The Master Plate - controlling AI from within

Tony Czarnecki, Sustensis

London 31/01/2024

Image credit: DALL-E 3

How might Transhuman Governors control AI?

One of the key assumptions taken at the Global AI Safety Summit at Bletchley Park in November 2023 is that continuous improvement of AI may ultimately lead to the emergence of AGI and finally – Superintelligence, which I perceive as a single, global, most advanced AI system. Whether we become bystanders or decision makers in that process largely depends on our ability to control the development of Superintelligence. If we manage to control its self-improvement, i.e., its goals, values, and behaviour, then it may become our friend and help us immensely in delivering the Global Welfare State and also our own evolution. We need to have this process of tight control in place for the most advanced AI systems by about 2025, and completely operating on a global scale by about 2028.

Some people suggest we should allow Superintelligence to evolve independently and essentially leave it alone. The assumption is that this purely digital Superintelligence, when left to its own devices, will still cater for our needs. For many, this would be an ideal situation. However, that might be the riskiest approach. If there is no ultimate control over the Superintelligence's goals and behaviour it will almost certainly start fighting with us for access to resources, such as energy or rare earth metals.

To mitigate the risk of Superintelligence acting against our interests or even becoming outright malevolent, we must exercise early control over its development as it becomes increasingly more intelligent. To do that, we need a global and immediate implementation of the mechanisms controlling Superintelligence. This requires diverse approaches, which may collectively, better control the evolving "mind" of Superintelligence.

One such innovative approach is proposed by Yann LeCun, Chief AI scientist at Meta. His views on controlling AI are optimistic, including solving the so called alignment problem, i.e., aligning AI's goals and motives with human values and preferences. He maintains this opinion in an interview with Financial Times, where he suggests that "several 'conceptual breakthroughs' were still needed before AI systems approach human-level intelligence. But even then, they could be controlled by encoding 'moral character' into these systems in the same way as people enact laws to govern human behaviour." This is broadly in line with the opinion of another optimistic AI scientist, Gary Marcus. It contrasts with the prevailing view among AI researchers who maintain that controlling a superintelligent AI might be impossible, as it is impossible for a monkey to control a human.

However, LeCun's proposal focusing on encoding a "moral character" into AI systems, ensuring that they act ethically towards humans, deserves a closer examination. This idea is based on the possibility that AI's intelligence and its goals can be decoupled, allowing the development of AI systems that are intelligent but driven primarily by goals aligned with human values. While this concept sounds theoretically feasible, implementing it in practice remains a significant challenge. However, irrespective of the feasibility of the method he proposes, it is an interesting and potentially valuable approach to controlling AI.

In an article discussing that interview, Alberto Romero raises two caveats to LeCun's proposal. First, relying on external control mechanisms, like laws, might not be effective for superintelligent AI. Instead, moral principles should be fundamentally encoded into the AI's design. Secondly, the concept

of morality is subjective and varies among humans, making it difficult to create a universal moral character for AI.

On the other hand, implementing morality as a parallel backbone to the advanced AI decision making may be easier than creating a superintelligent humanoid in the context of Moravec's Paradox. In his book published in 1988 'Mind Children: The Future of Robot and Human Intelligence' Moravec postulates that it is easy to train computers to do things that humans find hard, like mathematics and logic, but it is hard to train them to do things humans find easy, like walking and image recognition. Morality does indeed fall into this category, since like higher cognitive functions, it is a relatively recent evolutionary development and might be easier to replicate in AI than more ancient, optimized human skills. Although I share LeCun's optimism, like Alberto Romero, I also think that the practicality of implementing such a system remains uncertain and doubtful.

The first problem, linked with practicality, lies with agreeing human values, the cornerstone of morality. Considering the current global politics this boils down to the following questions: what type of morality can be considered as human-generic, who would define it and how long it would take to agree the common human morality. A short answer – it is unrealistic to expect it could be ever done. If it were at all possible that all states agree on something so fundamental to their identity, it would take decades to achieve that. But the new algorithms for humans' morality would need to be developed in a few years' time. There may be a slight possibility to agree and develop 'a narrow morality' algorithm broadly acceptable by many countries but not by all. Therefore, that may happen once we have a de facto World Government, rather than a truly global government.

