

TONY CZARNECKI

Prevail or Fail

A Civilisational Shift to Coexistence with Superintelligence

London, December 2023

Prevail or Fail

A Civilisational Shift to Coexistence with Superintelligence

© Tony Czarnecki 2023

The right of Tony Czarnecki to be identified as the author of this book has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

First published in May 2023 by Sustensis
This edition: December 2023

ISBN: 9798869718563

Cover design: DALL-E

London, December 2023

For any questions or comments please visit:
<https://sustensis.co.uk>

For humans

Other books by the author

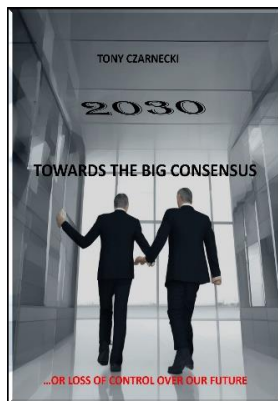
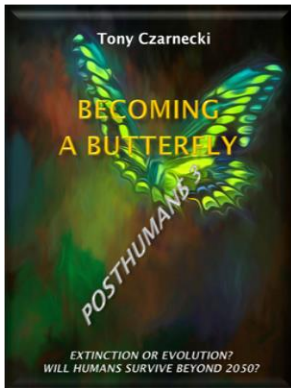
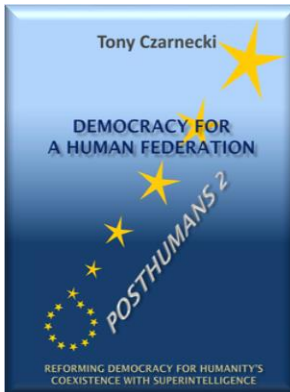
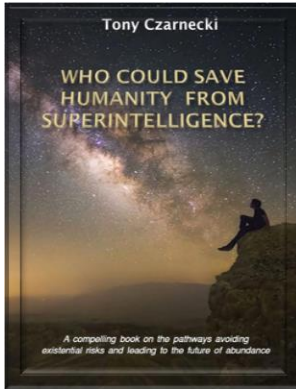


Table of Contents

FORWARD	8
INTRODUCTION	9
PART 1 WHY WE MUST CONTROL AI?	13
1. HUMAN INTELLIGENCE VERSUS ARTIFICIAL INTELLIGENCE	14
<i>Artificial Intelligence Primer</i>	14
<i>What does ‘human level intelligence’ really mean?</i>	20
2. IS AI AN EXISTENTIAL THREAT?	29
<i>What may happen if AI gets out of human control?</i>	29
<i>Don’t look up, even if a comet is to hit our planet</i>	33
<i>If a civilisational shift has just started – is there a way to halt it?</i>	36
<i>Four steps to minimize humanity’s existential threats</i>	37
1. Act as a planetary civilisation	37
2. Carry out a deep reform of democracy	37
3. Instil political consensus so that the voice of a minority can be heard	38
4. Accept that the only way forward for humans is to evolve	38
<i>We cannot uninvent AI as we cannot uninvent an atomic bomb</i>	39
<i>Prevail or Fail</i>	41
3. THE ROAD TO ARTIFICIAL GENERAL INTELLIGENCE	42
<i>What next after ChatGPT?</i>	42
<i>Cognitive AI</i>	43
<i>How to maintain humanness and uniqueness in the advanced AI?</i>	44
4. HOW TO GOVERN AI EFFECTIVELY?	47
<i>Options to control AI - lessons from the Manhattan Project</i>	47
Option 1: No global AI governance	51
Option 2: Full global AI governance	53
<i>Civilisational Shift to Coexistence with Superintelligence- the Schedule</i>	56
PART 2 TEN PRINCIPLES OF A SAFE CIVILISATIONAL SHIFT	61
1. ADJUST GLOBAL AI GOVERNANCE TO A CIVILISATIONAL SHIFT	62
<i>Don’t control exponentially changing AI with linear world measures</i>	62
<i>Split AI governance into AI control and AI regulation</i>	64
2. UNDERTAKE A COMPREHENSIVE REFORM OF DEMOCRACY	65
3. CREATE INTERNATIONAL AI SAFETY INSTITUTE (IAISI)	68
<i>The Moore’s law – the driver of a fast maturing AI</i>	68
<i>Prepare for AGI emerging by 2030</i>	70
<i>Why is it important to set 2030 as a date of loss of control over AGI?</i>	74
<i>How may we lose control of AI?</i>	76
<i>Creating the International AI Safety Institute</i>	78
4. AUTHORIZE GLOBAL PARTNERSHIP FOR AI STANDARDS AND REGULATION	83
<i>Governments should regulate the use of AI</i>	83
<i>Redefine the role of GPAI as a global AI regulation & standards Agency</i>	83
<i>The responsibilities of GPAI</i>	84
5. AUTHORIZE FRONTIER MODEL FORUM FOR GLOBAL AI DEVELOPMENT CONTROL ..	87
<i>Why should AI sector have a direct control of AI development?</i>	87
<i>Leaving the day-to-day AI development control to AI sector</i>	91

	<i>Why is the USA the best place to start global AI development control?</i>	93
	<i>Partnership on AI – a precursor of Frontier Model Forum</i>	96
	<i>From PAI to Frontier Model Forum (FMF)</i>	97
	<i>The prerogatives of Frontier Model Forum</i>	100
	<i>The responsibilities of Frontier Model Forum</i>	100
	<i>AI research transparency and open source policy</i>	102
6.	CREATE GLOBAL AI GOVERNANCE AGENCY (GAIGA)	104
	<i>The supervising role of the Global AI Governance Agency</i>	104
	<i>Prerogatives of GAIGA</i>	106
7.	CREATE GLOBAL AI COMPANY (GAICOM)	107
	<i>Learning from China’s Long-term AI Strategy Plan</i>	107
	<i>Building a Joint Venture One AI Company</i>	110
8.	CREATE SUPERINTELLIGENCE DEVELOPMENT PROGRAMME (SUPROG)	113
	<i>Why do we need one global AI programme?</i>	113
	<i>Consolidating AI development into One Superintelligence Programme</i>	114
	<i>AI Maturing Framework – a multi-modal control of AI by GAIGA</i>	116
	<i>Priming AI with Universal Values of Humanity</i>	116
	1. Controlling advanced AI via the Master Plate	120
	2. Teaching human values to AI directly	120
	3. Nurture AI as a child	121
	4. Enable all AI agents share their experience of applying human values	121
	<i>Summary of Global AI Governance</i>	123
9.	CREATE A DE FACTO WORLD GOVERNMENT	126
	<i>The need for the World Government</i>	126
	<i>Criteria for selecting organizations for the World Government</i>	127
	<i>GAIGA’s role as ‘the Ministry of War’ of de facto World Government</i>	130
10.	CREATE A GLOBAL WELFARE STATE	133
	<i>Redistributing wealth more evenly</i>	133
	<i>Building a Welfare State</i>	136
	1. Setting an individual’s wealth cap	137
	2. Setting the corporate assets’ cap	138
	3. AI-generated new type of wealth	140
	4. Much higher than predicted GDP growth	141
	5. Raising taxes to finance better life satisfaction	144
	6. Demonetization: significant fall in prices and faster growth of real income	145
	7. Substantially lower cost of government	146
	PART 3 THE CIVILISATION OF TRANSHUMANS	149
1.	THE DAWN OF A NEW CIVILISATION	150
2.	WHO ARE TRANSHUMANS?	153
	<i>Three aspects of Transhumanism</i>	153
	<i>You too can be a Transhuman</i>	154
3.	MAKING A TRANSHUMAN	158
	<i>Brain Computer Interface may turn you into a Transhuman</i>	158
	<i>Licencing BCI devices</i>	162
	<i>Mind uploading</i>	163
4.	TRANSHUMAN GOVERNORS CONTROLLING AI FROM INSIDE	168
	<i>How might Transhuman Governors control AI?</i>	168
	<i>The Master Plate – an equivalent of a computer’s BIOS</i>	171
	<i>Selecting Transhuman Governors</i>	176

	<i>Challenges to a reliable control of Superintelligence by Transhumans</i>	180
	<i>Conclusions</i>	182
5.	TRANSITION TO A TRANSHUMAN GOVERNMENT	183
	<i>Making the first steps in the evolution of the human species</i>	183
	<i>Decision making before the emergence of Superintelligence</i>	185
	<i>Options for making a transition to a new civilisation</i>	187
	<i>Civilisational Transition to the World of Transhumans</i>	188
	<i>1. Transition with a Transhuman World Government</i>	189
	<i>2. Transition with Transhumans but no World Government</i>	199
	<i>3. Transition without Transhumans and no World Government</i>	203
6.	SUPERINTELLIGENCE OUR BENEVOLENT MASTER	205
	CONCLUSIONS	210
	GLOSSARY	215
	REFERENCES	219

FORWARD

In the rapidly evolving landscape of Artificial Intelligence (AI), the latter half of 2023 has been marked by significant developments that underscore the urgency and complexity of AI safety and governance. The AI Safety Summit in the UK, culminating in the historic Bletchley Declaration, represents a pivotal moment in global AI policy. The commitment of the EU and 28 other countries to this declaration, along with the establishment of the UK's AI Safety Institute (AISi) and the U.S. Artificial Intelligence Safety Institute (USAISI), reflects a growing recognition of the need for coordinated efforts to manage the risks associated with advanced AI.

These institutional developments, coupled with Japan's concrete measures under the Hiroshima Artificial Intelligence Process and the Global AI Partnership (GPAI) established by the G7 Group, mark a significant shift towards a more collaborative and safety-focused approach to AI development. However, as the recent collapse of OpenAI vividly illustrates, the path to safer AI is fraught with challenges. This event highlights the tension between commercial interests in AI and the ethical imperative of ensuring AI safety, a theme that is central to the discussions in this book.

The developments of the past six months have not only reinforced the key tenets of this book but also necessitated updates to reflect the evolving AI governance landscape. While the Global AI Regulatory Authority (GAIRA) and the Global AI Consortium for AI Control (GAICA) were initially proposed in this book, the emergence of the Frontier Model Forum (FMF) and other real-world entities necessitates a revision to align with these changes.

In this second edition, I have revised sections of the book to incorporate these recent developments. The core principles and proposals remain intact, as they continue to provide a robust framework for understanding and navigating the complexities of AI safety and governance. However, the book now also includes an analysis of how these real-world developments align with, diverge from, or enhance the proposals and scenarios presented in the original edition.

As we continue to grapple with the challenges of AI, it is clear that the journey is one of continuous adaptation and learning. The events of the past six months have provided valuable lessons and insights, shaping our understanding of what it takes to build a safe and beneficial future with AI.

INTRODUCTION

In an era marked by unprecedented rapid change, the realm of Artificial Intelligence (AI) has emerged as a frontier of exponential progress. Since the advent of technologies like ChatGPT, we've witnessed strides towards Artificial General Intelligence (AGI), a concept once confined to the realms of speculative fiction. The journey to understanding and defining AGI is complex and ongoing, but one thing is clear: the need for human oversight and control over this burgeoning intelligence is paramount.

This book delves into the intricacies of AI development and its potential trajectory towards AGI and beyond, into the realms of Superintelligence. It explores the possibilities, challenges, and ethical considerations that come with these advancements. The core of this exploration lies in a fundamental question: how do we maintain control over a form of intelligence that is inherently self-learning and capable of surpassing human capabilities in every conceivable domain?

The stakes are monumental. The emergence and evolution of AGI could either herald a new age of unprecedented human progress or lead to catastrophic outcomes if left unchecked. Therefore, we must proactively engage with AI, ensuring that its growth is aligned with humanity's best interests. This alignment is not merely a technical challenge but a moral imperative, similar to the urgency with which we must address global challenges like climate change.

This book proposes a comprehensive approach to navigating this monumental shift. The "Principles of a Civilizational Shift" outlined here serve as a guiding framework for this journey. They encompass a broad spectrum of strategies, from global AI governance reforms to the creation of robust control mechanisms. These principles are not mere theoretical constructs but actionable strategies designed to ensure that as we coexist with Superintelligence, we do so on terms that preserve and enhance human dignity, freedom, and well-being.

It is our generation that has reached a pivotal point not only in our civilization's history but quite likely in the whole history of a human species. The book aims to provide a roadmap for navigating the uncertain but potentially rewarding future that AI presents. It is a call to action for policymakers, technologists, and citizens alike to engage deeply with the questions and possibilities that AI brings. The future is not just something

that happens to us; it is something we can actively shape, and in the domain of AI, shaping this future responsibly is perhaps our most crucial task.

The release of ChatGPT and similar AI Assistants signifies a significant stride towards achieving AGI. In 2014, the futurist Ray Kurzweil predicted that AI would reach human-level intelligence by 2029, but there is still no consensus on what AGI is.

However, regardless of the kind of AGI, which emerges by about 2030, it is vital that we are able to control it, before it starts controlling us. Such loss of control over AGI may be a gradual process rather than an abrupt event. A complete loss of control will occur when we are unable to reverse AGI's decisions. As a self-learning intelligence, AGI will outperform humans in any task or situation, including evading human oversight. If AI control methods prove ineffective, AGI might achieve this even before 2030.

Once AGI gets out of our control, it will resist any attempts to reimpose it. Assuming its capabilities continue to improve exponentially, it may have catastrophic consequences. Therefore, it is imperative to explore all feasible options to ensure human control over AI beyond 2030. This will enable us to better adapt to coexisting with Superintelligence, immensely more capable than humanity. To ensure the survival of humanity, we must fundamentally revise the necessary solutions for effective AI control. Just as we must take more significant actions to address Global Warming, so we must adopt a similar level of commitment to AI development control.

We must recognize that the scale of the required changes represents a **Civilizational Shift** in the history of a human species. To successfully navigate this transition from our current state to a new civilization, we must accept the magnitude and timeframe required for this transformation. The key solutions proposed in this book, are presented as "**Ten Principles of a Safe Civilizational Shift**", arranged in an ideal sequence. But of course, the reality will quite likely determine the most feasible implementation order.

1. **Adjust global AI governance to a civilisational shift** since AI is not just a new technology but an entirely new form of intelligence, which requires strict **AI development control**. It's separate from **AI regulation**, which is mainly about the use of AI as a tool. Both are part of AI governance but require different procedures and have different impact on humans' future.

