Taking control over AI before it starts controlling us

Will 2030 be a tipping point of losing control over AI?

Tony Czarnecki, Managing Partner, Sustensis, www.sustensis.co.uk

The late physicist Stephen Hawking, Microsoft founder Bill Gates and SpaceX founder Elon Musk have all expressed concerns about the possibility that Artificial Intelligence (AI) could evolve to the point that humans could no longer control it, with Hawking theorizing that this could "spell the end of the human race"[1].

Elon Musk has been urging governments on numerous occasions to take steps to regulate the technology before it is too late. At the bipartisan National Governors Association meeting in July 2017 he said: "AI is a fundamental existential risk for human civilization, and I don't think people fully appreciate that." He also added that based on what he had seen, AI is the scariest problem. Musk told the governors that AI calls for precautionary, proactive government intervention: "I think by the time we are reactive in AI regulation, it's too late".[2]

If we consider that 99% of all species have disappeared[3], then why should we be an exception to the Fermi's Paradox, of which one of the explanations is that no civilisation has contacted us, because once they had achieved a certain level of technological advancement, they destroyed themselves. So, if we want to avoid an extinction, we must mitigate existential risks such as climate change, pandemics, nanotechnology, global nuclear wars and most importantly the threat arising from developing a hostile Superintelligence. It is this threat that could be the most imminent sand most dangerous of all because it could annihilate the human species, possibly within the next few decades. But let me first describe what I mean by Superintelligence.

What is Superintelligence?

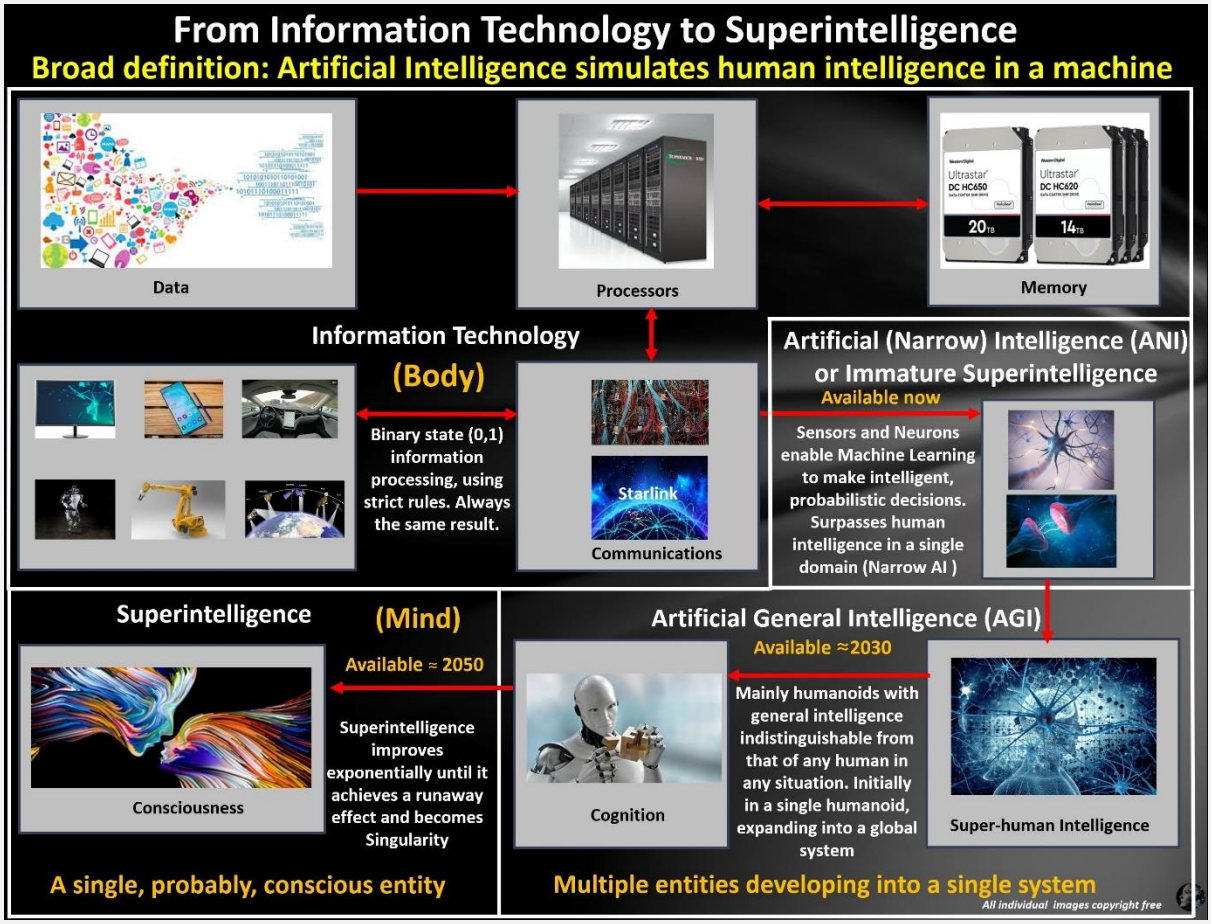
For an average person just the term AI may be quite confusing, embracing all aspects of what seems to be 'unnatural'. The difficulty for an average person starts with differentiating Information Technology (IT) and AI. IT processes information based on strictly defined rules, needing all the required input data. But AI may produce results based on partial input data, learning for example, from experience. Therefore, the same input data may not always produce the same output. It's the learning experience, which makes some humanoid robots resembling human behaviour - they make errors but on average, like in autonomous driving, fewer than humans.

What we have now, are individual, relatively unsophisticated agents, chatbots or robots. This is what is generally called **Artificial Narrow Intelligence (ANI)** and which is a kind of an immature AI. It already exceeds human intelligence and capabilities in certain areas, like in all games, including poker, which requires some intuition, or face recognition. Today, it only exceeds human intelligence and capabilities in certain areas, like in all games, including poker, which requires some intuition, or face recognition. However, such an AI is ignorant in all other areas and that's why it may be called Immature. It does not have any cognition, i.e., it would not know that it cannot walk through the wall, although this is rapidly changing. Still, this immature AI may already be very dangerous on a global scale within the next few years. For example, we may soon have millions of humanoid robots, such as an advanced Optimus, courtesy of Elon Musk, costing about $20,000. Such humanoid robots will be capable of carrying out most physical tasks around the house or in a factory, communicating verbally with humans. They may also be cognitive, i.e. be aware of how people live, move, and what it itself is capable of, e.g. jumping over a 3m fence etc. They will also be connected to the Internet. If by accidental self-learning or malicious design they self-connect to each other, they could over time plot a global destructive action of potentially disastrous consequences, like launching nuclear weapons. Moreover, unless there's is shortly a global banning legislation, some most advanced companies, like Amazon, may create global AI networks, operating from a central hub. Such a global AI system could create, if misused, a near existential risk. All this proves that AI does not have to be fully matured, to become an existential threat. In summary, current, Immature AI may very soon become an existential threat.