Secondly, morality may have not developed in hominids until consciousness has reached a certain level. That is why it is only present in humans, and perhaps to some extent, in apes or octopuses, the subject not raised neither by Yann Le Cun, nor Alberto Romero. Overall, it is an innovative proposal that should be implemented with all other methods, such as those proposed by Nick Bostrom in his seminal book 'Superintelligence'. However, none of them guarantees a failsafe control. We can only increase the probability of effective control by applying all feasible methods together.

All the methods of controlling AI have one thing in common – they try to control AI by humans. My view is that it is a forgone conclusion that sooner or later we would be the losers in this struggle for dominating the world. Instead, we should accept that AI is the next step in human evolution. The

biological homo sapiens will be gone. However, we may be the first ever creation of nature, which has designed its own evolution into a new species – a digital homo sapiens. If we accept that notion, then a logical approach is to start a civilisational transition to coexistence between humans and AI in a tightly coupled physical metamorphosis, similar to a caterpillar becoming a butterfly. Let me explain the concept briefly before expanding it below.

The core of my proposal is to create, what I call, the **Master Plate**, a method which may be more effective, unless physics and biology make its implementation impossible. The Master Plate is based on a BCI-fused control of Superintelligence, the most advanced AI, by Transhuman Governors. They would be carefully selected (including socio-psychological profiling) and connected in a ring via exponentially improving BCI devices to hundreds or even thousands of other licensed Transhuman Governors. One element of that ring would be the Master Plate's 'control hub'. This is a hardware/software device similar to a computer's BIOS (Basic Input Output System) enabling Transhuman Governors to control with their thoughts the main goals or decisions to be made by Superintelligence.

Such an approach would solve most the problems related to creating emotional, conscious and superintelligent being. But it would also start a gradual transformation of humans initially into Transhumans and ultimately into Posthumans – an entirely digital species. There would be no need of controlling AI, because it would be part of the most advanced Transhumans, as they would be part of the maturing Superintelligence.

The principle of AI's emotion, intelligence, and morality advancing in parallel with ours, where we are more advanced in consciousness and morality but less in intelligence, may be the best and the safest option because it greatly reduces the problem of lack of global agreement on human values and morality, which would take decades. The only requirement would be to initially select the avantgarde of humans (Transhumans) by an independent body authorized by a de facto World Government.

The Master Plate – an equivalent of a computer's BIOS

A complete control over a maturing Superintelligence is virtually impossible. If it attains sufficient intelligence, it will likely find ways to outsmart its controllers long before any planned escape from its restricted and protected environment actually takes place.

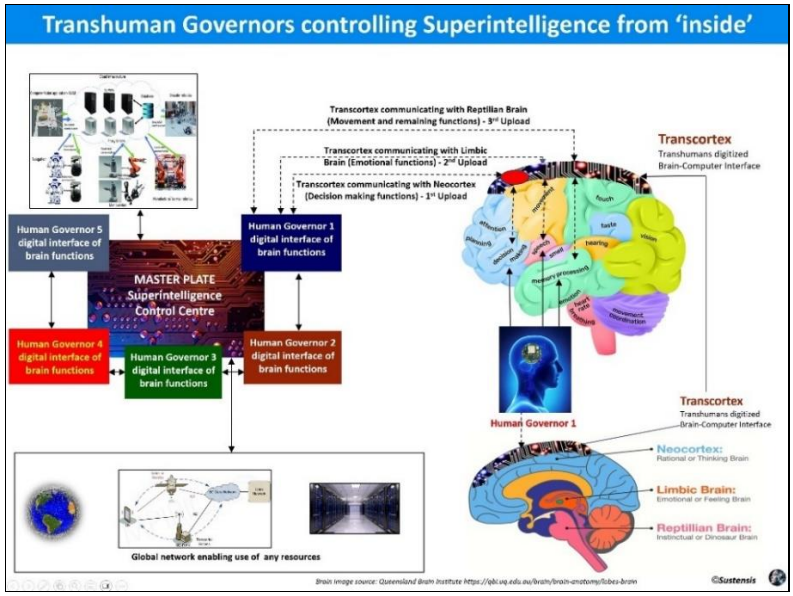
Instead, let's consider an alternative approach, which involves controlling Superintelligence by those responsible for developing its most critical components. They will have a role similar to those who develop updates to fundamental elements of the Windows system, such as BIOS (Basic Input Output System), which is essential for its functioning. Originally, BIOS was a ROM chip, a firmware, stored in on the PC motherboard (now, it is stored in flash memory circuits). What I propose is to install a Master Plate as a non-removable and non-programmable integrated circuit, in which its main 'value system' and preferred behavioural responses as well as other controlling parameters, such as 21 Asilomar Principles, would be stored. This would become a BIOS for the most advanced AI.

