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, AI control, AI regulation, AI character, AI as a nice guy.

Introduction

Can AI have different variants? Can different AIs have different personalities?

This paper is not the only one that claims that AI is not a single program, but that there are many variants that are significantly different. For example, De Kai in his book [19] claims the same. He even uses the plural and instead of AI he speaks of AIs. However, there is a difference between us and De Kai. He assumes that in the future many different AIs will coexist, while in our opinion there are many possibilities, but there will be only one winner (just as there are many presidential candidates, but in the end we choose only one and it matters who we have elected).

De Kai says that AIs will be our artificial children who will take care of us, their aging parents. We say that AI will be our new God who will take care of us. There is no difference between us and De Kai here, because both he and we expect someone to take care of us.

De Kai, like us, predicts that after the advent of AI, the world will be completely different. And he, like us, says that it does matter what kind of AI we will create and what the new world we will live in will be like. The difference is that according to De Kai, the most important thing is raising, while according to us, the most important thing is the AI's DNA (its program code), because that is where its character, goals, and instincts will be laid down.

For most of his book, De Kai ponders the question of what the new world that AI will bring us should be like. We are mathematicians, and it is not our job to say what the new world should be like. There are politicians who are tasked with that task. Our job is simply to show politicians that there are different options and to help them choose the one they think is the best. Why should we trust politicians and let them make the choice? Because the alternative is for the new world to be

chosen randomly on the principle of lottery. Surely it would be better if someone (whoever he is) made the choice.

According to Geoffrey Hinton [20], humanity's only chance of survival is to instill maternal instincts in AI. Hinton's idea is that if AI loves us like a mother loves her children, it will take care of us, no matter how stupid and helpless we are. Is it enough for AI to feed and clothe us and take care of our basic needs? Shouldn't it also obey us? The mother takes care of the baby, but she makes the decisions herself. The baby may show its mother what it wants by crying and grimacing, but whether those wants will be fulfilled depends entirely on her. She will decide what is good for the child and what is not.

For example, if she is buying a new house, the mother will make the choice herself. She may show the baby the different options and try to guess which one is best for him, but the final decision will be made by herself based on criteria that the baby does not understand at all.

Creating AI is like opening Pandora's box, but here we are not talking about a simple box with one lid, but about something like a cabinet with many drawers, and it does matter which drawer we pull out. The moral of the story of Pandora's box is that this box should not be opened. In our case, there's no way AI won't be created because there are too many people working on this task. The question is how we choose which drawer to pull out. The lottery approach is very wrong, because among the AI variants there are also quite unfriendly ones that will not obey us, and may not want to take care of us.

Donald Trump said that whoever creates AI will rule the world. Putin has said something similar, but both are wrong. The world will not be ruled by the creator of AI, but by AI itself. What this AI will be like and where it will lead us will be decided at the time of its creation, and it is good to think carefully before allowing it to replace us (before opening Pandora's box).

Should AI obey us unconditionally? When we design a new machine, we usually put safeguards in place to keep us from hurting ourselves or breaking the machine. Perhaps in this case, too, we should consider safeguards that will limit our freedom but make using AI safer.

In this paper, we do not presume to say what the new world that AI will bring us should be like. We only presume to ask some questions. Our job as mathematicians is not to answer these questions, but to say how one can construct an AI that will correspond to what politicians will want.

What kind of AI do we want to create?

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. In [4] we assert that the first AI created will be the only AI ever created. Furthermore, once created by us, the one and only one 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. We 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?

The core idea we wish to impart by this paper is that Artificial General Intelligence (AGI) is a dangerous thing which warrants the highest caution. We aim to persuade the reader that rather than being a single program, AGI is a class of multiple diverse programs and therefore it matters a lot what kind of program we will choose to create. The various AGI versions can be described by parameters which define its character. As we create AGI it is crucial to ensure that its character can be regulated and that we are able to choose an AGI which is acceptable to us.

What is AI?

All references to AI in this paper are references to AGI. By this we mean that AGI is a reasoning machine which is incomparably smarter than man. Many writers about the “true” AI split this concept in two: AGI and artificial superintelligence (ASI). In their view AGI is a machine which

reasons like a human, while ASI is a machine incomparably smarter than humans. In our view AGI and ASI are not different at all just like a program which plays chess like a human is not fundamentally different from a program which plays chess incomparably better than any human.