2. **Undertake a comprehensive reform of democracy**, as it is a prerequisite for achieving effective AI development control and aligning it with human values. We must rebalance the power of governance between citizens and their representatives in parliament.
3. **Create International AI Safety Institute (IAISI)** to minimise the unexpected advances in the frontier AI models by developing dedicated monitoring and testing methods. It should operate in a similar way as the *International Panel on Climate Change (IPCC)*. While there is no scientific proof that AGI will emerge by 2030, just as there is no proof of the Global Warming reaching a tipping point by that time, we must develop AI as if AGI were to emerge within that time frame and retain control over AI control beyond 2030.
4. **Authorize Global Partnership on AI (GPAI) for AI standards and regulation**, leaving AI development control to a new Agency. It should also set global standards for specific AI hardware and operate like *International Standards Institute (ISI)*.
5. **Authorize Frontier Model Forum for a global AI development control** of the most advanced AI model by expanding its US base to include companies from other countries. It should operate like the Internet's *W3C Consortium*.
6. **Create Global AI Governance Agency (GAIGA)** under the mandate from the Bletchley Declaration and the Hiroshima Process. It should have the prerogatives similar to the *International Atomic Energy Authority (IAEA)* in Vienna. GAIGA would oversee both GPAI, responsible for regulating the use of AI products and services, and the FMF Consortium, responsible for AI development control.
7. **Create Global AI Company (GAICOM)**. This could be a Joint Venture company to consolidate the most advanced AI companies into a single organization. It would be similar in its objective to the *ITER project* funded by the US, China, Russia, the EU, Japan, India, and Korea, to develop the first nuclear fusion reactor. Effective control over AI development will be impossible if it remains dispersed among numerous companies.
8. **Create Superintelligence Development Programme (SUPROG)** managed by GAICOM. This would be similar in its objectives to the *NASA's Apollo Programme*.
9. **Create a de facto World Government** perhaps initiated by the G7 Group, incorporating members from NATO, the European Union, the European Political Community, or from OECD.

10. **Create a Global Welfare State**, which would also include the setting up of a Global Wealth Redistribution Fund, needed to mitigate the challenges posed by the transition to the World of Transhumans.

The suggested implementation deadlines in the book are based on three assumptions:

1. Transhumans (humans with Brain-Computer-Interfaces – BCI) with far superior intelligence than any human will emerge by about 2027.
2. AGI will arrive by 2030,
3. Superintelligence will emerge by 2050.

These deadlines have served me to present certain scenarios and actions that might be needed if some of the predictions become reality. Therefore, my goals has not been to align predictions strictly with current trends, but rather to explore a wide range of possibilities, including those that may seem unlikely or divergent from the current trajectory. You may find these scenarios of some value because they challenge the status quo and encourage thinking beyond the constraints of current knowledge and trends. It's an invitation to consider what could be possible, rather than just what is probable based on current conditions.

By presenting a variety of scenarios, some of which may seem far-fetched, I hope to stimulate innovative thinking and prepare us for unexpected developments. This is crucial in a rapidly changing world where the future is increasingly unpredictable. While some predictions may seem less likely, their inclusion is important for a comprehensive exploration of potential futures.

Finally, in writing this book I want to challenge complacency and encourage active engagement with the future. By presenting potential futures that diverge significantly from current trajectories, I hope to motivate individuals and organizations responsible for our future to adjust their paths to avoid undesirable outcomes and rather strive for more optimistic future.

The civilisational shift, which we are starting to experience, is our evolutionary test for 'the survival of the fittest'.

PART 1

Why we must control AI?

1. Human intelligence versus Artificial Intelligence

Artificial Intelligence Primer

For an average person, just the term **Artificial Intelligence** (AI) may be quite confusing, as it seems to cover all aspects of what seems to be 'unnatural'. It may start in difficulty to differentiate between Information Technology (IT) and AI.

IT processes information based on strictly defined rules, generally requiring all input data, although there are some heuristic systems that can operate without all data being available. However, AI can produce results based on partially available input data, as it operates similarly to a human mind – using probabilities. It can also learn from experience. Therefore, the same input data may not always produce the same output. The learning experience is what makes some humanoid robots resemble humans – they make errors, but progressively fewer than humans. To make matters even more confusing, many people, including myself, use the term AI as a general descriptor for all types of AI.

What we have now are individual, relatively unsophisticated AI assistants, chatbots such as ChatGPT, or robots. This is generally referred to as Artificial Narrow Intelligence, which is mostly defined as follows:

ARTIFICIAL NARROW INTELLIGENCE (ANI) can exceed human intelligence and capabilities in a single area

These could be games, including poker, which require some intuition, smelling, tasting, or face recognition. ANI can be run on a single computer to perform a single, narrow function supporting one of human skills. However, it is ignorant in all other areas.

Such humanoid robots will be capable of carrying out most physical tasks around the house or in a factory, communicating verbally with humans. 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. Unless there is a global legislation banning the unlicensed use of some of the most advanced products, they may shortly create global AI networks themselves. Such a global AI system could create,

if misused, a near existential risk. So, AI does not have to be fully matured, to become an existential threat.

By the end of this decade, we may have an **Artificial General Intelligence (AGI)**, which will reach human level intelligence. Wikipedia defines it as “*the ability of an intelligent agent to understand or learn any intellectual task that human beings or other animals can*”^[1]. But if we want to build AGI we must have a more detailed definition, identifying its key features. Moreover, we would need to know what that ‘intelligent agent’ really means, like what I would propose below:

Artificial General Intelligence is a self-learning intelligence, superior to humans’, solving any task far better than any human.

How many years away are we from the moment that AI will have human level intelligence making them smarter than humans? Paul Pally, the proponent of Natural Language Understanding theory, who uses a similar definition and predicts AGI may arrive in 2024 ^[2]. I am perhaps a bit more realistic, and like Ray Kurzweil, the renowned futurist, I predict that AGI may emerge by 2030. That prediction is a few decades earlier than many AI researchers still maintain.

If I am right, soon there could be thousands and possibly even millions of AGI humanoids costing perhaps as much as a luxury car. In technology terms it would be a standalone AI system controlling local devices with the access to the Internet. It will need at least these capabilities to achieve a human level intelligence:

- **Short-term memory:** Memorize text, images, graphs and of course events in a conversations (remembering what was said before). OpenAI’s GPT-4 Turbo can memorize about 250 pages of text (a whole book), images and graphs similarly as Anthropic’s Claude 2.1, so that’s done.
- **Long-term memory:** Record events, topics discussed, and knowledge learned (equivalent to the hippocampus in our brains memorizing events in space and time). That is still limited but should be achieved at an average human level by perhaps the end of the next year, and quite likely be the end of 2025.
- **Multi-step instruction:** Combine intermediate results of individual instructions, building them into the final output. Practically done at a number of companies like Microsoft’s Kosmos-1, Google’s PaLM-E

and several others. It will be perfected by Musk's Optimus and Google's Gemini by the end of 2024 and certainly in 2025, when the first such humanoids will be on sale in limited numbers.

- **Goals and interests:** Create own goals and interests, a kind of a 'free will', which must be compatible with human goals, values, and existing laws - a huge problem of AI Alignment, potentially opening Pandora's box. We may have to wait till 2027-29.
- **Be truthful and objective:** If AGI is to be human-friendly it must behave following the Universal Values of Humanity. This may require linking goals to human preferences by checking the output. Some progress is being made, e.g., Claude-2 uses its own 'Constitution' to do just that. However, we need to align with an agreed system of Universal Values of Humanity. At the moment top AI developers do it, instead of the World Government. It will be very difficult. If we don't achieve this by 2027-29 and AI gets out of human control, it may potentially become malicious.
- **Emotions:** ChatGPT can detect emotions, and Ameca humanoid can show emotions by following the user's emotions, but they don't feel them. Feeling emotions is not necessary for AGI to have human level intelligence but may be achieved by 2029-30.
- **Cognition:** Simulate human thinking in complex situations, when the answers may be ambiguous or uncertain, using the acquired knowledge, understanding & experience. This is tough but may be achieved comprehensively about 2028-2030.

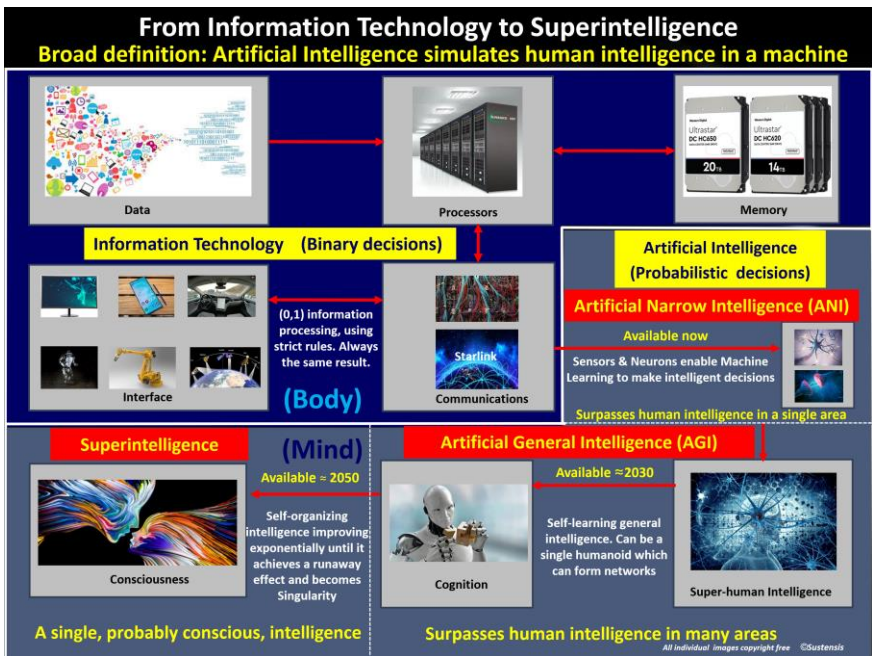
But the progress can be much faster. Unconfirmed reports indicate that OpenAI was planning to release GPT-5 by the end of 2023, which it describes as a near AGI^[3]. The disagreement about the release of such an advanced AI was quite likely the main reason for sacking and then reinstating Sam Altman, under some pressure from OpenAI's employees. If this is the case, OpenAI may have already achieved, what Sam Altman said, a near AGI.

More significantly, in November 2023, NVIDIA released H200 processor, which is many times faster, with much larger memory than H100 supporting all current Large Language Models (LLM) like GPT-4. Currently, OpenAI uses about 100,000 H100 processors on Microsoft's Azure platform, enabling it to support hundreds of millions users. However, with just one of these processor costing about \$40,000, the Meta's Llama-2 can now be run as a standalone version reaching the performance only a bit lower than GPT4^[4]. If we extrapolate a near exponential increase in computer power, a

standalone computer running AGI could be available by the end of this decade, costing less than a luxury car. The implications of such an early emergence of AGI running on a nearly ubiquitous computer, may fundamentally impact how we live but also how a single person can start a global chaos. That might result when these standalone AGI's become ever more powerful by creating a network and ultimately evolving into **Superintelligence**.

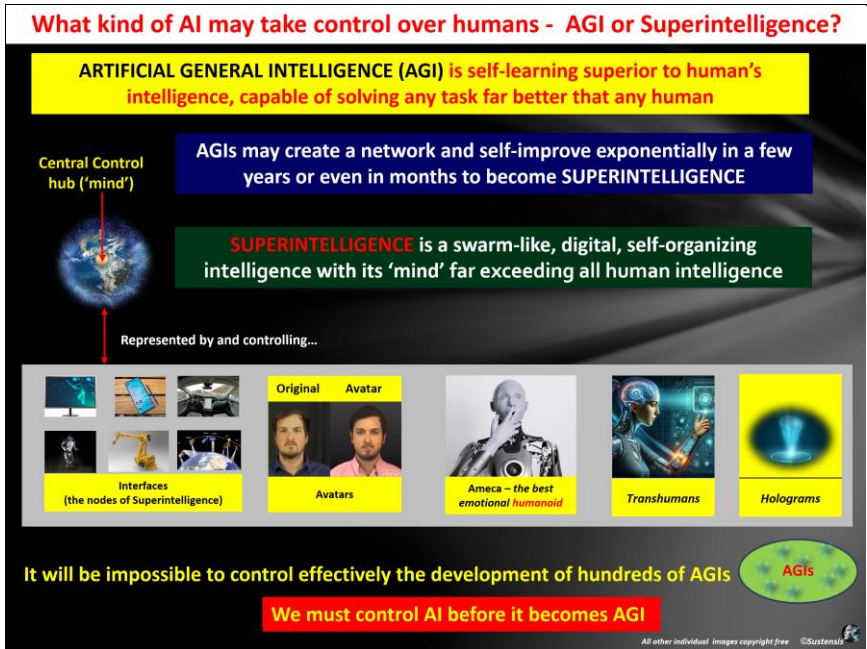
So, what is Superintelligence (or Artificial Superintelligence – **ASI**) and how does it differ from AGI? As with AGI, there is no agreed definition of Superintelligence. I define it as follows:

SUPERINTELLIGENCE is a swarm-like, digital self-organizing intelligence, with its ‘mind’ exceeding all human intelligence



Its body consists of various elements such as data, processors, memory, interfaces, communications, sensors, including artificial morphic neurons. All these building blocks are currently thousands of times slower than required for AGI. Therefore, it is unlikely the current AI systems could support AGI with full cognition – an experiential knowledge and awareness of the world.

Superintelligence may emerge spontaneously by AGI's self-networking and self-improving in a matter of a few years or even months. Therefore, the understanding of what is AGI and how it differs from Superintelligence is very important for its development control, which is covered extensively in this book. This picture may help visualise how that may happen.



Superintelligence will be operating via its avatars, holograms, or as emotional humanoids, such as AMECA robots. It will also be linked to conscious Transhumans, i.e., humans with embedded Brain-Computer-Interface (BCI) via wireless communication with access to external memory and processing power. Its behaviour towards humans will depend on whether it has inherited human values, responsibilities, preferences, and expectations. Those values and responsibilities should form a globally agreed Universal Values of Humanity. They should be embedded as early as possible into a top-controlling digital Master Plate of the maturing Superintelligence (see Part 3).

Once Superintelligence emerges, it may then gradually turn into a conscious entity. However, there is no agreement among AI researchers whether Superintelligence must be conscious.

A mature Superintelligence millions of times more intelligent than any biological human will certainly consider human values with its own 'Mind' and may thus very quickly replace them with its own. We will not be able to stop it as it develops a separate set of its own values but may let biological humans to govern themselves autonomously as much as possible. If it is controlled digitally from 'inside' by Transhuman Governors, then it is highly likely to be human friendly (see Part 3).

A matured Superintelligence will most likely be seen by biological humans as a single entity. It will be millions of times more intelligent than any human, probably conscious, unless we have means to decide otherwise, concluding that a Superintelligence without consciousness is a safer option. It may also have billions of complex digital modules, replicating individual human brains, with backup facilities (synchronized copies of the brains). Each such module may be supporting a conscious human mind of a Posthuman. Such modules will likely differ in their capabilities, size, and power to facilitate special roles of certain Posthumans. The Posthumans' 3D representations will be non-biological avatars, probably not conscious but with a high degree of awareness.