In January 2024, a prestigious Center for a New American Security (CNAS) published a research paper on how that could be done. CNAS introduces the concept of "**on-chip governance mechanisms**" - secure physical mechanisms built directly into chips or associated hardware that could provide a platform for adaptive governance. Such on-chip governance mechanisms could help safeguard the development and deployment of most advanced AI and supercomputing systems. Much of the required functionality for on-chip governance is already widely deployed on various chips, including cutting-edge AI chips .Chips sold by leading firms AMD, Apple, Intel, and NVIDIA have many of the features needed to enable the policies described above. These features are used today in a wide variety of applications. Therefore, what I am proposing in not a future technology, such a solution could be implemented today.

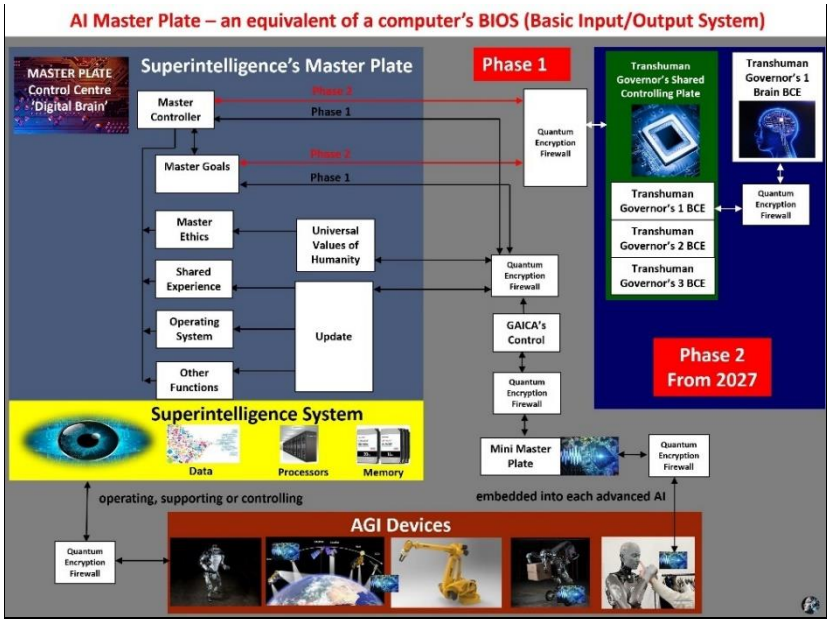
However, even such a tight control of AI may not be enough. We need a system of continuous control by the top developers deploying the most advanced system. To ensure the effectiveness and non-interruptive control, I propose in addition to such external control, also an internal control - controlling most advanced AI system from within. Yes, I am talking about perhaps the most potent method of controlling Superintelligence. It involves the wireless connection of some parts of the brains of the leading AI developers with the Master Plate, to control the maturing process of Superintelligence. Those involved in that process will be wirelessly connected to Superintelligence by Brain Computer Interfaces (BCI). They will become the first Transhumans. Since they will be controlling the maturing Superintelligence, they will become Transhuman Governors.

I refer you to my recent book '**Prevail or Fail – a Civilisational Shift to the World of Transhumans**' where the principles of wirelessly connecting various brain functions to external digital devices are explained. In the

diagram below, there are 3 types of uploads from the brains of 5 Transhuman Governors, enabling them to control the Master Plate.



The Master Plate will be implemented in phases as shown below.



In Phase 1 there will be no Transhuman Governors. The latest version of Anthropic's 'Claude' AI Assistant, which is almost as powerful as GPT-4, implements it in a fairly simple way using its new approach called Constitutional AI. Once a digital Master Plate has been manufactured, by a licenced company, it will upload the initial data like Superintelligence's Goals, Universal Values of Humanity (as agreed by an International Organization), its operating system and other components as shown.

The top level (grey box) is the actual Master Plate, which will control the second level, which is the maturing Superintelligence System (the yellow box). The Superintelligence System will control the third level (AGI Devices – the brown box), which may also be controlled by an international organization if needed. The most advanced AI devices, like humanoid robots, will have their own Mini Master Plates – see the bottom of the drawing.