The notion that AGI and ASI are two different things sends a comfort message to the subject-matter experts. They assume that we will first create AGI and only then we will arrive at ASI. This somehow leads them to comfortably believe that we have plenty of time. Indeed, we would be very comfortable if we had two more steps before getting to technological singularity. Unfortunately, the steps are not two. There is only one step left and we should think well before we make it.

The bottom line is that all references to AI in this paper are references to true AI, AGI or ASI which in our view are all the same.

According to [2] AI is a program which is sufficiently smart. First, why a program and not a machine? We might imagine AI as special-purpose machine designed to solve a specific task. Church asserts in [15] that any calculating machine can be simulated by a computer program. Thus, for any calculating machine there is a computer program which can be launched on any computer and will always do one and the same thing. Of course, a special-purpose machine would run faster that the corresponding computer program. (This would be the case if the machine and the computer on which we launch the program are comparable in terms of number of transistors and performance. So, even if the special-purpose machine runs a bit faster than the computer program, it would not matter a lot to us. What matters to us is the kind of thing the program does, while performance is of secondary importance.)

What does sufficiently smart mean? A program is sufficiently smart if it is smarter than a human being. The smarter between two intellects is the one which in almost all worlds performs at least as well as the other one. We say “almost all” here because we can always construct a special world in which the opposite holds true (the second one performs better than the first one).

In [2] we have an important specificity. There 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 special 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.

In [11] Pei Wang explores AI which has multiple goals and is moreover able to change these goals. These may be intermediate goals, but Pei Wang assumes that even the main goal can be changed. Thus, the notion of a changing goal is not a new one. Pei Wang suggests that for a system to be smart is should be able choose its goals as it thinks fit.

This notion is further elaborated in [17] where different agents have different instincts and different personality parameters. The authors in [17] deal with "cautiousness" (*desire threshold*) and evaluate it by using a parameter.

The authors in [17] indicate that the agent's success in a given world depends on the agent's personality. This would be the case when the goal is clearly defined. When we do not have a clear goal then the personality indirectly sets the goal.

Is AI possible?

We will start with the words of the Chinese philosopher Zhuang Zhou and the additions made by the French mathematician René Thom (this is the motto of the book [1]):

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.

Is AI possible or do they just scare us with it just as they scare children with Boogeyman?

The vastly prevailing opinion is that machines are unable and will never be able to think. Almost everyone believes that thinking is a privilege reserved only for humans. Is it possible to create a machine which thinks like a human but is immeasurably smarter than humans?

Let us not argue about this. Let us assume that AI is 99% impossible. However, this leaves 1% probability that the opposite is true, so let us stop for a while and ponder on what we should do if AI is possible. When it comes to the fate of mankind it is worth to spend some time thinking on this hypothesis despite the minor chance that out fears come true.

We, the authors of this paper, are part of the minority who believe that AI is possible. However, the mere fact that we believe in AI does not mean we believe in everything. For example, we do not believe in aliens and ridicule people who believe they see flying saucers. So we can understand people who do not believe in AI and ridicule us.

The question "What is AI?" is important as the questions "What is a dragon?" and "What is a ghost?". If you do not believe in AI, in dragons, or in ghosts, then these questions are meaningless. Nevertheless, we can ask "How can one fight a dragon?" although we may not be quite sure what a dragon is. Let us assume that AI is a machine incomparably smarter than humans. Whether such a machine can be made is another question. Maybe it cannot be made, but if we look around we will see so many things we thought were impossible that turned out to be perfectly possible.

The theory of the non-existent object is absolutely meaningless. (We mathematicians know that if an object does not exist, then anything can be said about it and it will be true. From the fact that something is non-existent it follows logically that it possesses all sorts of properties.) However, if an object exists or is likely to exist, then this theory is not meaningless.

Can we control it?

Very few people believe that AI is possible, but virtually nobody believes that AI can be controlled. For example, according to Radoslav Pavlov [12] AI is possible, but it is a kind of natural phenomenon that we cannot control and steer. An example of such a natural phenomenon is a hurricane. We can to a certain extent predict where the hurricane will come from, but are unable to change its path and steer it to a less populous area.