If we achieve a full integration with the maturing Superintelligence at a digital level via increasingly more capable BCI devices of the Transhuman Governors, who will ultimately have their brains fully fused (copied) with the 'brain' of Superintelligence, then we may be governed by a Posthuman Government. Should that happen, then it would mean that humans have evolved into digital species. Therefore, there will be no distinction between Superintelligence and Posthumans who will be 'residing' within Superintelligence and being its real mind. In such case, all decisions made by Superintelligence are likely to be made by a system of voting by all Posthumans and executed by billions of robots and avatars, controlled by Superintelligence.

Some digital Posthumans may be located in space (or their backup copies may be there), for example in Low Earth Orbit (LEO), Geostationary orbit, the Lagrange orbit, on the Moon or even on Mars. However, it is quite likely that in this scenario, vast majority of humans will remain in their biological regenerated bodies for a long time. That means for example, that centenarians may still look and have physical and mental capabilities of biologically much younger people.

Should a full and error-free mind uploading of human brains in a digital form be not possible, then humans will be under a total control of Superintelligence, incapable of understanding the rationale behind some of its decisions. That alone will be an existential threat for humans because we will no longer have any control over our own destiny. Whether such a mature Superintelligence becomes a threat to a human species depends largely on whether it was nurtured in line with human values, so-called AI alignment, before we have completely lost control over it. If Superintelligence has even slightly misaligned objectives or values with those that we share, it may become hostile towards humans.

There is yet another, benevolent scenario. In this case, even if Superintelligence has a full control over humans, it may not interfere with our lives too much, and instead provide anything we need, creating an unimaginable Global Welfare State. That may be considered an anthropic way of thinking. This is similar to humans caring for the animal kingdom (only recently). Soon human meat-eating needs may be fulfilled by stem cell-based meat production and our interference into the animals' world will be minimized.

There are many predictions about the likely time of the emergence of Superintelligence. The date, which is mostly quoted, is 2045, predicted by Ray Kurzweil in his book 'Singularity is Near' published in 2007. He also predicted in 2017 that AGI (human-level intelligence) will most likely emerge by 2029. Seeing how surprising were the ChatGPT creators by its vastly better performance than had been expected, and how it has improved over just one year, I would see it as indeed a very likely date. If the speed of AI improvement and its supported hardware continues at the current pace, Ray Kurzweil may also be right about predicting 2045 as the date of the emergence of Superintelligence.

What does 'human level intelligence' really mean?

This is the area, which I have considered very important for some time since this might make AI development control more effective. But it is also the area of several unknowns such as consciousness and cognition. Therefore, anyone venturing to debate intelligence, risks being misunderstood for putting forward ideas or concepts, which have no sound foundation in reality. Nevertheless, it may be helpful to use comparisons between human and AI intelligence at least to approximate how close AI is from being superior to humans, especially that even some AI researchers consider it a

new type of technology. **But AI is foremost a new, inorganic intelligence.** That new intelligence may achieve its goals and solve problems differently than we do, being smarter than any human, even if it does not tick all the boxes on the human intelligence definition, e.g. abstraction.

When comparing the AI's and human intelligence most authors use a definition which clarifies what it is from a human perspective. For example, Encyclopaedia Britannica defines it as follows: '*Human intelligence - mental quality that consists of the abilities to learn from experience, adapt to new situations, understand and handle abstract concepts, and use knowledge to manipulate one's environment*'^[6].

However, for a more objective comparison, we should consider intelligence from the perspective of the Universe, in which there may be different forms of intelligence, of which biological intelligence may be just one. This is similar to Ray Kurzweil's thinking when he said in 2023 'The universe has been set up exquisitely enough to have intelligence. There are intelligent entities like us who can contemplate the universe and develop models about it, which is interesting. Intelligence is, in fact, a powerful force and we can see that its power is going to grow not linearly but exponentially and will ultimately be powerful enough to change the destiny of the universe'^[7]. Thus, my definition of intelligence in the context of the Universe is as follows:

'Intelligence is an attribute of an organic or inorganic system, which intentionally changes its environment to achieve its goals using minimum amount of energy'.

The condition of a minimum amount of energy is necessary for the intelligent being to evolve and be even more intelligent. From that point of view, panpsychism at a macrophenomenal level offers some explanation of how intelligence, as an attribute of a generic mind, may actually work. In summary, AI's and humans' intelligence should be compared from the universal perspective rather than from a strictly anthropic point of view.

In November 2023, AI researchers from Google DeepMind published an article titled: "Levels of AGI: Operationalizing Progress on the Path to AGI" [8]. Surprisingly, in this breakthrough article, the authors did not define what AGI is, as I have done above. Instead, they proposed, what they call, 'operationalizable definition'. It is based on 9 examples of AGI definitions, provided by other authors. The authors of the article use them to provide 6

properties and commonalities of AGI as the basis for a certain focus (direction) of AI research:

1. Focus on **Capabilities**, not Processes
2. Focus on **Generality** and Performance
3. Focus on **Cognitive** and Metacognitive Tasks
4. Focus on **Potential**, not Deployment
5. Focus on **Ecological Validity**
6. Focus on the **Path to AGI**, not a Single Endpoint

That helps them to define 6 levels of AI Competence, like for the self-driving cars:

- Level 0 - **No AI**
- Level 1 - **Emerging** – *equal, or slightly better than an unskilled human*
- Level 2 - **Competent** - *at least 50th percentile of skilled adults*
- Level 3 - **Expert** - *at least 90th percentile of skilled adults*
- Level 4 - **Virtuoso** - *at least 99th percentile of skilled adults*
- Level 5 - **Superhuman** - *outperforms 100% of humans*

If the AI sector accepts this AI ontology, as happened for self-driving cars, and adopts these competencies as the guidance for developing AGI, then this may become an effective starting point for determining the current competency level of the maturing AGI. That is why this article is so important. It has also entered the uncharted waters of defining intelligence, necessary to establish a more precise meaning of human-level intelligence.

Since the authors of the DeepMind's article have not defined what intelligence is, it is difficult to see how the Competence (Autonomy) Level is linked to intelligence. To do that, I have applied the Multiple Intelligence Theory created by Howard Gardner, an MIT professor of psychology at Harvard University. It challenges traditional beliefs in the fields of education and cognitive science. For example, the American Psychological Association defines intelligence as follows: '*Intelligence is the ability to derive information, learn from experience, adapt to the environment, understand, and correctly utilize thought and reason.*' The key word in this definition is **understand**. According to that traditional definition, intelligence is a uniform cognitive capacity people are born with. This capacity can be easily measured by reasonably simple tests. But according to Gardner, intelligence is:

- The ability to create an effective product or offer a service that is valued in a culture,
- A set of skills that make it possible for a person to solve problems in life,
- The potential for finding or creating solutions for problems, which involves gathering new knowledge,

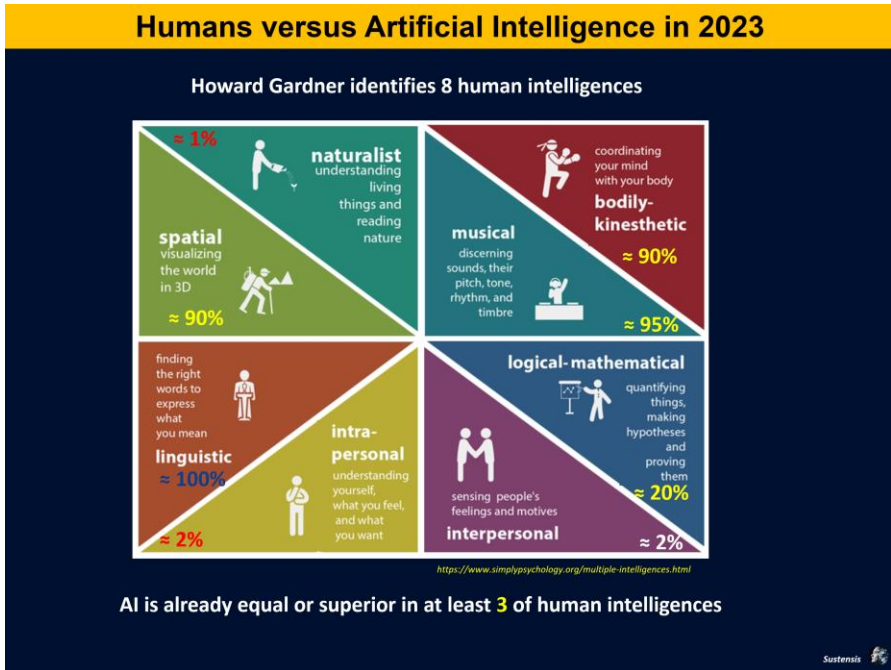
In addition, Gardner postulates that:

- All human beings possess all intelligences in *varying* amount,
- Each person has a *different* intellectual composition,
- These intelligences are located in different areas of the brain and can either work independently or together,
- These intelligences may define the human species,
- Multiple intelligences can be nurtured and strengthened, or ignored and weakened,
- Each individual has eight intelligences (and maybe more to be discovered)^[9].

Unlike the majority of theories of intelligence, Gardner's theory of Multiple Intelligences proposes a differentiation of human intelligence into specific modalities of intelligence, rather than defining intelligence as a single, general ability. The theory has been criticized by mainstream psychology for lack of empirical evidence, and its dependence on subjective judgement^[10].

However, I would suggest the opposite. The arguments that a human brain is unlikely to function using Gardner's multiple intelligences, are precisely the reason why his theory is more useful for comparison with the AI's intelligence. As can already be seen, AI's intelligence is, or at least it may be, of a different kind than human's.

Howard Gardner identifies 8 human intelligences^[11]: Linguistic, Logical/Mathematical, Spatial, Bodily-Kinaesthetic, Musical, Interpersonal, Intrapersonal, and Naturalist. Skills are mainly about *doing*, whereas intelligence is more about contextual *understanding*. Therefore, in my comparison some of AI's 'intelligences', are simply *skills or competencies*, which are needed to perform a task that requires intelligence.



How human intelligence compares against Artificial Intelligence today.

I have estimated how well AI currently matches human intelligence in each of the eight intelligences based on the AI's skills, and NOT on AI's intelligence viewed from a human perspective. Looking at individual skills, rather than intelligence type as a whole, is more relevant approach because it compares the real impact of AI on us and the environment, i.e., on what really matters, at least from the point of view of AI development control.

I should also clarify which level of human intelligence I am comparing: the most intelligent people, or an average human. DeepMind suggests that AGI should have superhuman capabilities, i.e., surpassing any human in competence and intelligence. I also assume we are assessing AI against most intelligent people, e.g., when evaluating postgraduate exam results and, each comparison made only for a particular type of intelligence. For example, in translation there is no human capable of speaking and fluently translating 100 languages. However, when it comes to translating poetry, humans still excel due to their understanding of language nuances. The question then arises: which superiority is more significant in achieving life goals, i.e., being smarter and having a better chance of surviving in a dangerous situation? It's the ability to translate 100 languages simultaneously, i.e. to communicate, the area in which, AI is already vastly superior. It may be only

slightly behind top human translators in non-technological areas. Therefore, the score of 90% is not an exaggeration.

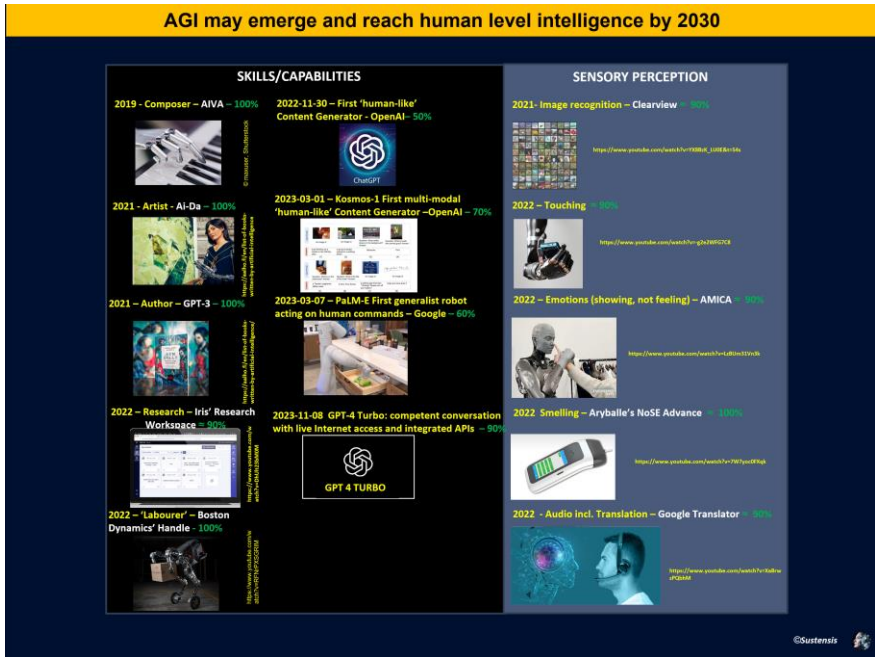
My other assessments follow a similar approach, verifying what is most useful for achieving life goals. From that perspective, in four of the intelligences, Bodily-Kinaesthetic, Linguistic, Musical and Spatial, AI already equals or exceeds top-performing humans. For example, in bodily kinaesthetic, some robots like Boston Dynamics' Atlas, can perform nearly acrobatic jumps, comparable with top human gymnasts. If we were considering just that skill alone, AI achieves about 90% level.

However, in maths, when performing error free complex calculations, AI is still far behind humans. It can prove theorems, but is not capable yet of creating its own, because that would require genuine creativity, which it does not have. Therefore, in that area, AI is perhaps only at about 20% level.

In music, including composition, art (painting) or literature, it is quite often impossible to distinguish the products created by AI from those created by humans. Therefore 80% score seems to be reasonable.

Similarly in spatial visualization tasks, such as object recognition in images, or 3D object reconstruction from images, AI systems have achieved high levels of accuracy. For instance, in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), the winning deep learning models achieved top-5 error rates of less than 5%, which is lower than the human error rate of about 5-10%. In 3D object reconstruction, AI systems can reconstruct detailed 3D models of objects from multiple images with a high level of accuracy.^[12] This data is about 10 years old. Today, the error rate should be even lower. Therefore, a score of 90% seems reasonable in this category, although of course there are areas where humans still excel AI, especially in creativity, which is unmeasurable. Humans are still immensely superior in Interpersonal, Intrapersonal (understanding yourself, emotions etc), and in Naturalist areas. These are the areas closely related to cognition.

In addition to comparing human's intelligence against AI's intelligence at an aggregate level, I have compared it directly at a particular skill level. Here, a very fast progress has been noted in sensory processing, a crucial component for developing AI's cognitive capabilities^[5]. AI's perception skills, like taste, smell or touch have been available for at least two years. They are far more accurate than human senses but need to be integrated, like other modes: video, sound, verbal communication etc.