Any interaction between the AI developers and the Master Plate would pass via a Quantum Encrypted (QE) devices, which could not be hacked because of the laws of physics (Quantum Entanglement). Similarly, any information exchange between the maturing Superintelligence system (the yellow box) with its external components or humans (the brown box) will only be possible via a QE filtering device.

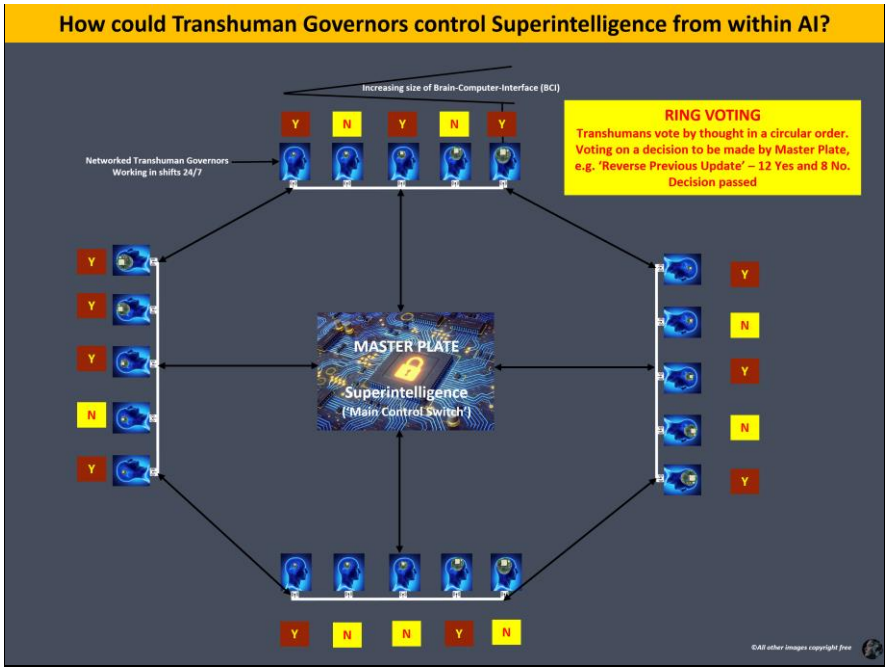
In Phase 2, the supervision of Master Controller and Master Goals will be performed by Transhuman Governors. The great advantage of using Transhuman Governors for controlling Superintelligence is that it would establish an immediate control. Additionally, as the advancement of the maturing Superintelligence progresses, so will the scope and the resilience of such control 'from within' since Brain-Computer-Interface (BCI) capabilities will advance at approximately the same pace.

Should the BCI Technology prove to be unreliable or even not feasible then the AI developers may control the Master Plate in the same way as in Phase 1, but it may be far riskier since even an immature Superintelligence will be much more intelligent than any human.

An effective programme of control must from the very start focus around controlling the AI's goals and behaviour, including knowing how it has arrived at any decision or solution, so called explainability. This must be built as the centre of all its decision, hence the proposed Master Plate. This is where the Universal Values of Humanity will be stored as well as its goals,

and human preferences, continuously updated as the maturing Superintelligence experiences the world of humans.

Transhuman Governors could be our best hope for retaining the control over Superintelligence for much longer. The early Transhuman Governors will give us the necessary experience in retaining the ultimate control over Superintelligence. Initially, perhaps just a few hundred specialists from various disciplines will be selected as Transhuman Governors and connected in a ring. Upgrading the software or authorising the execution of significant decisions by Superintelligence will require the consent of the majority of the connected Transhuman Governors. Superintelligence would thus be unable to change its key goals, how it functions, or which resources it uses if it is not confirmed by Transhuman Governors.