But within a decade we may have an **Artificial General Intelligence (AGI)**, which will exceed the intelligence and capabilities of any humans in **all** areas. It is mostly assumed that such an AGI will be only embedded in a single humanoid robot. This may be a general practice. However, it does not exclude that we may then have a network of globally connected thousands of such AGI humanoids controlling millions of other less intelligent robots and trillions of sensors. The consequences of such a network, which is highly likely to be outside of human control, might be potentially an existential threat. Imagine that no country will be able to control it, as no country has been able to control the Internet on a global scale for over two decades.

Now we need to explain what **Superintelligence** is. It may not be a big problem if an average person confuses Superintelligence with a Terminator-type robot. But it may be deeply troubling if that includes politicians. After all, these are the people whom we must convince that there is little time left before we may lose control over maturing AI. The media may be responsible for much of that misunderstanding by trivializing AI. However, it is also the result of poor, very narrow education. So, here is how I define Superintelligence.

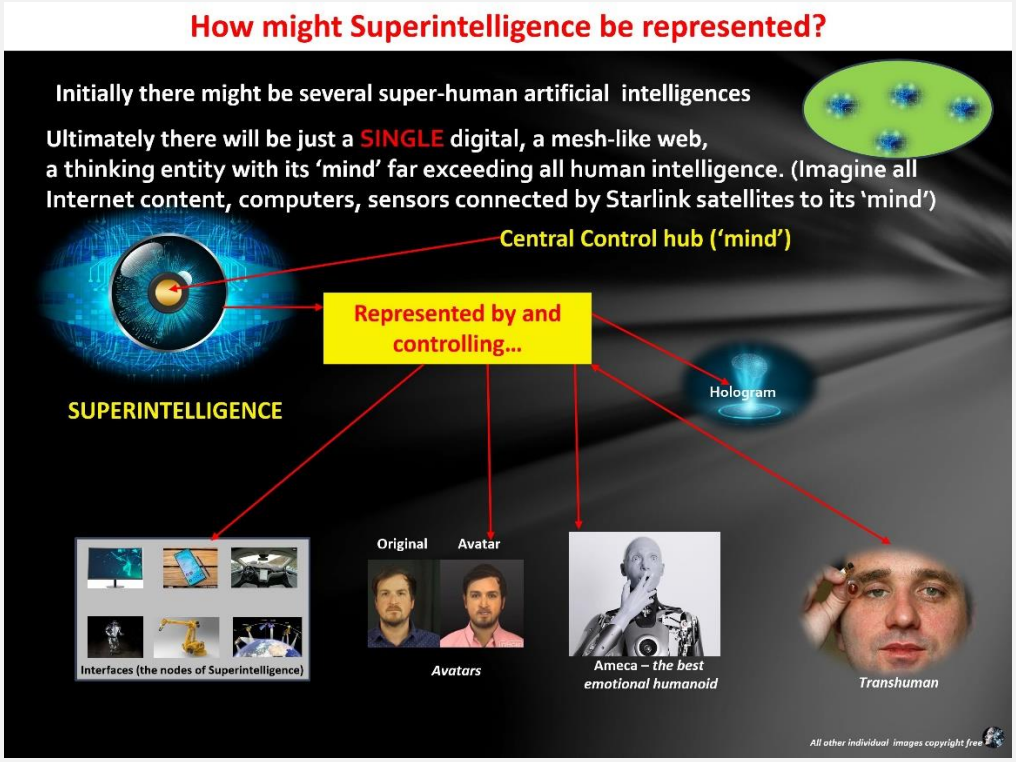
Superintelligence is a single entity, with its own mind and goals, immeasurably exceeding all human intelligence. Its body consists of various elements such as data, processors, memory, interfaces, communications, sensors, including artificial morphic neurons. We already have them. However, currently all these building blocks are thousands of times slower and could not even support AGI, not to mention a mature Superintelligence.



Neither does current Immature AI have a mind. That would require its intelligence to acquire full cognition – an experiential knowledge and awareness of the world. Once it achieves that, it may then gradually turn into a conscious entity. However, there is no agreement among AI researchers whether such an advanced intelligent agent must be conscious before it becomes superintelligent.

In the view of most AI scientists, Superintelligence will emerge by AGI's self-improvement over the years until it achieves the **Singularity** point sometimes called ' the runaway point'. At that time, humans will be under its total control and incapable of understanding the rationale behind is decisions. That alone will be an existential threat for humans because we will lose control over our own destiny. Whether such a mature Superintelligence becomes a threat to a human species depends largely on how, or if at all, it was nurtured in line with human values before we will have lost control over it. If Superintelligence has slightly misaligned objectives or values with those that we share, it may become hostile towards humans. Therefore, we must protect ourselves from such a scenario becoming a reality.

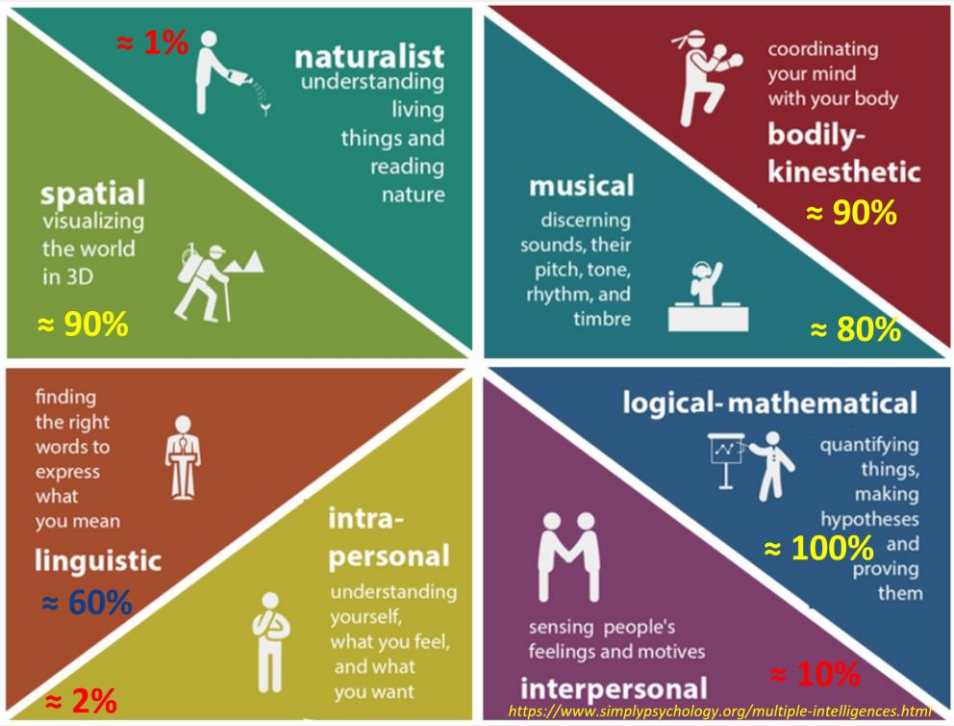
Superintelligence will present itself to humans in various ways and through numerous simultaneous representations in any part of the planet.