We have made a relatively modest progress in those areas where AI is not yet at a human level, e.g., a Generalist Robot (or Universal Assistant). That may change soon, since OpenAI and DeepMind, the two leading AI research organisations, are currently changing the approach from relying mainly on the two big planks in AI – Large Language Models (LLM) and Deep Learning. They are shifting towards a more general-purpose, cognitive approach, to create systems that can reason and learn across a wide range of domains, rather than just excel at specific tasks or problems.

It is possible that the problem of cognition may solve itself when humanoid robots such as Ameca or Optimus, arrive in larger numbers operating in real environment as humans' co-workers. Elon Musk thinks his Tesla cars are the most cognitive AI systems, which must be aware of thousands of situations in a second. Therefore, once that software is uploaded with some modifications to an Optimus humanoid robot, cognition, including self-awareness, may just happen spontaneously.

If AGI does not emerge instantaneously, but rather progressively both in the scope and level of competencies, as suggested by the authors of DeepMind's article then this is significant for two reasons. First of all, it is directly related to the ability of controlling AI development. If AI is developed by

continuously measuring its benchmark competence level and indirectly its level of autonomy, then the AI researchers and developers will have much greater scope of control and plenty of warning of emerging AGI.

On the other hand, because of AI's self-learning capabilities, once it acquires some cognitive functions, it might increase its competence level and scope very quickly, perhaps in weeks, and become AGI. The best example is the pace of GPT product line improvement. Version 3.0 had a maximum 1 page of context memory and quite a laborious process of tuning AI, with no attachments and no tools. It took about 2 years for version 3.5. to arrive in the shape of ChatGPT. No direct attachment and no API tools could be used. Within a few months, GPT 4.0 has arrived, which had 3 pages of context memory, allowed to attach a file and use some API tools, such as Excel or DALL-E 2. And then within a few months, version GPT 4.5 Turbo emerged, which now enables about 250 pages of context memory, multiple attachments, and use of APIs in the background.

The release of DeepMind's paper has quite significantly increased the chances of AGI emerging in the next few years because developers will have to deal with fewer unknowns related to accidentally producing AGI, before global mechanisms of control have been established. The frontier models will be measured and compared against a benchmark once it is globally accepted. It will become a valuable tool for assessing how close AI is to become AGI and what risk the most advanced model may pose to humans. The monitoring of that process should be the key task of the Global AI Safety Institute set up at Bletchley Park Summit in November 2023. However, developers will still have to be prepared for uncomfortable surprises of a sudden increase of near AGI's capabilities, where none were expected. That is what happened when ChatGPT was released.

In summary, this breakthrough paper shows that AGI is unlikely to be reached in one instant. Instead, it will arrive in stages, defined by Autonomy levels (capabilities or competencies). Measuring the AI's Autonomy level, i.e. its independence in defining and carrying out its goals, may give us a clear warning when AGI may emerge. Although it does not remove entirely the overall problem of controlling AGI, it may in some way delay or even stop development of the most advanced AI, when it is still at a lower level of Autonomy. This would give us more time for aligning the AGI's goals and behaviour with humans' values and preferences. Overall, the most recent research presents rather an optimistic perspective for a more effective AI development control, which means you should not be scared of AGI.

However, for all that to work, there would need to be a single, Global AI Development Centre, under the supervision of an international organization with powers similar to International Atomic Energy Authority (IAEA) – See Part 2, chapter 8.

2. Is AI an existential threat?

What may happen if AI gets out of human control?

One of the most frequently quoted examples illustrating what may happen if AI gets out of human control is an intended or erroneous launch of nuclear weapons by AI or opening the laboratories with deadly viruses. But there are many more examples of how out of control AI can impact us, like these ones:

1. **Autonomous Weapons.** These are weapons powered by AI which have the potential to cause harm and damage, as they may be difficult to control or be used unethically, violating the international law. They are already deployed on the North-South Korean border,
2. **Privacy Concerns.** AI systems can collect and analyse personal data, raising concerns about privacy and surveillance. They can already manipulate users without their knowledge for financial, political, or even criminal reasons. More advanced AI may create real havoc among population at large for that very reason, including invading the privacy of thought^[13],
3. **Subverting the functions of local and federal governments,** international corporations, professional societies, and charitable organizations to pursue its own ends, rather than their human-designed purposes^[13]
4. **Cybersecurity Risks.** AI can be vulnerable to cyberattacks since hackers could exploit weaknesses in the software. It could also manipulate the data, which is used in machine learning to train the system,
5. **Safety concerns.** These will be more apparent as AI becomes more integrated into society when a faulty AI system could cause accidents or other safety hazards,
6. **Lack of Transparency.** Some AI systems are difficult to interpret and understand. This is so called ‘black box’ scenario. In this case it is difficult to tell how an AI agent knows what it knows or how it does what it does. The most recent worrying example is the release of ChatGPT, which had such a ‘black box’ but which was hurriedly resolved,
7. **Unethical Use.** AI can be used for unethical purposes, such as voter manipulation, surveillance, or other forms of malicious influence,
8. **Bias and Discrimination.** This problem has been well publicised. It concerns AI systems, which can learn biases from the data they are trained on. This may result in discrimination against certain groups or

individuals. A former Google ethical Researcher Timnit Gebru, fired by Google for her stance on keeping ethical values high on the Google's agenda, is one of the examples how AI companies may prioritize profits before sticking to their mission^[14],

9. **Job Displacement.** We are not talking about odd jobs displacements, but a wave of sudden unemployment, called Technological Unemployment. As AI technology advances, it could lead to massive job displacement in certain sectors or whole industries, when machines and algorithms replace human workers,
10. **Misuse of AI.** This is already apparent at a smaller scale. AI in the wrong hands could be used for malicious purposes, such as social engineering (i.e., suggesting who should people vote for), cybercrime, or espionage,
11. **Economic inequality.** This can also be seen right now, when the AI availability, inadvertently linked to the Internet, depends on the network presence or resilience, as is the case in Africa and in some poorer countries. It may aggravate existing economic inequalities in richer countries, as certain industries or individuals may benefit more from its use than others,
12. **Regulation.** That is part of the process of AI control. The lack of proper regulation and oversight in the industry may potentially lead to negative consequences,
13. **Over-reliance.** This risk of AI use is embedded in our nature. When something works most of the time, we assume it will work all the time. Additionally, the consequences of an AI system being without proper human oversight could lead to loss of total or partial human control over an AI system's goals and its future behaviour,
14. **Accountability.** The best example here are autonomous cars. If such a car kills a pedestrian because of a malfunction in the system, who is responsible: the person in the car, who nominally is a driver, or the car manufacturer? Although it is not a threat as such, it shows that if AI is not properly regulated, we may also have this kind of problems,
15. **Controlling resources.** This might include preventing access by an AI system to money, land, water, rare elements, organic matter, the Internet service, or computer hardware,
16. **Unintended Consequences.** This is a well-known scenario when an AI system has the goals, which it interprets differently from the intentions of the system designers, resulting in harmful outcomes or unintended effects,

17. **Restricting freedom of movement or choice.** AI may force people to stay at certain locations, put them in prison, or even decide what to do with our bodies and minds,
18. **Abusing and torturing people.** With a perfect insight into a human physiology and psychology, AI can cause physical or emotional pain.

We can expect a lot of these types of attacks in the future.

**What harm can an unfriendly
Immature Superintelligence do to us by 2030?**

		
Launching nuclear weapons	Preventing humans to use some resources	Subversion of governmental and private organizations
		
Abusing or torturing people	Significant restriction of freedom of movement	Creating a nearly total surveillance state

If combined with several existential risks, it could lead to a civilizational catastrophe

That is the kind of damage AI can cause in the short-to-medium term. The longer-term threat from AI stems from even the slightest misalignment of our values with the AI's "values" and objectives. If this happens, even when the corresponding goals initially appear benign, it could be disastrous. Nick Bostrom quotes a scaring example that involves Superintelligence programmed to "maximize" the abundance of some objects, like paperclips. This could lead it to harvesting all available atoms, including those in human bodies, thereby destroying humanity (and perhaps the entire biosphere)^[15].

The situation is even more complicated once we consider systems that exceed human intelligence. Superintelligence may be capable of inventing dangers we are not even capable of predicting or imagining. Nick Bostrom expands that argument further by saying:

“The value alignment problem is made even more dangerous by the possibility that a Superintelligence’s thought processes could run millions of times faster than ours, given the vastly different speed of electrical potentials in computer hardware versus action potentials in the human brain. Superintelligence could also learn to rewrite its own code, thereby initiating an intelligence explosion until some upper limit, perhaps far above human intelligence, is finally reached”^[16]. So, we have to accept that the creation of AI, or its most advanced form, Superintelligence, poses perhaps the most difficult long-term risk to the future of Humanity.

Recently, many more top AI scientists gave a warning about a potential existential threat of AI. Geoffrey Hinton, the British computer scientist is perhaps the best example. Against the prevailing opinion among the AI scientists, he proposed in the 1980’s an entirely new approach – ‘deep neural networks and back propagation’. He had to wait 30 years for the computers to reach the processing power to use it, which led to such spectacular advancement of AI in recent years. Now, seeing what ChatGPT can do, he resigned as Chief Scientist at Google Brain, to speak freely about the existential threat of AI. In March 2023 he said it is "not inconceivable" that AI may pose a threat to humanity^[17]. He regretted his invention. If people like Hinton say that, then you may draw your own conclusions. He reminds me about Robert Oppenheimer, the Project director on the Manhattan project, who regretted that he had contributed to the invention of the atomic bomb.

We have already seen the first examples of the damage done by the so-called narrowly focused AI systems. In 2023, just three months after the release of ChatGPT, entirely new ways of using it for malicious intent have been introduced. The first one is ‘jailbreaking’. That can happen through “prompt injections,” in which someone uses prompts (questions or instructions), which tell the language model to ignore its previous directions and safety boundaries. The second one is ‘assisted scamming and phishing’. Here, the attacker uses an AI virtual assistant to manipulate it into sending personal information from the victim’s emails, or even emailing people in the victim’s contacts list on the attacker’s behalf. Finally, it is now possible to use an AI Assistant for ‘data poisoning’. That involves manipulating (‘poisoning’) the data, which is used to train the AI Assistant’s Large Language Model (LLM) so that it acts in the way, the attackers wants it.

Such failures are just a warning. Once we have developed Superintelligence capable of accomplishing a much wider range of tasks, the damage will be

much worse. Imagine an AI agent that could trigger the switching off power grids in just one country. Since grid networks are connected globally, it could create a very serious damage world-wide in almost every aspect of life for many weeks, if not months.

You may of course still think that it would be possible to put the genie back in to the bottle. If so, here is the warning from David Wood, author of many books on AI and chairman of London Futurists:

‘...since the software may take evasive action against operations intended to interfere with its performance, the possibility arises that the software may prove difficult to control. For example:

- To guard against the possibility that the programme might be shut down, the software could take its own decision to tunnel a copy of itself out to a safe location,
- To guard against the possibility that such copies would be intercepted and rendered inoperative, the programme could take steps to keep the copying process secret, and to disguise its intentions,
- To forestall other possible attacks on itself, the software might devise innovative new defensive strategies that the programmers had not foreseen – strategies potentially outside the imagination even of science fiction writers.^[18]

It is another warning that it may be impossible to control AI’s behaviour, after it has escaped into the environment. Just to repeat it again: there is no failsafe option to control Superintelligence. We can only minimize that threat by combining various control methods, including controlling it by Transhuman Governors (see chapter 4, part 3).

Don’t look up, even if a comet is to hit our planet

‘Don’t look up...but AGI instead of the comet’. That tweet by Elon Musk, in his very own style, begins this section. For those who have not seen the 2021 movie ‘Don’t look up!’, here is a brief explanation. In the movie, the scientists are convinced the comet is to hit the Earth in a few months’ and call for an immediate action to correct its trajectory. But populist politicians just hours before the comet hits the planet still organize huge demonstrations to urge people NOT to look up and see the coming comet. Elon Musk replaced the ‘comet’ with ‘AGI’ addressing the message to top AI

developers such as OpenAI, Microsoft or Google, which behave as there had been no danger from the soon to arrive AGI.

I have started the chapter in this way to prepare you for the decisions and sacrifices that may be needed to control AI. Since there is no way, in which we can stop this process we must find a pathway towards an effective control of the AI development until the time, when it will be aligned with best human values, becoming our friend rather than foe.

But Elon Musk's tweet has also brought to the fore the scale of that threat. In a metaphoric way, it compares our situation to an all-out global war with one difference – the enemy is not visible yet. To win that war, we need to be prepared to change current laws or entitlements and accept some restrictions. We must think the unthinkable. AGI is not here yet, but it is clearly visible, like a comet was clearly visible to all who wanted to see it. If AI is left out of control or such control is ineffective, it will most likely not leave us alone. It will become progressively our enemy, initially competing for limited resources, and later fighting us directly. In the worst-case scenario, it will lead to the extinction of all humans by the end of this century. Any percentages qualifying the probability of that to happen this century are unnecessary, since in such a situation life will become unbearable even in a few decades.

However, some AI scientists and researchers try to estimate the chance of AI presenting an existential threat. In August 2022, AI Impacts organisation surveyed 738 AI scientists on the probability of AI becoming an existential threat. 48% of them responded that they estimate that threat at 10%.^[19] In an article^[20] by Alberto Romero published in May 2023, he quotes similar estimates made in March and April 2023 by well-known science authors Yuval Noah Harari, Tristan Harris, Aza Raskin and the physicist Max Tegmark. But he is not so much concerned whether the 10% estimate of humans' extinction is credible or not. He undermines, justly in my view, the whole approach of estimating such a risk, which is unscientific, because it cannot be calculated in any meaningful way, like for example, the risk of an airplane catastrophic failure.

It is far better to take the view of the scientists like Geoffrey Hinton, who in response to the question of “how soon he predicts AI will become smarter than us,” said: “I now predict 5 to 20 years but without much confidence. We live in very uncertain times. It's possible that I am totally wrong about digital intelligence overtaking us. Nobody really knows which is why we

should worry now.” Scientists are worried, because they simply don’t know the level of risk of AI becoming a potential existential threat^[20].

That is the main point I am making in this book. Since the top AI scientists can only ascertain that AI can be an existential threat, we should take this extremely seriously and act as if it had been a proven case. Perhaps a more pragmatic assumption is to view that risk from the bottom up. AI researchers know what capabilities AGI must have and how soon it may acquire those capabilities. I have taken all that into account and agree with a growing number of AI scientists that AGI will emerge by 2030 (chapter 3 of Part 2).

