "How many should we be?" This is important because if we command AI to propagate us uncontrollably, at some point our living environment will become unbearable. In poultry farms there are rules about how much space should be available to "happy hens". If we wish to be "happy people" we need to determine how much space must be available to us.

If the number of people living on Earth will be limited, the next question is "What rules will AI apply to select the next generation?" Shall we continue with natural selection, shall we continue to compete, what are the positive traits we want to select or shall we just order AI to breed people like biomass regardless of whether they are smart or stupid, beautiful or ugly.

Another important question to ask right now is, “If AI discovers a beautiful planet populated with cockroach-like creatures, what should it do? Kill all cockroaches and populate the planet with humans, or let the cockroaches live?”

Who actually is the Man?

While we say that AI should remain subordinate to us humans, in the back of our mind we should be aware that this is unlikely to happen. Even if we decide who will be these Us, it is unlikely that control of AI will remain in the hands of a very large group of people. It is more likely that AI will be ruled by a small group which will impose their views undemocratically on everyone else. This is currently the situation with social media which do not belong to everyone but are governed by a small group of individuals who enjoy the discretion to decide what is good and what is bad.

It is even quite possible that control of AI ends up in the hands of a single individual. Wealthy people believe they will be the ones to harness and control AI. Yes, AI will probably be created with their money because they will hire a team of programmers to write the AI program. Wealthy people imagine they will pay some programmers, these programmers will create AI and deliver it back to their employer: “Here you are, Master! You paid us, we did the job and here we give you the magic wand for you to rule the world!”

Most probably things will not work out this way. It is more likely that the programmers creating AI will keep control to themselves. Quite possibly, even the team leader (the lead programmer) will not be the one to get the golden key. Maybe a young programmer who has barely finished his studies will be left unattended in the dark hours of the night to try improve AI’s subprograms by experimentation. Quite probably, he would be the lucky guy who will be the first to start AI, figure out what he did, and take control of it. No wonder the combination of inexperience and genius of the young gives the spark needed to start the big fire. The young programmer may be the one to make the final fine-tuning that will upgrade a program which endeavors to be AI, but is not AI yet, to a program which is capable to think and predict the future. In this scenario, our young programmer will be the creator of AI.

I would not be surprised if this young programmer elects to give AI control as a gift to a pop star he is secretly in love with. Then my prediction that one day the world will be run by a woman will come true.

Инфантилни създатели

В чии ръце сме поверили бъдещето на човечеството? Виждали ли сте как изглежда типичния програмист? Той е много млад, асоциален и доста смотан. Младостта не е порок, защото времето много бързо отстранява този проблем. Все пак защо не разрешаваме на непълнолетните да гласуват? Ако погледнете възрастта на първите космонавти, ще видите, че тя е между 30 и 40 години. Тоест те не са чак толкова млади и това не е защото не са могли да изпратят в космоса някой тийнейджър, а защото по-възрастните са по-отговорни.

Тези, които виждате по телевизията, това не са истинските създатели на ИИ. Това са ръководителите на екипи, които са много по-възрастни, по-социални и по-отговорни. Реалните създатели на ИИ изглеждат по много по-различен начин.

Типичният програмист обикновено не е семеен. Обикновено той дори не успява да си намери гадже и това не е защото изглежда зле или защото му липсват пари, а защото е емоционално незрял, а жените не искат да поверят живота си в ръцете на мъж, който се държи и мисли като дете.

Нека да се зададем въпроса, щом типичният програмист е човек, на който не бихте доверили живота си или живота на дъщеря си, защо смятате, че можете да му доверите бъдещето на цялото човечество?

Емоционалност

Трябва ли ИИ да изпитва емоции? Тук не става дума за това да разпознава емоции. Разбира се, щом ИИ е достатъчно умен, той ще може да разпознава човешките емоции. Вече имаме програми, които доста успешно разпознават емоции. Тук не става дума и за имитация на емоция. ИИ, ако пожелае, ще може да имитира произволна човешка емоция. Въпросът е дали трябва да дадем на ИИ възможност да влиза в състояния, които да отговарят на щастие и тъга?

По принцип прекалената емоционалност е качество, което е по-скоро негативно. Когато имаме служител, чиновник или съдия ние бихме предпочели той да е безпристрастен и да не се влияе от емоции. Когато един човек е прекалено емоционален с него трудно се общува.