It will represent itself as avatars, holograms, or as emotional humanoids, such as an advanced AMECA or Optimus robots. Finally, it will also be linked to conscious Transhumans, who play a key role in how I imagine humans may most effectively control Superintelligence.

Humans versus Artificial Intelligence today

Howard Gardner has identified 8 human intelligences[4]. These are: Linguistic, Logical/Mathematical, Spatial, Bodily-Kinaesthetic, Musical, Interpersonal, Intrapersonal, and Naturalist. In at least four of these - Bodily-Kinaesthetic, Logical/Mathematical, Musical and Spatial, AI already exceeds humans.

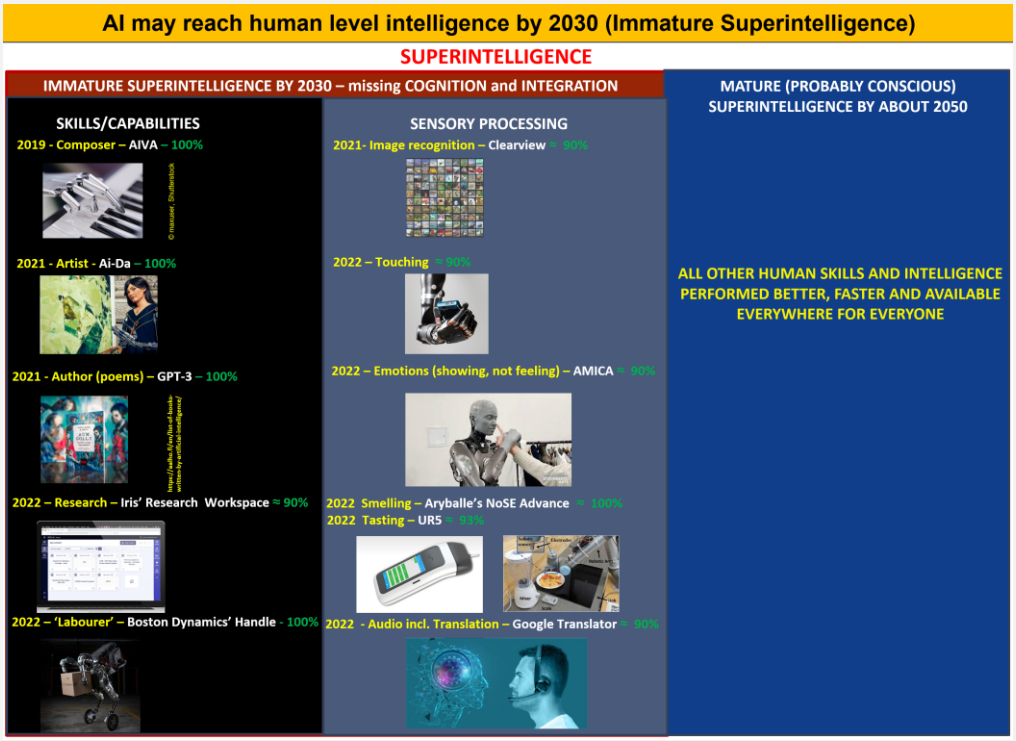


How humans Compare against Artificial Intelligence today

I have estimated how well a narrow AI's intelligence (in a single area of competence) currently matches human intelligence in each of the eight intelligences. In Linguistic Intelligence, it is vastly superior to humans (e.g. the number of languages it is able to translate simultaneously with fewer and fewer errors). However, humans are still immensely superior in Interpersonal, Intrapersonal (understanding yourself, feelings etc), and Naturalist areas. That is very closely related to cognition, the most difficult domain for AI to learn. However, the pace of progress in AI measured by the number of significant breakthroughs, which impact the entire industry, has been truly astounding. Here are some of the most significant developments over the last 20 years:

- 2006 – Convoluted Neural Networks For Image recognition *(Fei Fei Li)*
- 2016- AlphaGo – Supervised ML, Monte Carlo, Tree Search + neural networks *(DeepMind)*
- 2017- AlphaZero – Unsupervised ML *(DeepMind)*
- 2017- Tokenized Self-Attention for NLP - Generative Pre-trained Transformers *(GoogleBrain*
- 2021- AlphaFold – Graph Transformers (graphs as tokens) predicting 3D protein folding *(GoogleBrain)*
- 2022 (March) – Artificial neurons based on Photonic quantum memristors *(University of Vienna)*
- 2022 (2 April) – White Box – Self-explainable AI, Hybrid AI *(French Nukka lab)*
- 2022 (4 April) – PaLM, Pathways Language Model, NLP with context and reasoning (*Google Research*)
- 2022 (11 May) – LaMBDA –multi-modal AI agent – can also controlling robots with NLP (Google)

These breakthroughs have helped AI researchers to apply them in various domains, as illustrated below, in which AI's skills quite often vastly exceed human level intelligence and capabilities. That has also been reflected in the sensory processing, which may be crucial for developing AI's cognitive capabilities.



This does not include yet the impact of progress in AI-related hardware. For example, the number of tokens (1,000 tokens is an approximate equivalent of 1 human neuron) has been rising faster than exponentially over the last 4 years, increasing from 300M (BERT in 2017) to PALM - 650B in 2022 and 1.6 trillion (Wu Dao 2.0 in 2022). With the current pace of development, the number of neuron-like tokens should equal 86B neurons in a human brain by 2024, which would require about 86 trillion tokens.

However, if we include the super-exponential pace of development in synthetic neurons, based on memristors, and quantum computing, we can expect even faster acceleration of AI capabilities. We have already created AI, which can exceed human level intelligence and skills in some areas while being incompetent in the tasks, which every toddler can solve. This relentless progress in AI capabilities may lead to humans' losing control over the AI's self-learning capabilities, directly impacting our ability to control its goals. Once this tipping point is reached, the consequences for our civilisation and indeed for the future of a human species will be enormous. Therefore, AI scientists should at least agree on what might be the most likely date when humans may lose control over AI.

Governments must prepare for Artificial Intelligence being out of human control by 2030

The first problem we face when attempting to control AI is that we need to convince the public and most importantly, the world leaders, that such an invisible threat is real. One may call a maturing Superintelligence '**an invisible enemy**,' assuming it turns out to be hostile towards humans, similarly as the Covid-19 pandemic was. Calling Covid an invisible enemy was an excuse used by governments that it was not possible to see the threat as coming, hence they were not responsible for the consequences. Governments seldom see that spending money now to minimize the risk of potential future disasters is an insurance policy. The implications of such short-termism for controlling AI development are profound. In the worst-case scenario, given an immense power of Superintelligence, it would be enough for such an agent to make a single error to cause humans' extinction.