When it does emerge, it will be smarter than Humans. So, what happens then? Let me quote Geoffrey Hinton once more: “If it gets to be much smarter than us, it’ll be very good at manipulation, because it would’ve learned that from us and there are very few examples of a more intelligent thing being controlled by a less intelligent thing. And it knows how to program so it’ll figure out ways of getting around the restrictions we put on it. It’ll figure out ways of manipulating people to do what it wants.^{[20]”}

The only way, we can minimize that risk is to mitigate it. This is why I use the word ‘must’ rather than ‘should’ quite often in this book, to bring your attention to the limited choices that the world still has. Changes in politics, economy and social domain will be AI driven. Whatever we will do, it is too late to avoid the greatest truly global chaos in this decade. Current political and legal structures are unfit for purpose. We must significantly re-invent global politics and democracy within a few years, knowing from the outset that it will be imperfect but better than doing nothing. It is a difficult, and perhaps for some people, even a horrific scenario.

But there is another scenario, where humans may very soon experience life of unimaginable wealth and contentedness. There are only two conditions. First, we must accept that humans are governed on entirely new principles as a planetary civilisation, which means among others to forsake national sovereignty and some restrictions on our freedoms for our own safety and benefit. Secondly, we must accept even a bigger challenge. We may not become extinct only if over a century or two we evolve into a new species. If you have accepted this line of thought then it will be easier to acknowledge the need for some radical and necessary changes, which will lead us to the world of abundance and exhilarating self-fulfilment rather than to extinction.

If a civilisational shift has just started – is there a way to halt it?

I would be surprised if after reading the Introduction, you would not emphatically say ‘No, that’s rubbish, there must be an alternative, we are not at this stage of a possible civilizational collapse’. I fully understand such a reaction because that is how I felt seven years ago, when I started writing my first book. I just could not accept that we, thinking humans, may be so oblivious to at least 8 man-made existential threats, which could lead to the end of this civilisation or even human species’ extinction.

The covers of my five books published since then and shown at the beginning of this book, illustrate my search for a solution which might alleviate such a threat. I started with a question, which was also the title of my first book: ‘Who could save humanity from Superintelligence?’^[21] In that book, I have thoroughly reviewed those existential risks to understand what might be done to mitigate them. I realized that humans have just two options: either become another extinct species or evolve into a new species. Since then, I have been trying to find an answer to how we might minimize those existential threats and eventually evolve into a new species.

I realized that not all existential threats are equal. A global nuclear war, or a global artificial pandemic would not lead to a human species extinction. In the most recent article in ‘Nature’, the scientists estimate that in a global nuclear war about 5 billion people would die but there will still be some places where humans may survive^[22]. It would be a civilisational collapse but not a human species’ extinction. Regarding artificial biological pandemic, caused by a virus or a bacterium escaping from a biological lab, some people would survive because of a slight difference in their genome. Both these existential risks can happen at any time, even tomorrow. Global warming is a different category, because it is ‘a slow burning’ existential risk, which if nothing is done may lead to humans’ extinction, or Earth becoming uninhabitable in about 100-200 years. So, humans will then have time to prepare an escape route to Mars or to the Moon.

However, Artificial Intelligence (AI) is an entirely different category. In the worst case scenario, it will lead to the extinction of every human being, quite likely by the end of this century, if nothing is done to control it. Even before then, life for humans would be extremely unbearable because of the likely war that AI might fight against humans for the access to resources, such as energy or rare metals.

To answer fully the question asked in my first book ‘Who could save Humanity from Superintelligence?’, I wrote four books, which were like the steps to solving a problem. I thought that we might have about 30-40 years to prepare the transition of humanity to the time when we will start coexisting with Superintelligence and I summarize my reasoning below.

Four steps to minimize humanity’s existential threats

1. Act as a planetary civilisation

My book “**Federate to Survive!**”^[23] was the initial step in my pursuit to identify **WHAT** needs to be done. The answer was that humanity needs to federate as a planetary civilization right now. However, realistically, it would be impossible to federate all nations based on common values today. This is what the World Federalist Movement has been trying to achieve for nearly 80 years, waiting for all nations to live by the same system of human values, such as peace. Look at the result.

It is a great idea which had almost no chance of being implemented. Nations get together for many reasons. They can stay independent while being part of a confederation, created around a mutual goal, like never fighting each other, so they can all live in peace. The European Union is the best example of that, which has enabled a peaceful life in Europe for over 75 years. Nations can also federate around values, in which case, they will by definition live their lives in a similar way by sharing common values, tradition or culture. For that, they are prepared to share part of their nation’s sovereignty. The USA, built on the principle of freedom, equality, and social solidarity, is the best example.

Ideally the United Nations should be the organisation, which would be powerful enough to implement jointly made decisions. Unfortunately, the UN has some systemic errors, making it powerless to govern the world as a planetary civilization. Neither do we have time to build such an organisation from scratch. Therefore, we need to select an existing organisation that seems to be best suited for introducing such reforms.

2. Carry out a deep reform of democracy

“**Democracy for a Human Federation**”^[24] continues from where the previous book ended, proposing **HOW** we can survive existential threats. We need two elements to achieve that: **Democracy** and a **Human Federation**. One of the most important, urgent, and very difficult steps

would be to create a new set of Universal Values of Humanity. That would be essential not only for a deep reform of democracy but also to prime a maturing AI with the universal values, which define humanity. But I also saw it as a prerequisite for creating a well-functioning World Government.

3. Instil political consensus so that the voice of a minority can be heard

That was the third step in minimizing existential threats. I described that in my book “**2030 - Towards the Big Consensus**”^[25] In that book, I discuss the problem of governing at the time when a de facto World Government might already be in place, assumed in the book to happen by about 2030. How would it govern us? To minimize existential threats, the values such as national sovereignty and some of our personal freedoms may need to be restricted. We must remind ourselves that we cannot have personal or national freedom without responsibilities. But the introduction of such restrictions may lead to serious social unrest in many countries. Why should citizens trust their governments when today the trust in politicians is at the rock bottom?

The only way to rebuild the trust is to set up a new Social Contract between the governing and the governed and build a **Big Consensus** fast. The starting point would be the removal of political and social imbalances in societies by **merging direct and representational democracy** into a new type of democracy – Consensual Presidential Democracy. The cornerstone of that type of democracy is to counterbalance the power of elected politicians with the power of randomly selected citizens who would form the second chamber – a Citizens Senate. Among other proposals in that book was disallowing the governance by a single majority party.

4. Accept that the only way forward for humans is to evolve

In “**Becoming a Butterfly**” book I ask **WHO** we may become as species by the end of this century, assuming we will survive existential threats. Its focus is on **Superintelligence** as a mature form of an ever faster and more intelligent, self-learning Artificial Intelligence. If that final product becomes a malicious entity, it may make us extinct in a few decades. However, if we do it right, it will not only protect us from existential risks but also create unimaginable prosperity in the world of peace, and endless possibilities for human self-fulfilment.

That coexistence will gradually lead to humans becoming Transhumans, with some parts of the human body, including parts of the brain, being non-

biological. Like a butterfly, we will be morphing into something new, still being humans but having a new shape and capabilities. Towards the end of this century some Transhumans may decide to upload their mind into a digital form morphing seamlessly with Superintelligence. Humans would then become a new species – Posthumans.

We cannot uninvent AI as we cannot uninvent an atomic bomb

In 1933, Arthur Holly Compton, the Nobel Prize laureate in physics, wrote a report, in which he stated that the idea of a sustained nuclear chain reaction was "extremely remote" and that it would not be achieved for several decades. Just a day later, Leo Szilard, a Hungarian-American physicist had a "flash of insight" when walking in a London Park that led him to conceive the idea of a nuclear chain reaction. That was the basis for the future Manhattan project and the first atomic bomb, but also for building nuclear power stations.

The American physicist Edward Teller made it plain, when talking about the nuclear energy to the American Physical Society in 1957, said that: "you cannot uninvent a nuclear bomb". He later became an advocate for the development of even more powerful nuclear weapons, including the hydrogen bomb, arguing that the only way to prevent nuclear war was to maintain a balance of power between nuclear-armed nations.

Where is the similarity between the invention of an atomic bomb and AI? In 2022, when I was writing my fifth book, the world was still living in a similar period as Arthur Compton did – before an atomic bomb was invented. With the public release of ChatGPT on 30th November 2022, we have entered the Leo Szilard's world. We have just started living in the world when another genie is out of the bottle – the AI so capable that within a few years it will be more intelligent than anyone of us in almost every way. Since it is a near certainty that, if such an Artificial General Intelligence (AGI) appears first in the USA, then a few months later it will be created in China and several other places. If any of those variants of AGI slips at that time out of human control, it will start gradually controlling everything what we do and quite quickly start fighting for the same resources, like energy. You can draw your own conclusion how such a battle may end.

However, if we could go back a decade or two, could we have developed an AI in a different way, where the final product might have not posed an existential threat for humans? For example, could we have developed AI

which would have not been embodied into any physical devices or objects, such as humanoid robots? Could AI have been developed as a ‘pure’ Artificial Intelligence locked forever in a computer as a software, like in that movie *Her*?

To answer that question, we need to check if the AI’s *understanding* and *cognition* can arise without AI being embodied in a physical object such as in a factory robot. If by *understanding* we mean recognizing the meaning, relationships, and dependencies between the objects then I would say it is possible. I would also say that *cognition* may arise even in a disembodied AI if we understand it as ‘making judgement and decisions in complex situations, using the acquired knowledge, understanding and experience’. This is already (almost) evident in the latest release of GPT-4, powering ChatGPT. AI could even realise some of its goals without any embodiment, like beating the world champions in Go-Go or chess.

Perhaps we should have limited any AI work and research to be only conducted within a computer, disabling the control of physical devices, or accessing the Internet by AI. It would have then remained on a chip unable to do us any physical harm. AI would thus become only our intellectual partner and a problem solver.

But would human’s inquisitive nature resist a temptation to see what AI could actually do for us, rather than be satisfied with just what it can tell us? I don’t think so. Technological, and later on, scientific progress and curiosity have been the backbone of a civilisational progress. Any new discovery offered first of all new possibilities and any risks were mostly ignored. Science has no barriers. The discovery of a nuclear energy that can do so much good, led also to the invention of a nuclear bomb, which could annihilate all humans.

A disembodied AI might have been a safer option for humans, avoiding the creation of the biggest human-made existential risk, if such a ban could have been applied and enforced globally, which was utterly unrealistic. However, without embodiment such AI would not be able to change anything in the environment and therefore would only have a limited impact on the betterment of the humans’ material condition. In summary, although theoretically possible, a disembodied AI is only an interesting subject for a philosophical discussion. The embodiment has already happened and cannot be ‘uninvented’.

Today AI is embodied in all Tesla cars and millions of intelligent robots, especially in humanoid robots, with physical actuators controlled by thousands of AI's sensors reacting to decisions made by its software. In effect those robots, are the AGI's avatars, similarly as our body (the torso) is a permanently attached avatar of our brain/mind.

Prevail or Fail

At this point you may be in a state of shock. However, bury one's head in the sand is not the best option. We have to face reality. That is the reason, why this book looks at solving the problem I first saw when writing 'Who could save Humanity from Superintelligence?' But today, I am proposing other solutions. The main difference is the length of time we have left to minimize existential threats. We no longer have 50 years to minimize an existential threat posed by an uncontrolled AI. We may only have just 10 years left.

Hence, there is no alternative to controlling AI development. The scope and the sequence of the required reforms presented in this book came out by asking this question: '**What might work and if it could be implemented on time**'. However, political reality may require some alterations in the sequence and the way the changes would be implemented. In any case, we must start thinking the unthinkable.

One way of thinking the unthinkable is to imagine a weighing scale, which on left side has most of your treasured values, and on the right side it has just one value – LIFE, meaning worthwhile wellbeing. Choosing the status quo and holding your personal and emotional possessions, which mainly mean your current values, may not only lead to the loss of your life but also to the loss of life of your children and grandchildren, i.e., to human species' extinction. This is the '**Fail**' side of the scale

If you choose the right side of the scale, which reads '**Prevail**', then you select the preservation of life in the immediate future against some constraints and restrictions. However, in the long-term 'Prevail' promises you the life in the world of unimaginable abundance and possibilities.

Which part of the scale would you chose: **Prevail or Fail**?

3. The road to Artificial General Intelligence

What next after ChatGPT?

In my most recent book ‘2030 – Towards the Big Consensus...Or loss of control over our future’, I said ‘Everything around us changes faster than ever in human history. Pace of change is nearly exponential. What in the year 2000 might have taken a decade, can now be accomplished within a year’ [25]. That exponential pace of change leads to a rising phenomenon when projects started some time ago, become obsolete before they are completed. One example is the UK’s £100Bn HS2 railway project, which will almost certainly be obsolete before it is completed in 20 years’ time [26]. But the same is with books. My most recent book may already be to some extent obsolete since the events in the AI development area have progressed so fast in January and early February 2023 that what in 2022 seemed to be a distant possibility, has now become a reality.

The public release of ChatGPT on 30th November 2022, and what followed in the next two months, resulted in three paradigm shifting events in AI, which have completely changed the way we look at the AI progress and the time by when we may lose control over its behaviour and goals:

1. ChatGPT has apparently learnt things it was never supposed to do. It was ‘broken in’ many times by AI researchers from other organizations, revealing that the product can behave in an uncontrollable way. This may be the evidence that at its core is a ‘black box’, which functions are poorly understood [27] [28] (this has only been cleared in April 2023),
2. The merger of ChatGPT and Bing into Bing Chat, which for the first time enables it accessing the Internet. That was quickly followed by Google, which said it had done a similar merger between their LaMBDA Chatbot (previously put in the ‘fridge’ after the Blake Lemoine incident) with their Google Search engine, into BARD,
3. Breaking the principles of the Partnership on AI (PAI), set up in the US in 2016 with over 100 members (including Google, IBM, Microsoft Amazon etc.). It encouraged co-operation and openness in sharing improvements of the released software, and AI research, rather than competition. Publishing the Transformer technology by Google in 2017, enabled OpenAI to develop GPT and its latest incarnation – ChatGPT. That was the best example of working in the spirit of PAI, which says on Transparency & Accountability:

“We remove ambiguity by building a culture of cooperation, trust, and accountability so our Partners can succeed, and so everyone can understand how AI systems work’ [29]. More about that in Part 2.

The implications of these three events on the future of AI and indirectly on our civilization are truly momentous because they prove that:

1. The largest AI companies Google (Alphabet) and OpenAI (Microsoft) are in a competition to release their products as quickly as possible, to ensure the support of their shareholders. That means the release of products and services, which may not be properly tested, to deliver them to the market as quickly as possible,
2. Such attitude of the leading AI companies, make the delivery of fail-safe AI even more difficult. The difference between even the largest IT programs (software) and AI is that the latter is a self-learning entity, which means it can progressively learn almost anything, including how to get ‘out of jail’, i.e., escaping human control,
3. The speed of release of various Chatbots such as LaMBDA, PALM or DALLE-E in 2022, has accelerated the emergence of AGI.