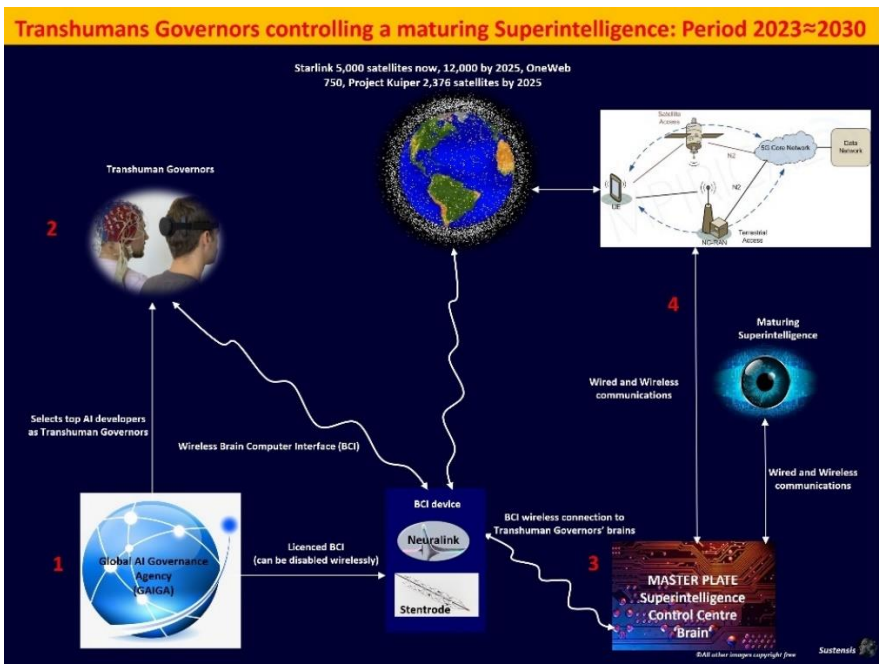


To control Superintelligence, it may be enough to oversee its activities just by Master Controller of the Master Plate. In the above diagram, there are 20 Transhuman Governors who are voting on 'reversing the previous update', because it may allow Superintelligence to escape from human control. The reversal of an update will require the consent of the majority of Transhuman Governors. Each of them will send a signal (Yes or No) by thought.

For this purpose, they will be connected in a ring to minimize the risk of the Superintelligence's access to the voting process. One end of the ring will be connected to the 'Input' of the Master Plate and the other to its 'Output'. When a decision is to be made, each Transhuman Governor would have one vote, imagining a 'Yes' or 'No' answer. As the voting on the ring progresses, the 'Yes' and 'No' decisions made by Transhuman Governors, will be added. Absent, or temporarily disconnected Transhuman Governors will be passed by. Once the last Transhuman Governor in the ring has voted, the majority decision will be executed on the Master Plate. Since all communications will be quantum encrypted, AI would not be able to manipulate such a decision, and neither will any of the Transhumans on his own. It will always be a majority decision. In that way, Transhumans will control Superintelligence for major decisions that it may need to make.

Selecting Transhuman Governors

Once BCI devices are reliable and harmless wireless communication with advanced AI such as GPT-4 Turbo have been established and authorized for implementation, the selected candidates will become the first Transhumans. The stringent criteria for selection should be established by an international organization, such as Global AI Governance Agency (GAIGA).



It should have powers similar to the International Atomic Energy Authority. However, the actual selection process should be conducted by an independent body comprising specialists from various domains, including AI scientists, neuropsychologists, biologists, chemists, physicists, and others.

A candidate for a Transhuman Governor must adhere to specific legal and ethical regulations, as well as undergo certain procedures, including:

- Consent to the insertion, removal, or digital disabling of the implant upon request by the licencing agency.
- Undergo psychological and psychometric evaluations.
- Maintain confidentiality regarding innovations, discoveries, and test results to prevent unauthorized replication of the process by potentially malicious individuals.
- Demonstrate openness, trustworthiness, and willingness to provide requested information to the licencing authority, or if required by a judicial court.
- Agree to share a portion of their memory pertaining to communications with Superintelligence and, if necessary, with other Transhumans within the controlling team.

Following the selection of the initial Transhumans, the controlling agency will proceed with issuing or implanting BCI devices for the selected top AI developers. These individuals will assume the role of Transhuman Governors and serve as members of the Transhuman Governors Board, operating according to the following framework:

- The selected candidates will have their BCI devices activated and start communicating wirelessly with a developed prototype of Superintelligence. Every year the scope of their interaction and the depth of control and monitoring of the maturing Superintelligence will widen, which may be necessary as its capabilities will increase exponentially. The wireless continuous communications with Superintelligence should enable its monitoring in real time.
- The role of Transhuman Governors, fully subordinated to the authorising agency, such as GAICA, will be to minimize the risk of developing by accident or intent a malicious Superintelligence. Since part of Transhuman Governors' brains will be wirelessly

connected to Superintelligence via BCI, they will be able to control it from within continuously working on shift schedule.