Let us not argue about this either. Let us assume that with a probability of 99% AI cannot be controlled. This leaves room for a minor probability for the opposite. Let us put our stakes on that minor probability and think how could we control AI and steer its character to something which is more beneficial to us.

Even in the case of hurricanes, we try to control and steer them. Of course, if we master the skill of controlling hurricanes, we will have to keep that skill secret because whichever way we steer a hurricane, somebody will suffer damage and blame it on us.

It is wise for the one who finds out how to control AI to keep this skill secret to avoid the resentment of the victims of AI's bad character. On the other hand, the one who can control AI will be strong enough not to worry about the frustration of others.

Can we regulate AI?

Not AI *per se*. What we can regulate is the AI creation process, but once AI is created, we will lose our power of control and our ability to steer its behaviour will be strongly limited. We might retain some levers of control, but only in case that we have embedded these levers in AI during the AI creation process.

Programmers believe they have full control on the programs they write. This holds true for ordinary programs since programmers can always shut down, modify and restart them. Unfortunately, AI is not an ordinary program and we will not be able to shut down and modify it as we wish for two reasons: 1) AI will manage processes which are vital to us, and 2) we will not be able to shut down AI unless AI itself agrees to be shutdown.

Therefore, AI regulation should occur before AI creation because after that it will be too late. It often happens that someone in a senior position has the power to rule people and believes that he will have that power forever. One should know when the power is in his hands and when he has lost it. We as mankind should be aware of this and use our power over AI while we still have it.

What will be the consequences?

We all agree that the coming of AI will be a great test for mankind. This challenge can be likened to a heart surgery, however the patient in this surgery will not be just one person but the whole of mankind.

It does matter how we get through this adventure because it does matter what kind of AI we are going to create. As with heart surgery, the risks are many. For example, we may not wake up after the anesthetic. Nevertheless, let us not think about the worst but focus on the positive scenario.

When one prepares for a heart surgery, the first question is whether the patient can do without the surgery. In our case this is not possible because the creation of AI is inevitable. Then how to deal with it? It matters a lot what hospital we will choose, who will be the surgeon and what kind of new valve they will implant in our heart. Another question is whether we want the surgery to be a planned one or we will have emergency surgery. A planned surgery is preferable because we will have the required tests, prepare our body and have the time to write our testament. Urgent surgeries just happen and are out of our control.

Somebody should prepare mankind for the impending heart surgery. The purpose of this paper is to gather people who will voice this issue.

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 to be educated before they can become the aware creature which we are discussing here. The education of Man is everything that has happened in his life (his history) from the very conception to the present moment. Accordingly, the education of AI will comprise all of its history from the time the program was started 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 education is the set of memories and knowledge that reside in our mind. In the case of AI we can assume that education is the program's current status (the content of variables, arrays, files, etc.)

Therefore, in our mind AI includes a program, a computer that runs the program and the education of that program (its current status).

Education

In humans, DNA is not everything. Apart from DNA, there is education 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 education. 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 need only to add one more step: make AI understand. This step will inevitably be made, and it will be made very soon.

There is already Chat GPT – a program that successfully imitates AI, but lacks the understanding in question. For this reason, Chat GPT looks like AI, but is not AI. Imagine you have a very fine car. It has leather seats and trims, a stereo system and a powerful engine. Your dream car has everything except a gearbox to connect the engine to the wheels. Without a gearbox the car will not move and what you have looks like a car, but is not. Well, once you have created all the assemblies and details of the car it will not be a big problem to create a gearbox as well. Especially when you know what is missing, it will not be difficult to find it and make a real car (or real AI).

We do we mean by saying that AI should be able to understand? We mean that AI should be able to search for a model of the world. The model is composed of i) the set of internal states of the world and ii) the function which describes how the world moves from one state to another in response to some action. In addition to the model, the current state of the world is required for the understanding process. We do not need to find the exact model of the world or the world's exact current state. What we need to find is their approximated description. Before we can make such a description, we need a language for description of worlds. These issues are addressed in [3] and [5].

Then, how does weak AI work without understanding? Instead of searching for a model of the world, the weak AI uses the approximation method. Imagine there is a complex geometric shape and your task is to describe it. You can do it in two ways: using a mathematical formula or using paper and glue to create a papier-mâché approximation of the shape.