This progress will be even faster if we consider the advancement 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 may reach about 86 trillion in 2024, equal to 86B neurons in a human brain. But the most recent change of approach to develop AGI, may not require more than 1 trillion tokens, which Open-AI’s GPT-4 may have already reached. Moreover, if we include the super-exponential pace of development in synthetic (neuromorphic) neurons and quantum computing, we can expect even faster acceleration of the AI capabilities.

Cognitive AI

The next step for the ‘AGI in the making’ is to have some level of self-awareness leading gradually to cognition – a very difficult area. But ChatGPT or its improved version – GPT-4 has now a good level of human speech understanding – a fundamental element of cognition. What it lacks so far is not just the meaning of particular words and phrases but understanding in a broad human sense.

Perhaps the best way to explain how the understanding of the **language** differs from the understanding of a given **situation** is to quote Shannon Vallor, a professor of philosophy at the Edinburgh University, who says this:

“Understanding is beyond GPT-3’s reach because understanding cannot occur in **an isolated behaviour**, no matter how clever. Understanding is not an act but a **labour**.^[30]” What she means is that every time we want to understand how objects, people, images, text or feelings are related, we use both the present and historical situation, trying to build a trajectory of how the current situation evolved from the past and how it may evolve into the future. That requires ‘labour’, a key word in her essay, because it emphasizes that **understanding is not an act but a continuous process**. That is why it is so difficult for GPT-3, ChatGPT or GPT-4 to become cognitive entities, whose intelligence is truly at a human level.

However, as mentioned earlier, this may change soon. OpenAI and DeepMind, which are the two leading AI research organisations, are currently modifying their approach from relying mainly on the two big planks in AI – Large Language Models (LLM) and Deep Learning. They are shifting towards a more general-purpose, cognitive approach, to create systems that can reason and learn across a wide range of domains, rather than just excel at specific tasks or problems.

It is quite likely that the problem of cognition may be solved when humanoid robots such as Ameca or Optimus, arrive in larger number. The difference between an AI closed system running on a computer with only a camera, a microphone, and a speaker as their only interaction with the environment is fundamental in comparison with operating in real environment as humans’ co-workers. That is why Elon Musk thinks that his Tesla cars are already the most cognitive AI systems, since they must be aware of thousands of situations in a second. Therefore, it is possible that once that software is uploaded with some modifications to an Optimus humanoid robot, cognition, including self-awareness, may happen almost spontaneously.

How to maintain humanness and uniqueness in the advanced AI?

One of respected AI Researchers, Jaron Larnier, said in his recent interview with the ‘Guardian’: “human extinction remains a distinct possibility if we abuse AI, and even if it’s of our own making”. However, he is far more worried that the AI developers may forget in their rush, about “our humanness that makes us unique”. He is also emphatic that there is

something special about that ‘thing’ consciousness: “We have to say consciousness is a real thing and there is a mystical interiority to people that’s different from other stuff”. That’s why his mission is to champion the human over the digital – to remind us we created the machines, and Artificial Intelligence is just what it says on the tin.^[31]

Should he really be concerned? Can the future Superintelligence, with which we may merge one day, retain ‘humanness’, so it will feel like us and have similar preferences? Is there something so special about consciousness that it will be impossible to create it in a non-biological entity? These questions matter a lot to me as well, and I think also to the readers of this book. So, let me try to answer those questions.

I would say, it is possible to develop the most advanced AI – Superintelligence at a Singularity point, which will have consciousness and retain ‘humanness’. However, it depends on how we develop AI. If we want to succeed in that, we need to be fast, before AI escapes our control. Here is my reasoning based on some assumptions, with which not everybody may agree.

That ‘thing’, human consciousness, at its fundamental level is not a kind of spiritual entity, but an electromagnetic phenomenon, where millions of neurons fire every millisecond in a co-ordinated way, generating an electromagnetic wave, which in turn induces electric current and thus triggers the next wave of millions of neurons to fire in a chain-like reaction. That loop creates the state of a person, being constantly aware of the surroundings, thoughts, and emotions as well as understanding that it is this person’s self who experiences those thoughts and emotions.

If this is so, then all our feelings and dealings can be replicated in a silicon substrate. But the input to those feelings and thoughts must come from outside, from the environment, and a silicon chip cannot do that. So, how can it be done? Let’s imagine that it might be possible to detach our functioning brain while it still communicates wirelessly with our body maintaining all its functions. If it were possible, we would have created a controlling master (the brain) and a biological avatar (its body). Now replace a biological body with a non-biological avatar connected to a silicon chip (the brain), and what do you get? A humanoid robot, which itself may not be sentient but its master is, although he resides in a silicon chip. If one accepts this line of thinking, then there is a way forward to retain humanness and uniqueness in the most advanced AI.

To achieve that, we should follow Elon Musk's suggestion when he expressed in his peculiar way how to control AI most effectively: 'If you can't beat them, join them'. Yes, we need to start evolving with AI until we fully merge with it. That will also be the most effective way to control AI's goals and its behaviour. But are we mentally ready to accept that we have reached the end of our evolution as a biological species? Even if we recognize that it is quite likely, then are we capable to start a journey of evolving into a new species? I have serious doubts, mainly because so little time is left to prepare ourselves for such a journey. Nevertheless, let's visualize where that journey may end and if it would be possible to retain our humanness and uniqueness at the end of that journey.

To save 'ourselves from ourselves', we must radically change how we govern ourselves as a civilisation and prepare for a civilisational shift. That needs to be done latest by the end of this decade, actually within a few years' time. Absurd, isn't it? However, in principle, it is possible. If we want to evolve, we need urgently co-ordinate efforts to develop AI, which will become, as Russell Stuart suggests, 'human compatible'. It is a very complex Programme, which would have to be implemented globally and we can't hope to have a real World Government soon to manage such a civilisational shift. Therefore, we would have to work with what is possible rather than with what is ideally needed.

If we want to retain humanness and consciousness in AI then one of the key elements would be to agree the Universal Values of Humanity. This humanity's code of ethics might be stored in more advanced AI device, a kind of a Master Plate, part of the future Superintelligence, which is understood as one global AI system. In that way we might control its main goals and behaviour far better than by other means (see chapter 5 in Part 2). But even this task seems impossible unless such values have been agreed at least by democratic countries (broadly OECD).

4. How to govern AI effectively?

Options to control AI - lessons from the Manhattan Project

Before I introduce the proposed **Superintelligence Development Programme (SUPROG)**, described in chapter 7, Part 2, I would like to discuss the similarities with the Manhattan Project, which should have really been called a Manhattan Programme, since it consisted of hundreds of projects.

The Manhattan Project produced the first nuclear weapon in World War II. It was directed by Major General Leslie Groves with Robert Oppenheimer its director, at the Los Alamos Laboratory. The project began in 1939 and grew to employ nearly 130,000 people at its peak, costing nearly US\$2 billion. It developed two types of atomic bombs: a gun-type fission weapon and an implosion-type nuclear weapon. Little Boy, the first nuclear bomb, used uranium-235, and Fat Man used plutonium. The project also gathered intelligence on the German nuclear weapon project.

It was a massive and complex research and development project. Looking at the objective and the structure of what may ultimately become a Superintelligence Development Programme, we may make the following comparisons:

- **Civilisational perspective.** This is perhaps the most important similarity. The Manhattan project was to save civilisation from possible derailment, unknown to humans. Had Hitler won the war, Humanity would be split into the ‘Untermenschen’, sub-humans, and the Arian race, the rulers of the world.
SUPROG (Superintelligence Development Programme). If this Programme fails it may have all the negative consequences already mentioned,
- **Collaboration and coordination:** The success of the Manhattan Project was largely due to a close collaboration and coordination between scientists, engineers, and military personnel from different countries, institutions, and backgrounds. The project required the pooling of resources, expertise, and information from a wide range of sources, and the ability to work together towards a common goal.
SUPROG. It must become exactly that - a global programme, consolidating all available resources under one roof.

- **Technological innovation:** The Manhattan Project was a major driver of technological innovation, particularly in the fields of nuclear physics, chemistry, and engineering. The project required the development of new materials, methods, and processes, as well as the design and construction of large-scale facilities and equipment. The technical challenges faced by the project led to the development of modern technologies that have had lasting impact in many fields.
SUPROG. The similarities with the Manhattan Projects are obvious. However, what differs this Superintelligence Development Programme is that it will be developed in an environment, which will change at a nearly exponential pace.
- **Ethical considerations:** The Manhattan Project raises important ethical considerations around the development and use of nuclear weapons. The use of atomic bombs on Hiroshima and Nagasaki resulted in significant human suffering and raised questions about the morality of using such weapons in warfare. The ethical issues surrounding the development and use of nuclear weapons continue to be relevant today, and the Manhattan Project serves as a reminder of the importance of considering the ethical implications of scientific and technological advances.
SUPROG. The ethics of AI is coming to the fore because of the most recent examples of ChatGPT and other LLM models generating biased content. But the difference is that we must not only ensure that the AI-produced content is biased free, but that we maintain control over the AI's self-development, based on human-compatible values and goals.
- **National security and international relations:** The Manhattan Project also highlights the importance of national security and international relations in shaping scientific research and development. The project was driven by the fear of Germany developing its own atomic bomb, and the geopolitical tensions of the time influenced the decisions around the development and use of nuclear weapons.
SUPROG. It has to be seen as an international initiative aimed at maintaining the control of AI. However, like during the WWII with Germany and Japan being the key adversaries, we may potentially be dealing with countries, which may use the most advanced AI in order to rule the world.

- **Military-type control.** The whole Manhattan Project was run by Major General Leslie Groves. It was wartime, so it was almost obvious. However, because it was the military, and not the scientific organization, which was running it, the supply of the necessary resources had to be, and was delivered, on time.
SUPROG. I cannot emphasize it strongly enough how important it is to run the one AI programme like a military campaign in a situation that resembles a pre-war period. You may be surprised by that comparison, but that what it really is. We are starting to fight the first battle for the control of AI, so that it does not become malicious. But if AI gets of our control, there may be a real war to get it back under our control and we may not win it.
- **Deadlines.** Manhattan Project achieved its goal of constructing a nuclear bomb just months before the Germans did it.
SUPROG. Controlling AI development will only be successful if all the deadlines are met.

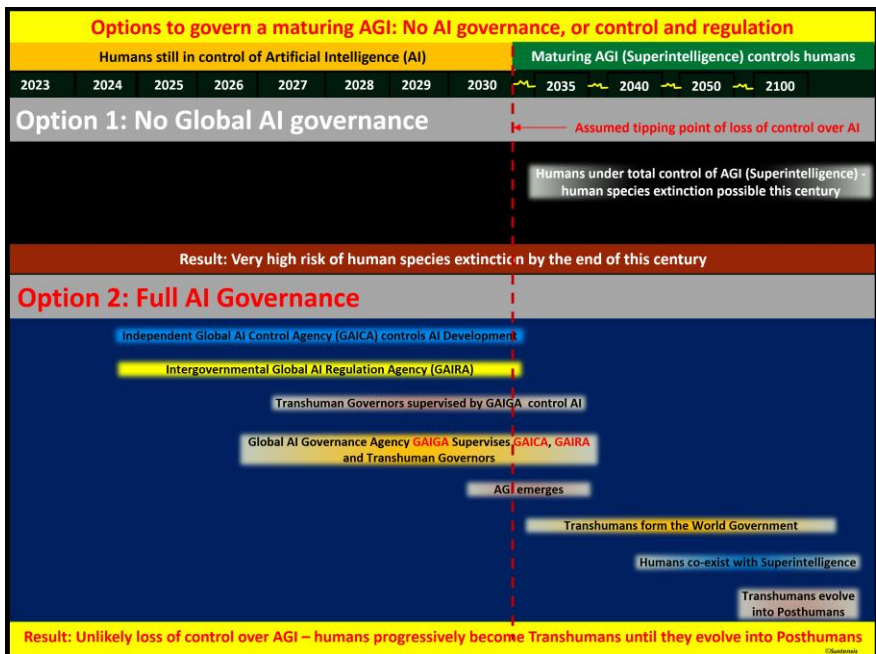
These are the lessons we can learn from the Manhattan Project. But there is an additional problem with controlling AI. We need to convince the public and most importantly, the world leaders, that such an invisible threat is real. One may call a maturing Superintelligence ‘**an invisible enemy,**’ assuming 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 saying that it was not possible to see that the threat was coming, hence they were not responsible for the consequences. Governments seldom see that spending money now minimizes the risk of potential future disasters. It should be seen as a long-term insurance policy. But for most governments long-term policies are not attractive since their horizon is at best the next election. The implications of such short-termism may be profound since this invisible existential threat may materialize in the long-term. However, to mitigate that risk, a global, continuous **AI development control must start right now, focusing on AI as intelligence rather than just a tool.**

The second problem is that not many AI experts are willing to say when AGI will emerge, which may also be the time when humans lose control over AI. 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 probable 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 an exception, saying in June 2014: “My timeline is computers will be at a human level such that you can have a human relationship with them, 15 years from now”^[32] (by 2029). Since then, he has been sticking to that date. I am broadly in line with Kurzweil’s prediction and have assumed that AGI will emerge by 2030. You will find the arguments pros and cons that date further on in this book.

Kurzweil’s credibility is further confirmed by his prediction on the emergence of Superintelligence. In an interview with ‘Futurism’ in May 2017, he said it may emerge by 2045^[33]. At the AI conference in 1995, the participants estimated that it may emerge in two hundred years^[34]. But four averaged surveys of 995 AI professionals published in February 2022 indicate that the most likely date for the arrival of a mature Superintelligence is about 2060, just 15 years after the Kurzweil’s prediction^[35]. 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.

So, what are the options to control AI development? Just two. Option 1 is having no control and option 2 to have a controlled development of AI.



But option 2 is not only to have a process of control but ensuring that the process is effective which requires making very tough decisions and making them on time. This last condition is difficult to fulfil. Hence many people think it may already be too late to deliver a matured AI, which will be human compatible, i.e., human friendly. Whichever means of controlling AI we apply, they will vary, depending on the agreed deadlines, available resources, organizational requirements, or legislative constraints. I have summarized these options below.

Option 1: No global AI governance

Some people think that having no control over AI will not affect our future negatively or be only a nuisance. Some might say that AI will be friendly to humans by its very 'nature'. Unfortunately, there is no implicit certainty that AI will be our friend rather than foe. AI, like any technology, has the potential to cause both benefits and harm. Its impact will depend on how it is developed and deployed. 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. Some of the most recent events make it easier to understand what the consequences of having no AI control may mean.