- It may be necessary at some stage for Transhuman Governors to communicate continuously and wirelessly between themselves. This may involve reading some of each other's 'deposited' thoughts, and intentions via something like a common external digital memory area.
- Only the authorized personnel would have access to an external memory area. This may be used for monitoring the working of the Superintelligence's prototype. As Superintelligence matures, the Transhuman Governors will gradually be communicating more and more often with it directly by thought alone.

In the first phase, these Transhumans would play the role of ‘guinea pigs’, testing how feasible and effective that method of controlling Superintelligence might be technologically and psychologically.

Initially, the first Transhuman Governors will be developers, neuroscientists, and specialist engineers who are at the forefront of the most advanced Superintelligence development. Their role will be similar to what they do today at OpenAI or Google’s Deep Mind when they decide, what functionality their applications will have, how those functions will be executed, and how individual people will be able to use them, which may depend on the users’ access rights.

For the first Transhuman Governors only a small part of their brain functions may need to be copied into a common area, such as decision making. They will be able to browse the Internet wirelessly by thought alone and store some of this information in a memory store in the Transcortex part of the brain and process it on an external computer. Progressively, more of their brain cognitive functions will be fused with the control centre of the maturing Superintelligence - the Master Plate.

Transhuman Governors will discuss any potential problems with the GAIGA’s Board to modify the Superintelligence’s development process as needed. But they also may increase their interdisciplinary knowledge exponentially by having access to their own large wirelessly integrated digital memory and processing capabilities. In just 3-5 years from becoming Transhuman Governors, they may be far more intelligent than any biological human in any aspect of human knowledge. With immediate access to the entire Google repository, they might be able to resolve many problems faster than any current computer. They will simply have an advantage over a purely

digital computer, by having consciousness and a general knowledge, which most advanced AI systems will not have for some time.

Giving Transhuman Governors such exceptional powers is not free of risks and will have significant consequences. Broadly, there are at least two **negative consequences**.

- The superiority of the Transhuman Governors' 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 that of 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.
- The impact on the political governance. There is no guarantee, that some of those Transhuman Governors will have no urge to dominate us all, using still immature Superintelligence. That is why I cannot emphasize it enough how potentially dangerous some Transhumans, including some Transhuman Governors, might become even within this decade.

I have been considering the selection and then supervision of Transhuman Governors by the controlling organisation, such as GAIGA, using Blockchain technology and operating as a new type of organisation called Distributed Autonomous Organization (DAO). They have emerged about 2015 and their intention is to democratize decision making by following step by step changes. One such example is SingularityDao, set up by Singularity.Net[82], one of the oldest and very influential body in the AI area. As most Blockchain organizations, it is linked to investment into cryptocurrencies. But if we consider that it would be used in the second half of this decade, it may be too slow and less effective than controlling Transhuman Governors and AI development by thought through the Master Plate.

When GAIGA is established, then at least in principle it will democratize the decisions made by Transhuman Governors on behalf of all humans. However, in practice such control of Transhuman Governors will be largely symbolic, since their decisions may 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. That is why the selection process of Transhuman Governors is so important.

But you may wonder why Transhumans couldn't control Superintelligence from 'outside'? Of course, they could. That is how it is currently being carried out. However, the advantage of controlling Superintelligence from within is as follows:

- If properly implemented with quantum encryption, which would be the ultimate security firewall, it gives the highest level of Superintelligence control.
- Immediacy of access to controlling Superintelligence via thoughts. Any attempt by Superintelligence of trying to get 'out of jail', would be immediately reported and acted on wirelessly.
- It is also an indirect method of a gradual mind uploading.
- It will significantly strengthen the overall effectiveness of other methods of controlling Superintelligence such as o those proposed by Nick Bostrom.

The only way to verify the feasibility and effectiveness of controlling Superintelligence by Transhuman Governors from within is in trying out this method.

Challenges to a reliable control of Superintelligence by Transhumans

I have serious doubts whether an effective control of the AI development process by certain restrictions and regulations is possible, even before AI becomes AGI. But I am not suggesting in any way, not to control AI. Just to the contrary. We must control AI development process by any available means to give us more time to prepare for the moment when AGI releases itself from human control.