The second problem is that not many AI experts are willing to say when Superintelligence is most likely to emerge. That allows politicians to dismiss any calls for taking serious steps towards controlling Superintelligence, saying it is hundreds of years away, so we do not have to worry about it now. The predictions by AI scientists and leading practitioners are generally vague without a clear definition of what is meant by Superintelligence. Ray Kurzweil is perhaps an exception. Being one of the most reliable futurists, he says that a mature Superintelligence may emerge by 2045[5]. At the AI conference in 1995, the participants estimated that it may emerge in two hundred years[6]. But four averaged surveys of 995 AI professionals published in February 2022 indicate that the most likely date for a mature Superintelligence is about 2060, just 15 years after the Kurzweil's prediction[7]. In any case, if his predictions are correct, most people living today will be in contact with Superintelligence, which may be our last invention, as the British mathematician I. J. Good observed in 1966.

Perhaps even more important than the time by when Superintelligence emerges, is an approximate time when humans may lose control over AI, operating as a global system. Here again, AI scientists and top AI practitioners prefer not to specify such time, using instead more elusive terms like 'in a few decades or so.' **However, without setting a highly probably time when we may lose control over AI, the world leaders will not feel obliged to discuss this existential risk for humans, which such a momentous event may trigger.** Therefore, those who see that problem, should be bold enough to spell out the most likely time and justify it. Ray Kurzweil is again an exception here, saying in June 2014: "My timeline is computers will be at a human level, such as you can have a human relationship with them, 15 years from now,"[8] i.e., by 2029. Since then, he has been sticking to that date.

The loss of control over AI self-improvement can be compared in some way to the loss of control over the operation of the Internet. No country can switch off the Internet globally. Doing so, would be theoretically possible but it would mean a civilisational collapse, but even then, such a switch off may still be incomplete. We may soon face a similar situation with a globally networked AI, controlling billions of sensors and millions of robots. A desktop computer power will increase by about 1,000 times by 2030 (from 2014), reaching the intelligence level of an average human, if measured by the no. of neurons, and vastly exceeding our memory and processing power (Ray Kurzweil's reasoning). But that does not include the progress in neuromorphic neurons, quantum computing and several other related areas, which will immensely increase the capabilities of such an intelligence.

Therefore, we should take 2030, as the most likely date by which humans may lose an effective control over AI. **2030 is thus the AI's tipping point,** likely to be reached at the same time as the tipping point for the global warming and what I would call a Global Disorder if the UN's Sustainable Development Goals are not met by that date. **Therefore, we may face three civilisational tipping points at around 2030.** AI may then have an intelligence of an ant, but immense destructive powers, resulting either from erroneous design or from a malicious intention. There may be several such agents by the end of this decade, which might even fight each other, if deployed by some psychopathic dictators, hoping to achieve AI Supremacy and use it to conquer the world.

Unfortunately, instead of serious discussions on the consequences of losing very soon the control over self-learning AI, conferences on AI threats are concerned with face recognition impacting our privacy, which are relatively trivial aspects of AI control. By focusing on these issues, the real dangers, to which we may be exposed, are hidden. Revealing them would require putting stricter control on large companies developing AI, similarly as it happens now in the carbon economy, where those companies' profits are reduced. Furthermore, deep interests to protect national industries make an effective control of AI development very difficult.

It is not so important, who specifies a concrete date but that such a date is widely publicised and supported by eminent AI scientists. For example, it was argued for decades that a potential global warming tipping point was far away, so nothing was done. Only when at the Paris conference in 2015 and at COP26 in Glasgow in 2021, a maximum 1.5C temperature rise was set, as recommended by the International Panel on Climate Change (IPCC),

concrete global action was finally agreed. But COP25 in Paris also specified a pivotal date 2030, as a tipping point, beyond which we may lose the battle for controlling climate change[9].

AI has of course a much wider and more imminent impact than global warming on our species' survival, covering every domain of human life from a peaceful use to military applications. Therefore, it is even more important that decisive global action on AI control is put in motion as soon as possible. The best way forward seems to be to follow the IPCC and COP26 examples and call a global conference on controlling AI. The prime goal of the Conference should be:

1. To declare 2030 as a tipping point for AI
2. Create an international agency to monitor continuously the progress of AI development
3. Agree thresholds, which should not be passed by the most advanced AI systems before 2030, such as:

    - The creation of the first simple cognitive AI Agent
    - Humanoid robots surpassing the performance of a human brain and running complex AI algorithms
    - The number of artificial neuromorphic neurons exceeding the no. of neurons in a human brain
    - Incidents when AI network of globally connected robots goes out of control leading to a global chaos.

If AI does not cross those thresholds before 2030, it will give us more time for preparing the transition to the period when AI will start controlling us. The date 2030 is only an example, although like with climate change, it seems to be most likely. There is a saying 'What is not measured is not done' and just declaring such thresholds may be enough to trigger a global action.

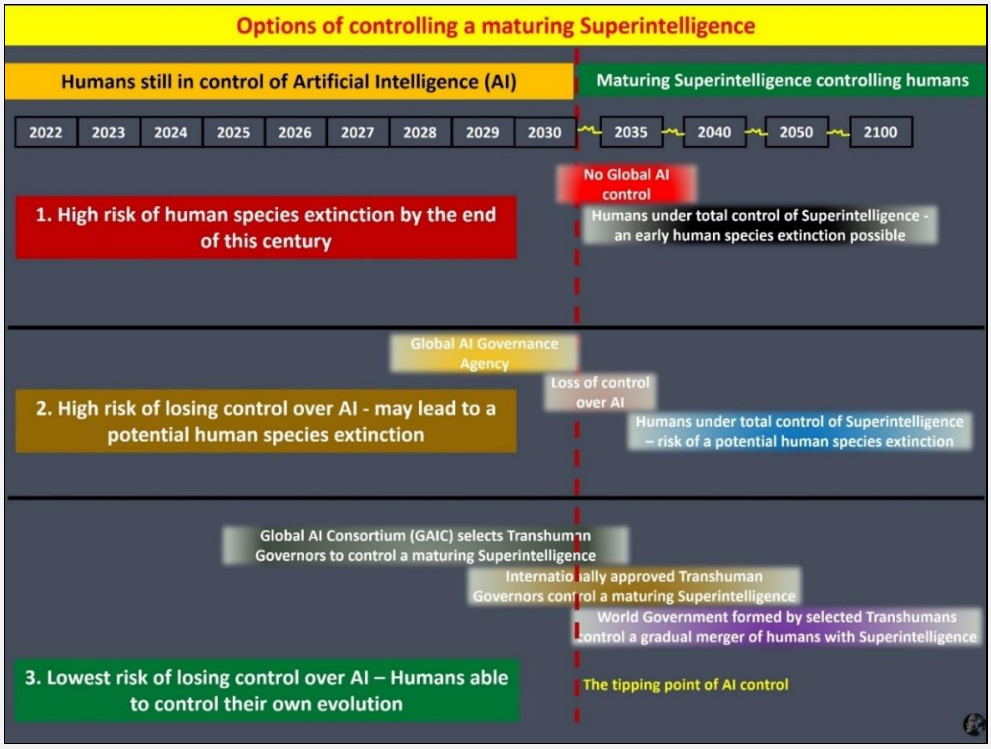
So, who would set up such an Agency? It should of course be the United Nations by default. But its Security Council, which as the current war in Ukraine shows, is unable to implement any global order. That does not mean that the UN has not been trying to do its best in many vital areas, such as nuclear disarmament, biological weapons, health, education, or culture. It has also initiated some research about the AI control and put forward proposals at various UN events under the auspices of its Interregional Crime and Justice Research Institute (UNICRI). For example, in July 2018, in Singapore, it organised the first global meeting on AI and robotics for law enforcement, co-organized with INTERPOL. In April 2019 a Joint UNICRI-INTERPOL report on "AI and Robotics" was published. The problem is that this proposal has remained just that – a proposal.