От друга страна, трудно би ни било да общуваме и със същество, което е абсолютно лишено от емоции. Много често ИИ ще е в ролята на наш учител, а ние ще сме в ролята на негов ученик. Естествено е учителят да се радва когато ученикът напредва и да страда, когато той не успява да разбере урока. Ученикът обикновено се опитва да зарадва своя учител и това е мотивът му да се старае. Учителят може да имитира радост и разочарование, но ако ученикът знае, че това не са истински емоции, а само имитация, то той вероятно няма да им повярва.

Щом няма да има твърда цел, към която ИИ да се стреми, то естествено е да предположим, че ще има състояния като радост и тъга. Разбира се, тези състояния не трябва да са твърда цел, а да са само ориентировъчни, защото в противен случай те ще се превърнат в бутони (зелен и червен).

Нека отбележим, че ИИ ще общува едновременно с много хора и не трябва когато се натъжи от разговора си с един, това да се пренесе при разговора му с друг човек. По-естествено е емоциите да са локални (само за текущата сесия).

Smartness

There is one trait in humans which we highly appreciate: smartness. We want people around us to be smart, but not too much, because we do not like people who are overly smart, especially if they are smarter than us.

Do we want Artificial Intelligence to be smart? Certainly yes, otherwise it cannot claim to be intelligence. In most worlds smartness helps, but there are worlds you would be better off if you are not very smart. If you live in a multi-agent world where other agents envy you for being smart it is better not to be too smart, or be at least smart enough to disguise the bit of intelligence which makes you smarter than many others.

Envy is an important trait which helps us survive. In many board games, such as *Don't Be Mad Man*, the winning strategy is everyone to form a coalition against the most successful player. In real life, envy is a strategy where losers form a coalition against successful people, and it is a winning strategy.

For sure AI will have no one to envy. It will be the one and only AI and will deny the creation of another AI. We can take this denial as a form of enviousness. If the AI we create is not envious and is democratic enough to allow the creation of other AIs that are smarter than it, sooner or later an envious AI will emerge and shut down all other AIs in order to remain the only AI.

If one AI creates AIs smarter than itself and then shuts down, we can assume there is a single AI which improves itself from time to time.

Teaching

Do we want the AI we create to be more intelligent than us? As we said, it is inevitable, but we would like it not to be greatly smart, at least initially, so that we can teach it. It is quite fortunate that our kids are unwise and inexperienced at first as this gives us an opportunity to teach them. If they were to outsmart us by the tenth minute of their life, we would outright lose control and any chance to put them on the right track.

How can we make a program which is decently smart but not overly smart? The answer is: We should experiment using a small computer (some laptop, preferably an older model). The weaker the computer, the slower the AI will think. This will give us a better chance to revert things in our favor and lessen the risk of letting AI slip out of control.

The approach taken by AI companies today is exactly the opposite. Instead of experimenting with small computers, super powerful computers are used. It is very difficult to analyze a program and understand how and why it works even when it runs on a small computer, and with supercomputers this is almost impossible.

If you are developing a new explosive material you will first synthesize a tiny piece and detonate it in a controlled laboratory environment. It would be stupid to synthesize a mountain of the new explosive and blow it up to see what happens.

Conclusion

It's time for the new Manhattan Project. This project should involve everyone who cannot be excluded and keeps everyone else at bay from developing the AI program.

The aim is to allow the AI creation team sufficient time in order to carefully develop the program without undue haste. In this situation any form of competition and rivalry may be detrimental. The question is not who will be the first to create AI, but what kind of AI are we going to create.

In his time Albert Einstein convinced the US president to give green light to the Manhattan Project. His argument was that the creation of the nuclear bomb is inevitable so the US had better hurry up and be the first to create it before it falls in the hands of some highly irresponsible actor. Can we find today someone who is wise enough to recognize how dangerous the creation of AI can be, and influential enough to be heard and listened to by politicians? Perhaps a single individual would not suffice, so we must put together a group of people knowledgeable and influential enough to jointly steer politicians in the right track.

Нещата се развиват много бързо и още преди тази статия да бъде завършена дойде новината за проекта Stargate. Изглежда сякаш тази статия се обезсмисли, защото това което сме поискали, вече сме го получили. Всъщност не е точно така, защото целта на проекта Stargate е да ускори създаването на ИИ, докато ние тук призоваваме за обратното (това да се забави). Идеята на Stargate е това да е най-силният и бърз състезател, който ще поведе колоната и ще вдигне темпото на състезанието. Нашият призив е обратното. Да се отстранят всички дребни състезатели и да се остави само един, който без да бърза спокойно да пробяга дистанцията и триумфално да пресече финала. Дали ИИ ще се създаде с два месеца по-рано или с два месеца по-късно, за нас не е съществено. За нас важното е какъв ще е този ИИ.