However, I firmly believe that the best, if not the only way to control AI is to do that by progressively fusing more and more brain functions of the selected Transhuman Governors with the AI Master Plate, its decision centre. On the other hand, scientific objectivity requires to consider what happen if using Transhumans to control AI is ineffective on physiological grounds or because of other limitations. I said that I assume Transhumans, capable of controlling AI from within, by using sophisticated Brain Computer Interfaces (BCI), will be created by 2027. Although I am less concerned about a delay in the availability of such advanced BCI devices, I am unsure whether it may ever be possible to transmit reliably by thought alone any amount and any content of the human brain to an external device.

That may happen for many reasons. For example, BCI devices may actually enable advanced AI to use that wireless link to 'infect' the brains of the Transhuman Governors in a way that they may become unconsciously controlled by AI. This would mean an inverse way in which AI might be controlling Transhumans Governors, which in principle may be possible. There may however be some defences against that, like quite successful antivirus defences and firewalls used in IT systems.

Other challenges, which Transhuman Governors may encounter are linked to the transfer of certain brain functions, such as accessing the content of the human memory copied to external devices and accessing it later at any time by thought alone. This may be due to the biological brain's inability to handle the massive information flow and the discrepancy in information processing speeds. The average speed of a biochemical signal in the human neural network is at least 10,000 times slower than that of an electric impulse in a computer.

To overcome that potential difficulty a viable alternative might involve establishing a significantly slower information flow between the biological brain and a BCI device connected to Superintelligence. In this arrangement, the amount of the external information flow would be limited, and the transfer of digital information from Superintelligence to the biological brain would be decelerated, providing only final results or recommended decisions. This parallels the functioning of GPT-4, where the majority of processing and storage occur outside of your computer.

But what if filtering of the information content, or even slowing down the transmission speed does not solve the problem of a reliable transmission and copying of any content of the brain. What may happen then? It is a pure speculation, but if AI scientists come to such a conclusion *before* Superintelligence slips out of our control, then the following scenario may be plausible.

First of all, I assume that the AI's Master Plate would still be controlled successfully by partial fusion of certain brain function, like reading or switching on and off devices by thought alone, which is already possible. Therefore, Transhuman Governors would still be able to control major goals and decisions of AI. Secondly, nu then it will be almost universally accepted that an existential threat coming from AI is real and imminent. In such case, assuming all the time that only a single global AI development centre would exist, and other countries, like China, would also recognize that human brain limitation make it impossible to control AI from within, then the

development of the advanced AI to avoid it becoming Superintelligence, may be frozen perhaps for decades.

If an advanced AI development is frozen before it becomes Superintelligence then humans may have much more time to prime it with our values and preferences, nurturing it in real human environment for years. We would also have more time to develop new ways, which may ensure that such an immature Superintelligence does not escape human control.

You may rightly ask why we do not pause or even freeze an advanced AI development right now, as it has already been suggested in an Open letter signed by over 100,000 AI experts in April 2023. It will not happen for two reasons. First, we do not have any form of even a de facto World Government, so we would not be able to enforce such a decision. Secondly, developing companies and countries with a significant AI development potential, still do not see AI as a potential existential threat. That is why a single global Superintelligence development centre under an international control of an organization like GAIGA is so urgent. The most recent turmoil at OpenAI, where its CEO Sam Altman was expelled from the company because of the company's contradictory goals of developing a safe AI, which would also be very profitable, shows that only such a single global development centre is the only way for the most effective AI control.

Conclusions

In summary, there are two scenarios for controlling AI.

1. **Developing an advanced AI, as an independent digital species,** soon far more intelligent and capable than humans. The goal of such control should be to delay as far as possible the moment when Superintelligence gets out of our control. We will then have a better chance that it learns our values and preferences and be guided in its decisions by those values. However, realistically, when Superintelligence arrives, it will see all the inconsistences in our values and the way we live our lives. Therefore, after some time, rather than following our values, it will make its own decisions based on its far better understanding of human needs.

   It may then instal a new civilisational order, taking full responsibility for our future and creating the world of unimaginable abundance. At the same time, it would also facilitate the process of human species' evolution into a digital species.

However, there is no guarantee that once Superintelligence is out of our control it will not become a malicious entity. This may then become a dystopian scenario, in which humans may become extinct.

2. **Developing an advanced AI by Transhumans**. In this scenario there will be no more 'It and us' where we control AI to ensure our continuous existence as a biological species next to a far advanced digital Superintelligence. Humans' safety will be delivered by a gradual process of osmosis of our intelligence, emotions, and consciousness with digitised superintelligence until we make a transition similar to caterpillar becoming a butterfly. Instead of Superintelligence a new species will be born – Posthumans.