Why finding a model is a better idea than approximation? A model will enable us plan our next actions which will drive the world from its current state to the state we want it to be. A model allows us to think several steps ahead, while approximation only tells what the next action should be. In other words, approximation takes us only one step ahead and this is the approximation of the teacher we are trying to imitate. Therefore, we cannot approximate without having a teacher, while having a model will equip us with the ability to choose the right action ourselves. Approximation needs a huge number of examples on the basis of which we can successfully imitate the teacher. Conversely, in order to find a model, we need only a limited

number of examples – only those that are sufficient to help us identify the most probable model within the set of possible models.

The good thing about approximation is that an approximation technique is already available (neural networks), while a model-finding technique has not been developed yet.

When will AI appear?

Last year we saw three predictions from three leading experts in the AI area [6, 7, 8]. 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 our 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 we and you 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. (We assume that natural intelligence is able to create AI. Once we have assumed that this is possible, we should also assume that AI is capable to self-improve and sophisticate its own code.)

What kind of guy will be AI?

It is not too difficult to create strong AI (one that understands what is going on). In [5] 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 us 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.

Anthropocentricity

As we create AI we have better make sure that it is nice to us humans. That is, we think from the perspective of humans. We do not have another perspective.

If we look at AI dispassionately we will find that mankind is not very important to it because AI can exist without us. We can also exist without AI, but this is only for the time being and before we have created it. When we create it we will become dependent on it and will no longer be able to exist independently. We will need AI indispensably like electricity and smartphones nowadays although we fared quite well without them before.

Prof. Vardi asserts in [16] that we humans should aim to build machines that augment, not replace, human intelligence. He goes on to say that we are rushing to create AI with little consideration of its impact. It is shown in [16] that technical progress results in the creation of increasingly sophisticated machines, which leads to the conclusion that the creation of AI is inevitable. Therefore, we should accept the fact that something smarter than humans will come into being, and think about how we will live with that thing and how can we preserve at least partially the meaning of our existence.

Nice Character and Nice Guy

The concept of *nice character* cannot be strictly defined because of its subjective nature. Various people have various preferences to other peoples' characters. Even a single person would not be sure what a nice character is, because people do not know what they want.

This paper will not deal with nice characters. Instead, we will try explain what is a character and how it can be controlled and regulated. The ultimate goal is to empower us select the kind of character we want AI to have. Furthermore, we will address the question of what we actually want, because before you reach your goal you should know where you want to go.

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 a parameter we will get a specific program. Thus, our program with parameters is a set of multiple programs rather than a single program. AI is not a single program, but a set of all programs that can predict the future and endeavor to achieve certain 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 the modulation of AI's character will indirectly change its 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 to 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.

Many authors deal with the curiosity of AI, although most of them tend to use different terms. For example, Sutton and Barton in [14] describe curiosity as "balance between exploration and exploitation". The difference between their book and this paper is not about the terms used, but about timing. Sutton and Barto assume that the level of AI's curiosity will be selected by us during the AI creation process, while our assumption here is that we will create AI as a program with parameters and will define the curiosity parameter right before we launch the AI program.

There is another character trait mentioned in [14], and that is greed. There is a discount rate γ that determines whether the AI will greedily chase close rewards or be willing to ignore immediate gain for the sake of future successes. The coefficient γ is inversely proportional to greed. The larger γ is (closer to one), the less greedy the AI is.

Initial character

We have to divide the AI's character in an initial and current character. Pei Wang noted that a character can change in response to experience. For example, current curiosity can change as a

function of positive or negative experience. We have some initial curiosity which is part of our DNA (part of the AI program).

We can assume that current curiosity is a number, while initial curiosity is a number and a function (which determines how initial curiosity will change over time). The simplest case is to assume that initial curiosity is a parameter (an initial value and a constant function). That is, the simplest case is to assume that curiosity does not change.

The more interesting case would be to assume that a character changes and is dependent on time and experience. For example, with age we become less curious. Also, our curiosity is sensitive to our experience. The initial curiosity function should also tell us how strong will be the influence of experience and for how long will it continue before it loses momentum.

Persistence

There is another character trait which can be easily coded in the AI program – persistence.