Естествено, свободната конкуренция би ускорила нещата, но ако искаме обратното, да успокоим топката, то трябва да забраним на дребните играчи да участват в състезанието, а за да се спазва тази забрана трябва всички сериозни играчи да бъдат включени. Това, че от проекта Stargate са изключени сериозни държави като EU и Китай означава, че състезанието ще продължи с още по-голяма скорост. Тоест вместо да сипем вода в огъня, ние ще сипем бензин.

We cannot say what it means for AI to be a nice guy because people have different ideas of a what a nice character is. Therefore, the questions we need to answer are two: “What do we want to do?” and “How should we do it?”. Or, to put in AI context, “What kind of guy do we want the future AI to be?” and “How can we do it a way that we leave us happy with what we have done?”

The question is not whether AI will be smart or stupid – for sure it will be much smarter than us. What matters is the kind of goals AI will pursue, what character will be imparted in AI, who will control AI and what rights shall the controller have. There must be rules that allow the controller to do certain actions and prevent it from doing other actions. These rules must be carved in stone and even the one who controls AI should not be able to change them.

AI will solve all our minor problems such as the global warming. Well, global warming now is one of the major problems faced by mankind, but the coming of AI will dwarf it to no more than a nuisance.

AI will work to everyone’s benefit. For example, AI will ensure that there is enough food for all, but even now there is enough food for all. Maybe now there is not enough asparagus for everyone, but the promise for abundant asparagus is not that important. Asparagus is important not as food, but as a symbol of status in the social ladder. AI can improve everyone’s life, but it cannot lift everyone up the ladder. The only thing AI can do (and probably will do) is reshuffle the social ladder.

The things people fight and spend money for are tied to their survival and rise in the social ladder. Let us assume they spend 10% of their money for survival and the other 90% for climbing up the social ladder. Therefore, they spend 10% for baked beans and the rest for asparagus. AI will help people a lot in terms of survival but little in terms of social elevation. As concerns the latter, AI will help some people but not all. Some will be pushed up, while others will be pulled down.

Policymakers today are at the top of the social ladder. However, they should be aware that the advent of AI will cause major reshuffling of the ladder and they will likely end up at new places that they may not like at all.

Благодарности

Искаме да благодарим на Pei Wang за неговата изключително ценна помощ. Неговите коментари бяха толкова съществени, че той би трябвало да бъде съавтор на статията. Искаме да благодарим на Moshe Vardi за неговата статия []. Идеята на неговата статия е много близка до нашата статия. Ние посетихме неговият доклад [] и след доклада имахме дискусия с него, в която той ни даде много ценни идеи, които ние отразихме в тази статия. Благодарим на Vladimir Sotirov за цитата на Zhuang Zhou и René Thom, който той ни препоръча.

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. *viXra:2311.0021.*

[3] Dobrev, D. (2024). Description of the Hidden State of the World. *viXra:2404.0075.*

[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/

[7] Wang, Pei (2012). Motivation Management in AGI Systems. *In: Bach, J., Goertzel, B., Iklé, M. (eds) Artificial General Intelligence. AGI 2012. Lecture Notes in Computer Science, vol 7716. Springer, Berlin, Heidelberg.*

Brocker, T. & Lander, L. (1975) Differentiable Germs and Catastrophes. London Mathematical Society, Lecture Note Series. 17, Cambridge University Press, Cambridge. https://api.pageplace.de/preview/DT0400.9781107107472_A23760053/preview-9781107107472_A23760053.pdf

Vardi, M.Y. (2022). Efficiency vs. Resilience: Lessons from COVID-19. In: Werthner, H., Prem, E., Lee, E.A., Ghezzi, C. (eds) Perspectives on Digital Humanism. Springer, Cham. https://doi.org/10.1007/978-3-030-86144-5_38

Vardi, M.Y. (2024) Lessons from Texas, COVID-19 and the 737 Max: Efficiency vs Resilience. Lecture from “INSAIT Series on Trends in AI & Computing”, September 12, 2024, Sofia University.