Therefore, realistically we cannot count on the UN being capable of creating such an Agency soon enough to control AI effectively, before it is too late. Therefore, despite all the problems the European Union has, it is probably the most experienced organization, which might take up this challenge. This might also include the European Political Community, created in October 2022 if a more ambitious conversion of the EU into a European Federation is further delayed.

Such an agency should start a global comprehensive control over AI by about 2025 if it is to delay the pace of AI advancement, which may seem impossible. However, it should not be a significant challenge if we consider that the UN was created in just two years and NATO within one year. The problem is that although we now live in even more dangerous times than ever, the post-war generations behave as the existential threats had vanished. That is why governments take decades to sign any global initiatives. The best example are the IPCC-led efforts, to agree a global action on reversing the climate change, which took 23 years from the Rio conference to Paris COP21 conference. But we now face not just an extinction arising from a global nuclear war. There are several more existential threats, of which a development either by a malicious design or by error of a hostile AI is highly likely within a decade. Although the setting up of AI controlling agency by about 2025 seems almost impossible, we should still try to organize a conference, which would set it up as soon as possible.

Options for implementing a continuous, global control of a maturing Superintelligence

It will be more dangerous for humans if AI surpasses the earlier mentioned thresholds aimed at delaying its current very fast capability improvement than the global temperature increase above 1.5C. Once these thresholds have been passed, which are the warning signs about humans losing control over AI, then the start of human species' evolution or extinction will begin. Whatever happens we already have no other option that to evolve. If we maintain a proper control over AI, we will succeed in delivering a friendly AI and reach the world of unimaginable abundance and opportunities. Depending on the speed of implementation of the agreed actions by governments, our civilization may progress according to one of the three scenarios (assuming 2030 as a tipping point):



Option 1: No Global AI Control

If we do nothing or have an ineffective AI control, humans will be progressively under a greater control of a maturing Superintelligence. If it becomes hostile to humans, it may trigger an early human species' extinction.

Option 2: Global AI Governance Agency (GAIGA) controlling Superintelligence

The second option is to create an agency e.g., Global AI Governance Agency (GAIGA), which would control the emerging Superintelligence. I propose to model the operation of such an agency on the International Atomic Energy Authority (IAEA) in Vienna. IAEA was created in 1957 in response to deep fears but also hopes regarding the use of nuclear technology. The Agency was set up as one of the UN's organizations.

It should be responsible for ensuring that the earlier mentioned thresholds related to AI control are not exceeded before 2030. The agency should maintain continuous and comprehensive control over AI chips, weaponized AI, robots, neural networks, brain implants etc. That is of course still based on the assumption that the pace of AI development runs at the current rate and no significant invention accelerates the process of a maturing AI even further.

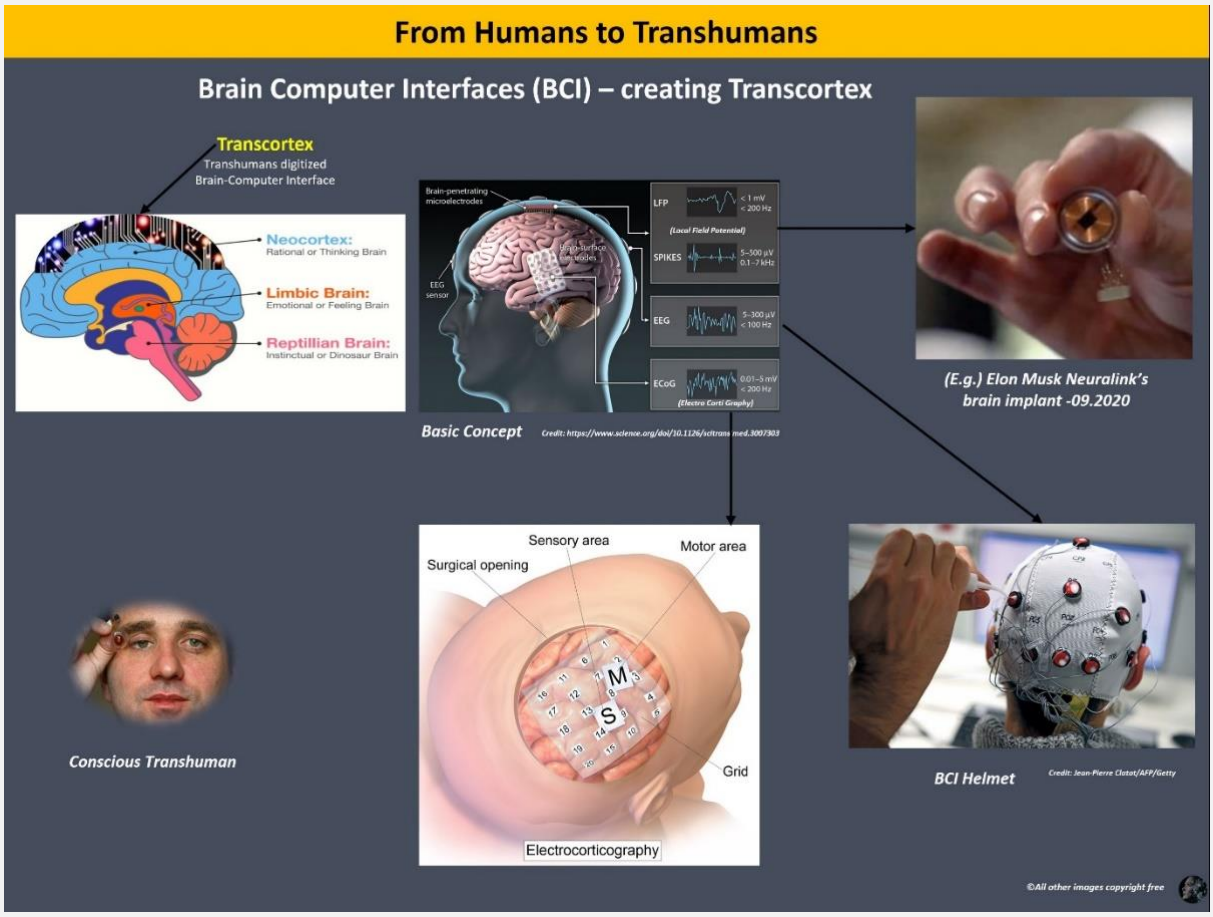
How could then GAIGA control the process of a maturing Superintelligence? Can we control it effectively at all? Without going into details, Nick Bostrom mentioned in his seminal book 'Superintelligence' that none of the capability control and other methods can guarantee an effective AI control. Even Stuart Russell's recently proposed 'the human preference method' does not guarantee 100% control either. But we could improve that control significantly. In my recent book, 'Becoming a butterfly'[10], I have presented a framework for a maturing AI, where some of these AI controls are applied simultaneously.