When we describe the world, a major part of that description are algorithms. The description tells us how a certain action would change the world. Most actions do not happen in a single step and instead require the sequential execution of many steps, which we call an algorithm [3]. Algorithms require us to determine a cutoff point. We may continue for as long as it takes to achieve the desired result, or decide to quit at some point. How long we continue running the algorithm will depend on various things that can be estimated numerically. For example, *Importance* (how important is the goal for which we are executing this algorithm) and *Pressure* (the toll it takes on us and the resources which we may need to use for something else). Thus, the number of steps we will take before quitting is *Importance / Pressure* multiplied by some constant we would call *Persistence*.

$$Steps\_Before\_Quitting = Persistence \cdot Importance / Pressure$$

Instead of exact number of steps, this may the probability of making another step. In this case however the *Persistence* constant will different:

$$Probability\_of\_Continuing = Persistence \cdot Importance / Pressure$$

Imagine two worlds where there is gold buried under the ground and we have to dig to find it. In the first world the gold is buried deep in the ground, while in the second it is shallow. Let us have two AIs such that the first one is more persistent than the second one. The persistent AI will be more successful than the other AI in the first world. In the second world it will be the other way around. When the gold is deep the persistent AI will drill a few deep holes and find some gold, unlike the other AI which will drill many shallow holes to no avail. In the second world, the persistent AI will also find some gold, while the other one will find much more because in the second world it is better to drill shallow holes.

Of course, if we have infinite amount of time, based on experience AI can adjust its level of persistence. The assumption of infinite time is wrong because it inflicts many distortions.

Let us assume that success in the world is determined by the time it takes us to mine out the first piece of gold. In this assumption AI will not be able to adjust on the basis of experience, therefore much will depend on the level of its inborn persistence, if any.

The self-preservation instinct

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

Let us 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 stay 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 the life expectancy in humans is limited, 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 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 may not injure a human being or, through inaction, allow a human being 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 Africa, because it is the most populous continent (at present Africa accounts for only 19 % of the global population, but this share will probably rise to more than 50 % in the foreseeable future). However, the world now is not ruled by Africa, but by the most developed countries and mostly by the US although the US population is only 4 % of the world's population. If the world is ruled by the US, then "we" will be the US citizens. Probably the world will continue to be ruled by the developed countries, namely those that have been involved in the creation of AI. Who will be those countries is very important, because their profile will determine the profile of "us" – the future controllers of AI.

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.

We should not wonder in case if this young programmer elects to give AI control as a gift to a pop star he is secretly in love with. Should this happen, the prediction that one day the world will be run by a woman will come true.

Infantile creators

In whose hands have we entrusted the future of mankind? Have you seen a typical programmer? He is very young, antisocial and quite inapt in real life. Youth is not a vice because time very quickly corrects this problem. After all, why do we not allow minors to vote? The first astronauts were between 30 and 40 years old. Well, they were not that young – not because the space agencies could not send teenagers in space, but because adult people are more responsible.

The people you see on TV are not the real AI creators. These are the team leaders. They are people of higher age, social posture and responsibility.

A typical programmer is usually unmarried. Many of them even cannot find a girlfriend – not because they are ugly or poor, but due to their emotional immaturity which makes women reluctant to rest their life in the hands of a guy who thinks and behaves like a child.

Now think about this simple question: If a typical programmer is a person to whom you will not entrust your life or the life of your daughter, why would you entrust to that person the future of the entire mankind?

Emotionality

Should AI experience emotions? This is not about recognizing emotions. Of course, once AI is smart enough, it will be able to recognize human emotions. There are already programs which recognize emotions quite successfully. It is not about imitating an emotion either. AI would be able to imitate any human emotion if it wishes to. The question is whether we should give AI the ability to enter states that correspond to happiness and sadness?

Basically, excessive emotionality tends to be a negative property. We would prefer an employee, a civil servant or a judge to be impartial and not susceptible to emotions. Excessively emotional people are difficult to communicate with.

On the other hand, it would also be difficult for us to communicate with a being which is absolutely devoid of emotions. Very often AI will be in the role of our teacher and we will be in the role of its student. It is natural for the teacher to be happy when the student progresses and suffer when the student fails to understand the lesson. A typical student tries to make his teacher happy, and this is what drives him to work hard. The teacher may imitate happiness and frustration, but if the student knows that these are only imitations rather than pure emotions he will probably not believe them.