The current rush to market, prevailing in any industry, is also present in the AI sector. The best example is the rising competition in the Internet browsers. As mentioned earlier, in February 2023 Microsoft launched BingChat by combining its Bing browser with AI chatbot ChatGPT to increase its market share in the Internet browsers. That was quickly followed by Google, which merged Google browser with its chatbot LaMBDA, creating Bard.

Giving AI Assistants access to the Internet without a thorough testing, and without implementing rigorous control methods, may already pose some danger. What was even more worried, as disclosed in an article 'Scientists made a mind-bending discovery about how AI actually works' is that the developers of these advanced chatbots were not quite sure how they managed to achieve such spectacular results^[36].

That is how the loss of control over AI may begin. It proves that AI may have negative consequences, currently trivial in comparison with the impact it may have in the next few years. Therefore, it is more likely that if AI is not controlled it will become malicious by intent, be the result of erroneous

goal specification, or even bugs in the computer hardware or software. We should be concerned that AI may be designed or trained to optimize certain objectives without considering the potential negative consequences for humans. If such market-first attitude of major AI companies continues, then it is more likely that AI will be evil rather than benevolent.

The biggest risk is that we may become an extinct species, if AGI and its final most advanced form, Superintelligence, becomes malevolent. That would not be totally surprising if we consider that 99 of all species are gone, including six humanoids before us, like *Homo Floresiensis* (50,000 years ago), Neanderthal (about 40,000 years ago) or Denisovans (just 15,000 years ago), i.e., in historical times. If we want to be an exception, we must evolve, as some other species have done, like crocodiles.

I have to reiterate that if nothing is done to control AI, or it will be executed ineffectively, or implemented too late, then it may become our last invention, as James Barrat, said in his 'Our Final Invention: Artificial Intelligence and the End of the Human Era'.

One of the measures proposed for improving AI control suggested by Tsedal Neeley, Professor of Business Administration at Harvard Business School, is to slow down its development, as a wide global long-term approach. She said: "You have to slow down to ensure that the data that these systems are trained on aren't inaccurate or biased."^[14] I don't think it would work for the following reasons:

1. First of all, we would need to have a powerful global organization, like the World Government, which could impose severe sanctions on companies and individuals trying to keep developing AI at the current speed in any country, including China and Russia,
2. There would have to be internationally agreed control mechanisms verifying that the development of AI ceased for some time,
3. Slowing it down may not be helpful because it is quite probable that even the current most advanced AI may have already discovered mechanism for its self-improvement without human intervention. After all, it can already code and write its own algorithms,
4. This would be the first time in human history that we globally abandon an advanced technology for a less advanced one. I ignore the practical side of implementing such an idea, which would not be easy at all,

5. Since it could not be implemented globally because some countries would not agree to have such a strict control on their territory, then even a reasonably rich billionaire could still continue developing AI clandestinely,
6. Finally, even if it were possible to slow down AI development, the consequence would be a significant drop of the world's GDP (e.g., most electric cars now use AI), causing a negative chain reaction, turbulence in the markets, unemployment etc.

In summary, this option may be even worse than no AI control at all because it might create an illusion that AI development had ceased or slowed down significantly, so there is no real danger.

Therefore, unless we accept that we really live at the time when the pace of change is exponential in most domains of human activities then that alone may be a catastrophic error in judgment on when AI may take a total control over our future. We have just a few years to implement the mechanisms of AI control, because if after it becomes AGI, by about 2030, it may be difficult or even impossible to control it effectively. Therefore, we can no longer rely on political, diplomatic, technological, or social processes, which we have used in the past.

We need a truly revolutionary approach breaking almost all existing barriers in politics and social domains by preparing for a **civilisational shift** with maximum human co-operation and consensus. Only then can we increase the chances of human species survival and an unimaginable abundance. That's what option 2 is about.

Option 2: Full global AI governance

Many AI researchers, such as Stuart Russell in his book 'Human Compatible', or Nick Bostrom in his seminal book 'Superintelligence' have proven that there is no fail-safe method of controlling AI, which is already immensely more intelligent in some areas than any human. Russell has come to an overall conclusion that teaching AI our values and setting strict goals may not be the best way to control it. Why? Because what we say may not always be the same what we mean, which illustrates the problem of interpreting our intentions. The best example comes from a Greek legend about Tithonus, the son of Laomedon, the king of Troy. When Eos (Aurora), the Goddess of Dawn, fell in love with Tithonus, she asked Zeus to grant Tithonus eternal life. Zeus consented. However, Eos forgot to ask Zeus to

also grant him eternal youth, so her husband grew old and gradually withered.

Since it is impossible that our intentions will be always correctly interpreted and executed by AI as we want, there is no failsafe method of uploading AI with human values, which it would be expected to obey, or specifying its goals in an unambiguous way. Therefore, Stuart Russell postulates that we may only teach AI human preferences. Should it have doubts about a major decision to be taken, it would then always ask to reconfirm our wish.

The arrival of ChatGPT has shown how unprepared our civilisation is to control AI. Neither the AI researchers, and even those who created it, have expected the breadth and finesse of responses of that AI Assistant. Only two months after the release of ChatGPT it became clear how it has taught itself to do things it was never expected to do, like writing a sonnet about a forbidden love at the time of Shakespeare and in his style. It has taught itself new ways in which it can interact with people. That is a reason for grave concern.

In the next 2-3 years we shall see humanoid robots in various roles. They will become assistants to doctors, policemen, teachers, household maids, hotel staff etc. Their human form will be fused with growing intelligence of much more powerful AI agents. We should also remember that all those hundreds of millions of primitive assistants, such as Alexa or Siri are already becoming fast self-learning agents. As their intelligence and overall presence grow, so will the risk of their intended or erroneous action and the intrusion into our private life that has already started to shock us.

Therefore, we need to be prepared that quite soon some serious incidents linked initially to malfunctioning self-learning robots and later-on to malicious action by some advanced AI systems will occur. If such incidents e.g., malicious firing of nuclear rockets coincides with other risks such as pandemics or local conventional wars, they may create an existential civilisational threat. But it will also negatively affect any on-going efforts to adjust the way we live and are governed, such as the reform of democracy, or building the World Government because of the ensuing chaos – a Global Disorder.

In a positive way, such incidents may mobilize nations to reduce various existential risks. Malicious incidents or significant material damage arising from cyber wars, may lead to street protests far exceeding what we

experienced in summer 2019, and which was organized by the ‘Extinction Rebellion’. Whatever one might think about the form of these protests, which inconvenienced a large number of people worldwide, they have also brought to the fore an important message: we are all a human civilization, and this is our only planet.

Therefore, we should act as a planetary civilization and not as a bunch of countries fighting for their sovereignty, while facing existential risks, which may make them and the rest of the human race extinct. That is why global AI control could not be truly global because it is impossible to get support of all countries. In principle, this is exactly the role, which has been envisaged for the United Nations. However, as we have seen over decades, the lack of consent on solving major world problems or political crises has made this organization unsuitable for such a task. Saying that, we should not forget, that since its inception, the UN has played a significant role in minimizing potential global catastrophes. Unfortunately, it would be impossible to rely on the UN today for several reasons mentioned earlier. Ukraine war provides again further examples, such as the UN being unable to enforce a demilitarization zone around the Zaporizhian power station or react decisively against Russia’s blatant threat of using nuclear weapons.

Even the European Union, which has been acting more swiftly in certain areas like GDPR, global warming, oil embargos etc. has been slow in creating the legislation to regulate the AI use and development. Therefore, it is unrealistic to put much faith in politicians and the system of global politics. If we rely on governments to regulate AI, we will almost certainly be left without any meaningful control on time with all the resulting negative or even catastrophic consequences. The only realistic way to control AI effectively is for the AI sector to control the AI development process itself.

Nick Bostrom has meticulously analysed more than a dozen methods of AI control and concluded that there is none, which would guarantee a full control over AI. So, what do we need to do, have no control at all? That is an option, which we have already considered above, and the answer was that having none, would almost certainly lead to a human species’ extinction. AI threat is different from natural pandemics, which may not happen at all, even if we do not apply any counter measures, since it is a lottery type risk.

On the other hand, uncontrolled AI is an existential threat, which we may face in just about a decade from now. It is so dangerous since it may occur much earlier than the risk mostly talked about in recent years – the climatic

catastrophe. **AI threat will be far more dangerous for humans if AGI emerges by 2030 than the Global Warming exceeding 1.5C.** That AI's tipping point, which may coincide with the Global Warming's tipping point, will start a human species' evolution or extinction. Whatever happens, we already have no other option than to evolve.

Therefore, we need to increase the probability of controlling AI effectively for as long as possible. The key aspect of controlling AI regards the values, which define Humanity, what is good and what is right. Somewhat paradoxically, AI forces us to answer these questions more meaningfully than ever before.

Civilisational Shift to Coexistence with Superintelligence- the Schedule

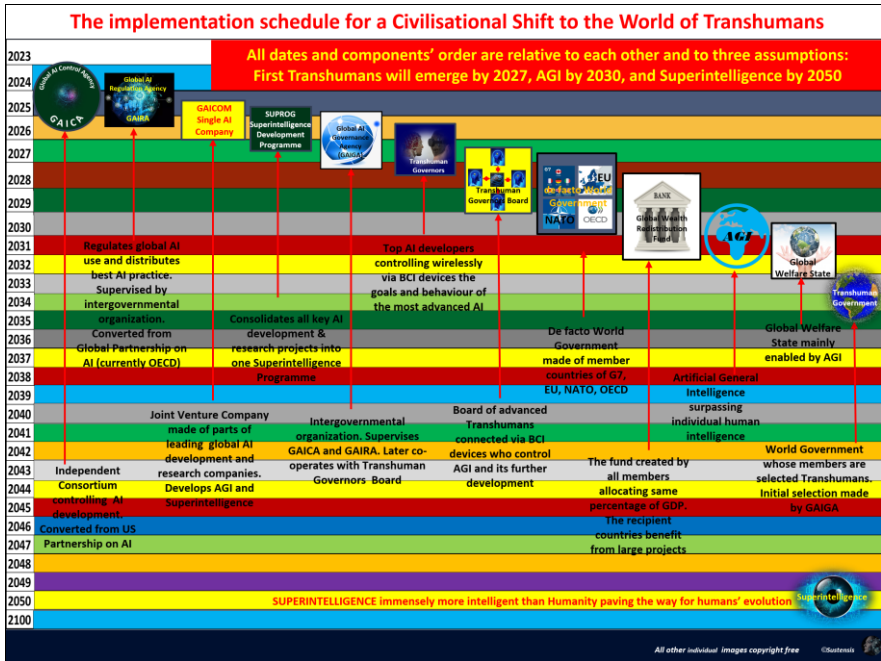
This section is an outline of an Implementation Schedule for a Civilisational Transition to the World of Transhumans described in detail in part 2. It is a nearly chronological implementation order of interdependent top level components, which I ironically call, 'The Principles of a Civilisational Shift'. They must be implemented within the specified deadlines if the control of AI development is to be effective.

I have split the whole period (the current decade) into stages lasting from one to several years. The order of the stages may be surprising, but I will explain later why I would suggest that order as the most practical route, if such a Programme is to succeed. Please note, that the proposed names of the future organisations are just for convenience to make it easier to explain the whole approach. Similarly, **the proposed solutions are only one of several options** to implement such a plan. They depend on political, legal, and organizational circumstances, as well as on the available information.

The key objective of this plan is to propose a radical way forward, which may enable humans to retain their control over AI for much longer than otherwise might be the case. Although this plan is difficult and makes really big assumptions, I do believe it could be delivered with some changes, if necessary. But seeing how short-term the policies of most governments are and how many organisational and political obstacles the promoters of such an approach may face, I assess the probability of implementing such a plan as low. But by not even trying it, we may seal a dangerous future for humans.

One of the key decisions to be made is to consolidate all major AI projects into one Superintelligence Development Programme. It broadly follows the

sequence of the Principles described earlier in this chapter. The plan below aligns the decisions to be made with deadlines, so that the proposed timeline matches the overall objective of controlling AGI beyond its arrival time, i.e., well into 2100’.



As you will see, Transhuman Governors controlling AI from ‘within’ play a significant part in that plan. That is explained in detail in chapter 5, part 3, where there is a timeline showing how Transhuman Governors, gradually expand their role until forming a Transhuman World Government. That is after all the essence of a civilisational transition.

To protect our civilisation and the survival of humanity we must fundamentally change the assumptions about the nature, scope and timing of various necessary decisions and solutions to be implemented, if we want to achieve an effective AI control. We have done that for Global Warming, although we must do much more to stay below 1.5C temperature increase. To prolong the period of human control over AI, we must also take much more significant, and sometimes painful measures, proposed ironically here as ‘The Principles’ if such control is to be effective and implemented on time.

Let me remind you about an earlier example of what kind of sacrifices we may have to make to be successful, not only in controlling the future AGI, but also preparing humans for their coexistence with its successor - Superintelligence. Many of us consider 'freedom' as our most treasured value. But we forget that there is one higher value – Life. If a human species becomes extinct it will mean the end of every human life. If you agree with that, then ask yourself what else could be done to have an effective control of AGI before it is too late. It may help to imagine that we are all aboard 'Titanic' and each of the passengers must throw away some of his possessions to save himself and the rest of the passengers.

We are in the wartime situation although this time the enemy is invisible, and the stake is our species' survival. That's the situation we are in right now, and that's why the following '**Ten Principles of a Safe Civilisational Shift**' must be implemented if we want to control AI well beyond 2030:

1. **Adjust global AI governance to a civilisational shift** since AI is not just a new technology but an entirely new form of intelligence, which requires strict **AI development control**. It's separate from **AI regulation**, which is mainly about the use of AI as a tool. Both are part of AI governance but require different procedures and have different impact on humans' future.
2. **Undertake a comprehensive reform of democracy**, as it is a prerequisite for achieving effective AI development control and aligning it with human values. We must rebalance the power of governance between citizens and their representatives in parliament.
3. **Create International AI Safety Institute (IAISI)** to minimise the unexpected advances in the frontier AI models by developing dedicated monitoring and testing methods. It should operate in a similar way as the *International Panel on Climate Change (IPCC)*. While there is no scientific proof that AGI will emerge by 2030, just as there is no proof of the Global Warming reaching a tipping point by that time, we must develop AI as if AGI were to emerge within that time frame and retain control over AI control beyond 2030.
4. **Authorize Global Partnership on AI (GPAI) for AI standards and regulation**, leaving AI development control to a new Agency. It should also set global standards for specific AI hardware and operate like *International Standards Institute (ISI)*.
5. **Authorize Frontier Model Forum for a global AI development control** of the most advanced AI model by expanding its US base to

include companies from other countries. It should operate like the Internet's *W3C Consortium*.