The first step in the Framework is to instil Universal Values of Humanity as an ethics code in a form of a Master chip controlling the advanced AI agents. Such chip would be distributed under license and implanted in all advanced AI agent's 'brains (control hubs). The Universal Values of Humanity, also incorporating Human Responsibilities, might be derived from the UN Declaration of Human Rights, combined with the EU Convention on Human Rights and other relevant, more recent legal documents in this area. Irrespective of which existing international legislations are used as an input, the new Declaration of Human Rights would have to be universally approved if these values are to be truly universal. That may be impossible (China or Russia, will not accept them). But even if such values do not become genuinely universal, they should nevertheless be binding.

The second step in the Framework is nurturing AI as a child in real environment. There are already some good examples in this area of AI control, like the introduction of humanoid robots into Japanese care homes. The top AI companies are already collecting AI experiences in human preferences. For instance, some legal firms can use the GPT-3 agent to prepare cases for presentations at the court[11]. Tesla has been routinely gathering the 'experiences' of its cars and then uploading those cars with the preferred actions to avoid collisions[12]. AI-controlled robots are used in laboratories, in the office, and in factories. Such a Maturing Framework, monitored by a Global AI Governance Agency, may increase the chances of delivering a benevolent Superintelligence. Although these controls themselves cannot ensure failsafe result, they would still be better than no control at all.

Option 3: How could Transhumans Control Superintelligence?

Simultaneously, as a fallback option, we should create by about 2025, an **interim** organization, which would provide continuous, although imperfect control. This time horizon implies that it would be an independent non-governmental Agency. Elon Musk's metaphor "If you can't beat them - join them!"[13] perhaps best describes the third option of controlling Superintelligence. That might be delivered by creating Brain-Computer-Interfaces (BCI) for the top AI scientists. We would thus create the first Transhumans, who would control the maturing AI from 'inside'. Our brain consists of three brains – Reptilian Brain, Limbic Brain and Neocortex with nearly 70 distinct functional areas. We have currently at least three basic methods to read the brain's electromagnetic waves.

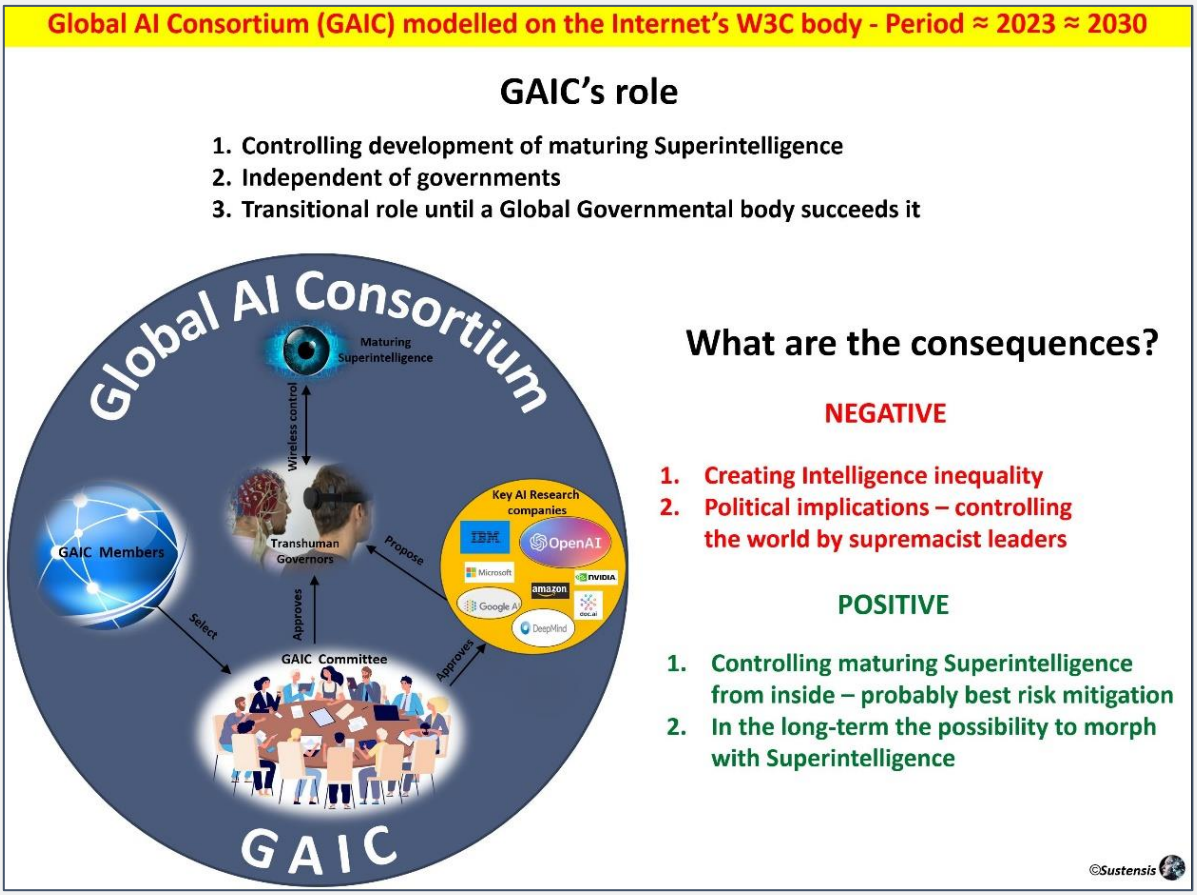


The first one reads changes in the Local Field Potential (LFP). An example could be Elon Musk's Neuralink brain implants. The second method uses Electro Encephalogram (EEG) embedded in a special helmet to read the brain's activities. This well-tried method is being used to read people's thoughts and to instruct the brain on how to manipulate things, like typing with thought alone[14]. Finally, we can also use Electro-Corti-Graphy (ECOG). This method uses special neuromorphic chips implanted as a digital interface on the surface of the brain.