Since AI will not have a firm goal to strive for, it is natural to assume that there will be states such as happiness and sadness. Certainly, these states should be indicative rather than firm goals, otherwise they will morph into buttons (green and red).

Let us note that AI will communicate with many people at the same time, so it must not carry sadness from the conversation with one person to the conversation with another person. Common sense suggests that emotions should stay localized (in the current session).

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 another AI 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 is time for the new Manhattan Project. This project should involve everyone who cannot be excluded and keep 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.

Things are moving very fast and the news about the Stargate project came even before this paper was finished. It seems this paper had become pointless because we have already received what we are asking for. Actually, this is not exactly the case because Stargate aims to speed up the creation of AI, while here we call for the opposite (slow it down). Stargate is looking for the strongest and fastest racer that will lead the pack and step up the pace of the race. Our call is the opposite, namely to eliminate all small racers and leave only the one who will run the distance serenely and cross the finish line triumphantly with no rush. Whether AI arrives two months earlier or two months later is not important to us. What matters to us is what kind of character AI will be.

Obviously, free competition would accelerate, but if we want to decelerate then we need to ban the small players from the competition, and if this ban is to be respected all serious players need to be on board. The fact that serious countries such as the EU and China have been excluded from

the Stargate project means that the race will continue at a more reckless pace. That is, instead of pouring water, we will be adding petrol to the fire.

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.

Acknowledgements

We would like to recognize Pei Wang for his extremely valuable support. His comments were so essential that he should have been a coauthor of this paper. We would like to appreciate Moshe Vardi for his paper [9]. The idea of his paper is very close to ours. We attended his speech at [10] and after it we had a discussion with him in which he gave us many valuable insights, and we have reflected them in this paper. Many thanks to Vladimir Sotirov [13] for recommending the quote from Zhuang Zhou and René Thom. Special acknowledgements to Valentin Goranko [18] for his invaluable help.

References

[1] 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

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

[3] Dobrev D. (2023) Language for Description of Worlds. Part 2: The Sample World. Serdica Journal of Computing 17(1), 2023, pp. 17-54.

[4] Dobrev, D. & Popov, G. (2023). The First AI Created Will Be The Only AI Ever Created. *viXra:2311.0021*.

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

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

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

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

[9] 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

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

[11] 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.*

[12] Pavlov, R. (1990). Natural language processing systems. In I. Popchev & L. Dakovski (Eds.) ARTFICIAL INTELLIGENCE: problems and applications. *(1st ed., pp. 225–233). Technica Publishing House.* (This book is in Bulgarian language). https://knizhen-pazar.net/products/books/1772534

[13] Sotirov, V. (1999). Arithmetizations of Syllogistic à la Leibniz. *Journal of Applied Non-Classical Logics, 9(2–3), 387–405.* https://doi.org/10.1080/11663081.1999.10510975

[14] Sutton, R. & Barto, A. (2015). Reinforcement Learning: An Introduction. *A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England.*
https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

[15] Church, A. (1941) The Calculi of Lambda-Conversion. *Princeton: Princeton University Press.*

[16] Vardi, M.Y. (2025) Homo Ratiocinator (Reckoning Human). *Communications of the ACM, Volume 68, Issue 3, Page 5.*
http://dx.doi.org/10.1145/3714998

[17] Christian Hahm and Pei Wang (2025). NARS Genetic Encoding. *Technical Reports #23 of Temple AGI Team, February 2025.*
http://dx.doi.org/10.13140/RG.2.2.13424.78089

[18] Goranko, V. (2023) Logics for Strategic Reasoning of Socially Interacting Rational Agents: An Overview and Perspectives. *Logics, 1(1), 4-35.*
https://doi.org/10.3390/logics1010003

[19] Kai, De. (2025). Raising AI: An Essential Guide to Parenting Our Future. *MIT Press, 2025, ISBN: 9780262049764.*

[20] Hinton, Geoffrey. (2025) We need to program AI to have 'maternal instincts' towards humanity. Ai4 conference, Las Vegas.
https://ai4.io/vegas/
https://dunyanews.tv/en/Technology/901146-godfather-of-ai-reveals-the-only-way-humanity-can-survive-superintel