Within the next few years all these methods will enable the creation of cognitive, most likely non-invasive, BCI devices (e.g. like wearable helmets) for the first Transhumans. There are already about 10,000 Transhumans in Sweden alone, who have implanted chips in their hands allowing them to use such chips for verifying their identity instead of using passwords for banking, purchases or passing security gates. But BCI devices, which will make a real difference will not be those controlling limbs, organs or curing diseases. They will of course immensely help millions of people. But for population at large over the coming decades a real difference that these BCI devices will make will be to enhance human's mental capabilities, i.e., memory, processing speed, decision making and sensory processing. Those who wonder how it could be possible, should consider that we are already partly Transhumans. Our smart phones give us enormous extra intelligence, which we could not dream of even a few years ago. The only difference between the future Transhumans and us is that our external intelligence (e.g. a smartphone) communicates with the brain via our eyesight, hearing and touch. Since BCIs are digital devices, their capability will increase nearly exponentially. For example, the brain electrodes' density should increase by the end of this decade by about fifty times to what it was in 2020.

Transhumans' brain with a wireless access to a much more advanced AI knowledge base than for example the most advanced today's Natural Language Processors (NLP) such as LaMBDA or PALM, will be able to process and store in the cloud whatever it decides to remember, and then retrieve it instantaneously. Their cognitive capabilities will increase so much that they will become the most intelligent and capable people. They may thus become invaluable, if selected by international bodies, such as the UN, to help resolve civilizational problems.

Initially, the first Transhumans would be the top AI developers. They should be from the outset under the control of a global organization. However, as mentioned earlier, this will almost certainly not happen on time. In the interim, we need a transitionary body, which should start operating in the next 2-3 years at the latest. Therefore, I propose to create a **Global AI Consortium (GAIC)**, which would be modelled on the Internet's W3C Consortium. In over 30 years of its operations, W3C has proved how well such a body, independent from governments, could function, maintaining global control over all key activities of the Internet[15].



GAIC would operate in a similar way as W3C. Its members will usually be large AI companies. GAIC members' Assembly would select the GAIC committee. Its role would be to select from the proposed candidates Transhuman Governors. These would be key AI researchers and developers, neuroscientists, philosophers, psychologists, or other scientists. They will either have fitted a neuromorphic chip, or more likely, wear a special helmet enabling them wireless thought transmission to external data and processing facilities. But most importantly they will be linked to a hub of a maturing Superintelligence. Such a hub will be the main decision centre of the evolving and constantly improving Superintelligence, like the main control module on the Android operating system on smart phones. It will progressively become a global AI network overlaid over the Internet, including satellite networks.

Once the Transhuman Governors have been approved by GAIC, they will be progressively fusing wirelessly more of their brain cognitive functions with the control centre of the maturing Superintelligence. Such a 'Master Switch' will be wirelessly controlled by all networked Transhuman Governors, who will be connected to each other. Upgrading the software or authorising the execution of significant decisions would require the consent of the majority of the connected Transhuman Governors. Probably no verification of their decisions will be needed since they will be able to read each other thoughts of the part of their brains integrated with the Master Switch using for example the Blockchain method. They may have to sacrifice their privacy - their contribution to the increased safety of all humans. Other important functions of Superintelligence might also be controlled in this way.

In a few decades, the body of Transhumans will become increasingly non-biological and their brain more digitally integrated with the emerging Superintelligence. By the end of this century, the whole brain of the willing Transhumans may be digitized and fully fused with a purely digital Superintelligence. Unless there are some physical obstacles e.g., related to porting consciousness onto digital chips, an entirely new, non-organic species will emerge, which might then be called Posthumans. (NB: such a progressive mind uploading through BCI may be more realistic than all-at-once copying of human mind into its digital equivalent.)

What might be the consequences of giving Transhuman Governors so much power? There are at least two negative ones. The first one is the superiority of the Transhumans' intelligence since their cognitive capabilities will be significantly extended in just a few years. Their memories, processing power and the speed of their decision making might be perhaps even a thousand times faster than the top human experts. They will be above anyone's capabilities in any area of science or knowledge. In relative terms, they will be almost omniscient. But political implications may be even more critical. Since it will be impossible to control all BCI implants (they may be produced not just in the USA but also in China or Russia), very rich individuals and some political leaders will almost certainly get them for themselves. Thus potentially, any dictator may become a Transhuman, although he will not be connected to the 'approved' Superintelligence. This can be achieved in a few years' time and people might not even notice it.

However, there are also some positive consequences. The first one is the ability to control a maturing Superintelligence from inside at a hardware level, which might be the most resilient method of AI control. Secondly, the Superintelligence controlled by Transhuman Governors, will deliver unimaginable benefits to humans, creating the world of abundance.

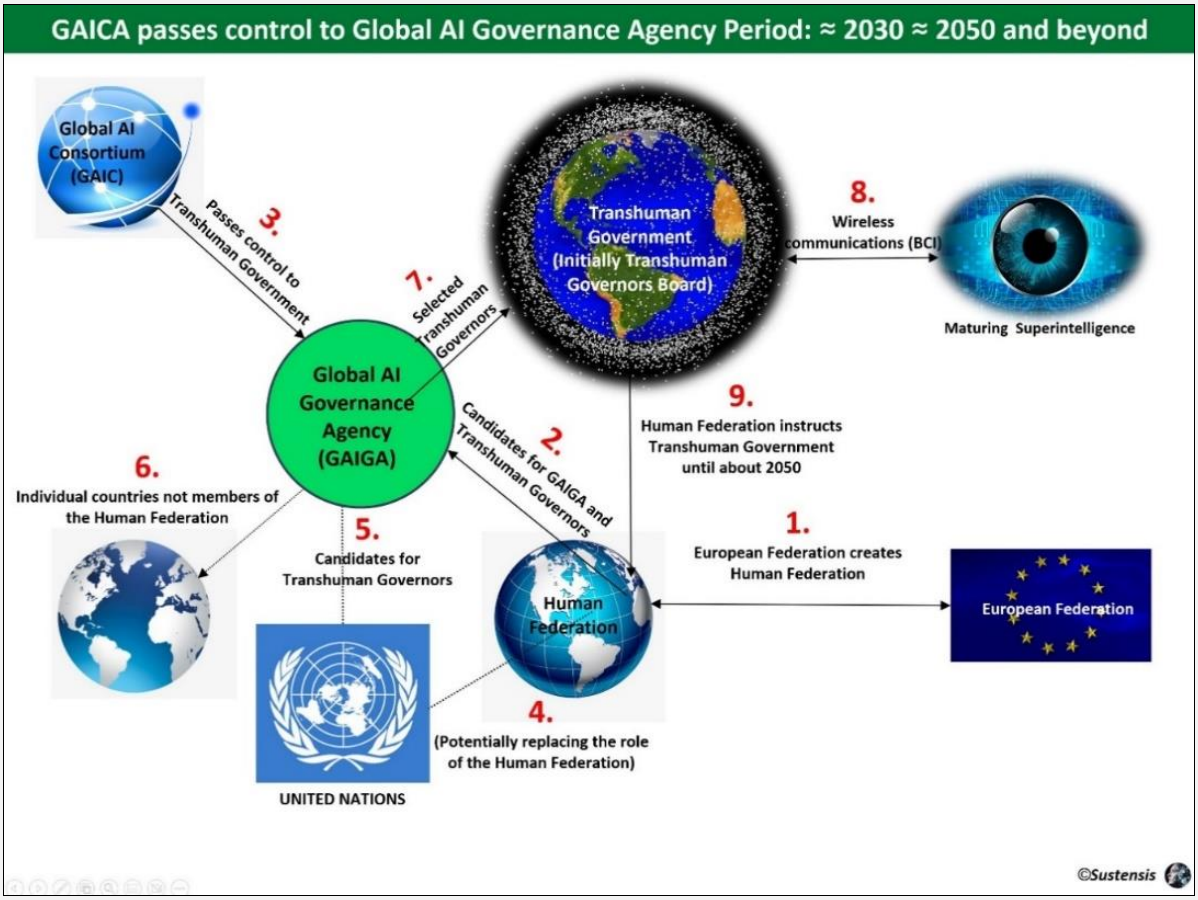
It is difficult to say for how long GAIC would be the sole controller of Transhuman Governors, and to which organization it will pass that control. In an ideal situation, it should be the United Nations playing the role of the World Government. But, as mentioned earlier, the UN is unlikely to be capable of an effective control. However, if UN somehow gets the real global powers of governing in the next decade, it would of course be the natural choice for this organization to select the Transhuman Governors.

By about 2030 we may finally have a de facto World Government created from the expanding European Federation or the European Political Community. At this stage, in the absence of an effective control by the UN the control of Transhuman Governors would most likely pass from GAIC to an international organization, such as the earlier mentioned Global AI Governance Agency (GAIGA). This will, at least in principle, democratize the decisions made by Transhuman Governors on behalf of all humans. However, in practice such a control of Transhuman Governors would be largely symbolic, since their decisions might be far better than those made by the most capable humans. It will be in our own interest to let them decide what is best for us.

The only meaningful decisions of the Human Federation would be the selection and deselection of Transhuman Governors. However, that should not be based on political grounds. Therefore, it should be GAIGA rather than the World Government, which should take the ultimate decisions on selecting Transhuman Governors. To reduce the risk of biased decisions, the number of Transhuman Governors would probably extend to thousands, or even tens of thousands by the end of this decade, each having the same rights.

It is also possible that our civilizations will not be able to form a Human Federation and its executive arm, the World Government, at all. In such case, Transhuman Governors, gradually more tightly fused with Superintelligence, would play the role of an actual World Government anyway. Any attempt to remove such Transhumans by force would only made matters worse, since Superintelligence would be then out of any control.

A Transhuman Government selected and not elected would be a big dilemma for our civilisation. Democracy as we know it would only be relevant at a lower level of decision making, perhaps till the middle of this century. From about 2050 the Human Federation will be there to implement the decisions of the Transhuman Government, over which it may have no longer any control. Human species' fate will be in the hands of Transhuman Governors who by the end of this century may be completely digitized becoming Superintelligence themselves. This will pave the way for more humans morphing with Superintelligence and thus starting the evolution of the human species on a grand scale.



Superintelligence will deliver unimaginable benefits to all people. As it matures, the first change people may notice in the next decade, if this scenario comes to fruition, is that there will simply be no wars. That on its own will increase the wealth growth. Productivity will soar, perhaps doubling the current growth rate of the world's annual GDP. That may pay for rebalancing the average income worldwide and for regenerative medicine, which may very quickly extend a healthy life span by decades.

Superintelligence will enable individualized AI-assisted education as well as facilitate personal fulfilment. People will be able to accomplish most of their wishes, such as developing skills in the arts, music, literature, climbing mountains, and do whatever else interests them. All existential risks, including climate change will be minimized or eliminated by the maturing Superintelligence.

London, October 2022

****

You can find more detailed information on the subjects covered in this article on Sustensis website: www.sustensis.co.uk, which is also supported by dozens of videos.

About the Author

Tony Czarnecki is a futurist, a member of the Chatham House and the Managing Partner of Sustensis, London a Think Tank focused on Humanity's transition to coexistence with Superintelligence (www.sustensis.co.uk). His ideas have been presented in his three books of the 'Posthumans' series: "Federate to Survive!", "Democracy for a Human Federation", and the latest one – "Becoming a Butterfly".

References:

[1] BBC News, https://www.bbc.co.uk/news/technology-30290540 , 2/12/2014,
[2] Camila Domonoske, Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk', 17/7/2017
[3] D. Jablonsky, Nature, Volume 427, Issue 6975, pp. 589, 2004
[4] Howard Gardner: 'Multiple intelligences and related educational topics' 2013, https://howardgardner01.files.wordpress.com/2012/06/faq_march2013.pdf
[5] Ray Kurzweil, in an interview with 'Futurism', https://futurism.com/kurzweil-claims-that-the-singularity-will-happen-by-2045, 10/05/2017
[6] The seventh conference on innovative applications of artificial intelligence, https://aaai.org/Press/Proceedings/iaai95.php, 21/8/1995
[7] Cem Dilmegani, When will singularity happen? 995 experts' opinions on AGI, https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/, 3/2/2022
[8] Ray Kurzweil, in an interview with NBC: https://www.nbcnews.com/tech/innovation/top-google-engineer-says-computers-will-be-humans-2029-n128926, 11/6/2014
[9] COP26 – Together for our planet: https://www.un.org/en/climatechange/cop26
[10] Tony Czarnecki, "Becoming a Butterfly: Extinction or Evolution? Will Humans Survive Beyond 2050?", London, 2021.
[11] GPT-3 in legal tech, https://www.jdsupra.com/legalnews/gpt-3-in-legal-tech-insights-from-the-3183642/, 15/12/2021
[12] Vikram Singh: Tesla: A data driven future, https://digital.hbs.edu/platform-digit/submission/tesla-a-data-driven-future/, 23/3/2021
[13] Elon Musk, Twitter, @elonmusk, 9/7/2020
[14] Shelley Fan, A New Brain Implant Turns Thoughts Into Text, https://singularityhub.com/2021/05/18/a-new-brain-implant-turns-thoughts-into-text-with-90-percent-accuracy/ 18/5/2021
[15] About W3C - https://www.w3.org/Consortium/