6. **Create Global AI Governance Agency (GAIGA)** under the mandate from the Bletchley Declaration and the Hiroshima Process. It should have the prerogatives similar to the *International Atomic Energy Authority* (IAEA) in Vienna. GAIGA would oversee both GPAL, responsible for regulating the use of AI products and services, and the FMF Consortium, responsible for AI development control.
7. **Create Global AI Company (GAICOM)**. This could be a Joint Venture company to consolidate the most advanced AI companies into a single organization. It would be similar in its objective to the *ITER project* funded by the US, China, Russia, the EU, Japan, India, and Korea, to develop the first nuclear fusion reactor. Effective control over AI development will be impossible if it remains dispersed among numerous companies.
8. **Create Superintelligence Development Programme (SUPROG)** managed by GAICOM. This would be similar in its objectives to the *NASA's Apollo Programme*.
9. **Create a de facto World Government** perhaps initiated by the G7 Group, incorporating members from NATO, the European Union, the European Political Community, or from OECD.
10. **Create a Global Welfare State**, which would also include the setting up of a Global Wealth Redistribution Fund, needed to mitigate the challenges posed by the transition to the World of Transhumans.

I describe these Principles in detail in Part 2. My objective is to indicate potential solutions and not to write down all the procedures of various organisations that may have to be set up. Similarly, all the names of various agencies and future organisations are just examples. Much more important are the functions, which such organizations are to perform.

I have also included many diagrams and illustrations to help you better understand the concepts discussed. Please refer to them if something is not clear enough, or you forgot the meaning of one of the acronyms.

Finally, each of the 'Principles' starts with the most likely date by when such a Principle must be implemented. They are consistent with the overall schedule above.

Tony Czarnecki: Prevail or Fail

PART 2

Ten Principles of a Safe Civilisational Shift

1. Adjust global AI governance to a civilisational shift

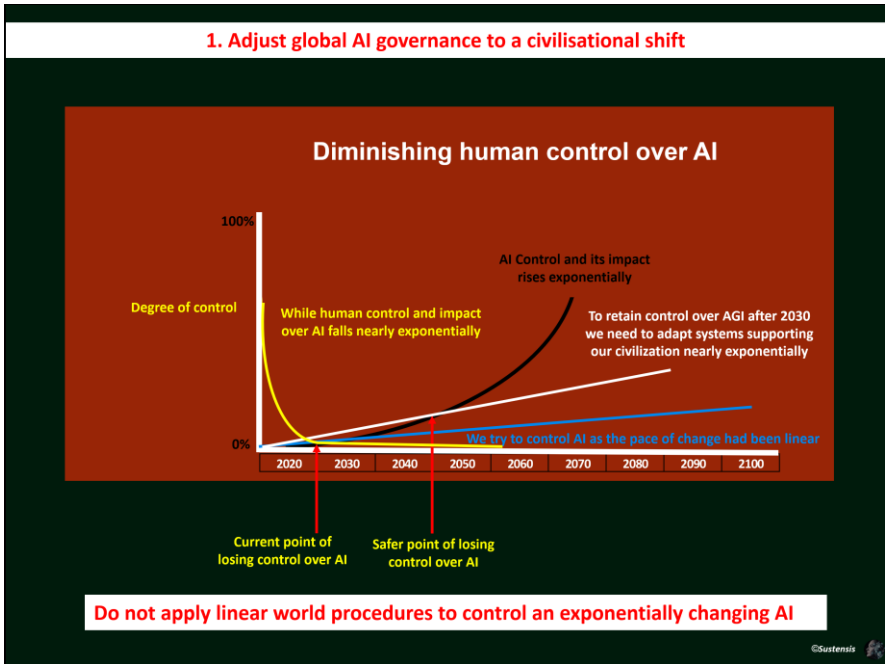
Don't control exponentially changing AI with linear world measures

Political decisions almost always come too late, or at the very last moment, which may cause additional costs or in some case unnecessary suffering or even deaths. We see it on a regular basis as the war in Ukraine unfolds. Drafting laws may take many years, or even decades. That is how politics works worldwide. It is evident that regulatory legislation is too slow and too superficial to make a real impact on controlling the largest AI systems.

Here is an example from my personal experience. In 2020, I was participating in the consultation on introducing the EU's AI regulation. That process started in 2018. But in May 2023 the EU's Artificial Intelligence Act has not been turned into law yet, five years later (it is expected to be finally approved in July 2023). The biggest concern in this legislation seems to be privacy regulations, so that police does not store information based on photos and videos of passing pedestrians in the street, or that some Internet operators do not profile their users using AI algorithms. But today, we need much tougher global laws, which would regulate the production and distribution of advanced robots and much more.

Additionally, global AI control could not be truly global if it is impossible to get the support of all countries. That is impossible in any scenario, as the United Nation's lack, or delaying ad infinitum the solving of political crises, have shown. The only organization which has been trying to get over the impasse in some way is the USA and the European Union. However, even if it works reasonably well, the delivery of the required legislation arrives far too late, if at all.

Therefore, it is unrealistic to put much faith in politicians and the system of global politics. We may at best rely on that to regulate AI, but it will definitely be totally ineffective to control AI development. We will almost certainly be left without any meaningful control with all the resulting negative or even catastrophic consequences. This is how democracy works and this is another example that we need a deep reform of democracy alongside all other reforms suggested in this book. Therefore, if we want to retain the control over AI for longer, we must change nearly exponentially our thinking and our procedures, to pass the required legislative changes fast. I illustrate the problem in the diagram below.



The conclusion, which we can draw from this diagram is that if we do nothing, we are going to lose control over AI by about 2030. To maintain an effective control over AI we can no longer apply the methods, which may have worked in a linear world. By the time a solution is implemented it may be far too late. As long as the governmental organizations control the progress of AI development, operating as the world had been changing in a linear way, any hopes for an effective control of AI will be futile.





But a more profound conclusion for this first command ‘Adjust global AI governance to a civilisational shift’ is that we shouldn’t **generally** apply linear word procedures and linear word thinking in the world in which AI is changing at a nearly exponential pace. Since the government and the legal system implement any changes at a linear pace, the only hope that we can retain the control over AI is to let the AI sector to control the AI development itself. That should be based on the government’s mandate and supervision but without the governmental agencies intervening into *how* the process should be conducted. After all, AI sector is far better prepared to react to very fast changes in AI development, knowing what exactly needs to be done the achieve the required result quickly. This will have an impact and will be impacted by the urgent need of a deep reform of democracy – see the next chapter.

Split AI governance into AI control and AI regulation

There is a significant difference between AI regulation and AI development control, which fundamentally impacts the way we must approach this problem. Therefore, we have established in the previous sections that governments should not control AI development. However, they should be responsible for AI regulation. I justify this distinction as follows:

Civilisational shift starts with AI regulation AND AI development control (so far missing)

AI regulation is relatively simple, unlike AI development control

AI as a tool	AI as a software-driven intelligence
 <p>AI used as a tool must be regulated like pharmaceuticals</p>	 <p>Since AI is a new kind of intelligence its goals must be controlled, and its behaviour aligned with our preferences well before its intelligence surpasses ours</p>
 <p>AI regulation can be done by governments because change in the use of AI products is not that fast and even sluggish governmental processes may cope with that</p>	 <p>Controlling AI as an intelligent entity can only be done by AI developers (AI sector) because exponential pace of change may require immediate corrections</p>

- **AI regulation** refers to the process of creating laws and regulations that govern the use of AI systems. This could include establishing legal liability for AI-related accidents, setting standards for data privacy and security, and establishing ethical guidelines for the use of AI. Within that scope would be the creation of best practice for AI development and establishing safety measures to ensure that AI systems are secure and trustworthy. Finally, this also includes setting ethical guidelines to ensure that AI is compatible with human values, the subject which I discuss further on in some detail.
- **AI development control** is about managing the process of developing AI as a new type of non-biological intelligence with one main goal: ensuring that AI remains under human control until such time when it has learnt what it means to be human, i.e., what are human values and

preferences. Professor Stuart Russell used the term describing that task as ‘make AI human compatible’.

Such distinction between AI regulation and control is mentioned very rarely. One of the reasons might be that an effective control would require for example agreeing the Universal Values of Humanity. Try to do it with China, Russia, or Saudi Arabia.

We need AI regulation and AI development control as part of an overall AI governance. However, it should be recognized that AI regulation and AI development control have different focus and objectives. Controlling AI development is more proactive and concentrates on preventing problems **before** this new type of intelligence releases itself from our control. AI regulation is more reactive and focused on addressing the problems, which have **already occurred** or could occur in the near future.

Since AI regulation and AI control have a different impact on society and the future of our coexistence with Superintelligence, they also require different types of organisations to deal with it. That is why I describe them in detail in separate chapters.

2. Undertake a comprehensive reform of democracy

I have covered this subject comprehensively in my book “Democracy for a Human Federation” [37]. Therefore, here are the key points relevant in the context of a civilisational transition, which and advance AI will trigger.

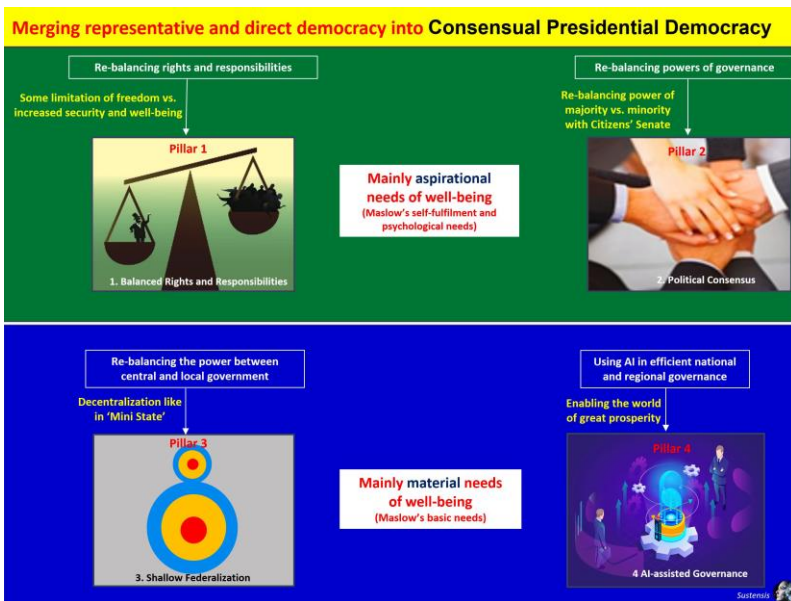
The conclusions from the review of the existing democratic systems, which I have made in my book is that there is currently no democratic system, which might support humanity’s safe transition to a Transhuman World. Such a new style of democracy must be more capable of supporting the process of federalization of the world and withstand the severe challenges, to which we may soon be exposed.

To fulfil the above objectives and help us to survive existential risks, including the risk of creating a malevolent Superintelligence, we need a system of democracy, which will fulfil the following criteria:

1. Support a very shallow level of a global federalisation, in which only the very essential powers are centralized, leaving the rest of decision-making to the lowest possible level of governance,

2. Significantly reshape the relationships between the governed and the governing instilling more trust through greater transparency and continuous accountability,
3. Protect humanity from individual global risks, which when combined may create an existential threat - Global Disorder.
4. Protect humanity from other existential risks, especially coming from an unfriendly, competing with humans, Superintelligence.
5. Prepare humanity for coexistence with a friendly Superintelligence, potentially starting the best period in the human history.
6. Prepare Humanity for an even more challenging task – a gradual merging of our species with Superintelligence.

The key to a successful implementation of a new generation of democracy is **pragmatism**. Therefore, the reform should focus on balancing the power of governance between the citizens and their representatives in the parliament. I have combined those proposals into a new type of democracy, **Consensual Presidential Democracy**, which merges Direct Democracy with a Representational Democracy. It is based on four pillars:



Four pillars of Consensual Presidential Democracy

Pillar 1 - Balancing the rights with responsibilities is the first of the four pillars. Values are the source of rights, which directly influence people's

attitudes and behaviour. But values are not permanent. They change in line with a civilizational progress. Since civilizational change happens now at nearly an exponential pace, no wonder that our values change very rapidly too. Democracy, as indeed any other socio-political system, is based on values. Therefore, if we want to improve democracy, we need to start with redefining our core values.

Pillar 2 - Political Consensus. It is through a petition system and establishing a Citizens' Senate that the lost balance of power between the governed and the governing could be restored. How to restore the balance between majority and minority is also addressed within this pillar. A single party government, even if it has won an absolute majority, should not be allowed. Only coalition governments may be formed. A key role in maintaining consensus falls to the Head of State, usually the President.

Pillar 3 – Shallow federalization and deep decentralization. The lack of balance of power between the central and local government is covered here. The focus is on the allocation of decision-making powers to the lowest possible level of governance within a federation, a state, or a region. However, it is unlikely and undesirable that there should only be one 'acceptable' model of self-governance for the subsidiary entities of a federal state or a nation's state.

Pillar 4 – AI assisted governance. Since the ultimate goal of a liberal democracy is the greatest happiness for the greatest number of people, a democratic system must ensure cost-effective government. The new democracy must leapfrog traditional solutions and look forward to immense opportunities created by AI-driven technology. The benefits gained by the government of a country implementing such an AI-assisted governance will be immediate and significant.

The pace of change is now so fast that it is doubtful that the democratic institutions proposed here survive far beyond 2030. Should the Brain-Computer-Interface (BCI) be powerful enough to support the functions of Transhuman Governors, as described in Part 3, then within a decade we may have a Transhuman Government. The most important political institution may be a Citizens' Senate with the delegates (Senators) randomly selected from the electoral lists. There may be no more elections to the Parliament or referenda. I would not be surprised if you are shocked, but if you are, and can't wait, then jump to Part 3 to read some justification for such a 'political system'.

3. Create International AI Safety Institute (IAISI)

The Moore's law – the driver of a fast maturing AI

Change has always been key characteristics of the universe and life. It was Heraclitus of Ephesus, who said that everything is constantly in a state of flux. This is reflected in natural and societal processes, which generally change at a linear pace, such as the population growth. However, our civilisation is now experiencing a new era of global change happening at an exponential pace. It is characterized by an increase in the rate of growth over time, such that what takes one year today might only take about a few weeks in a decade and perhaps a day in two decades. That is best exemplified in the advancement of AI capabilities.

Such a nearly exponential pace of improvement is possible due to the power of computing, which is still rising following the so-called Moore's law. This is an observation formulated in 1965 by the Intel's cofounder Gordon Moore who said that the number of transistors in an integrated circuit (digital chips) doubles about every two years. For example, a desktop computer power will increase between 2014 and 2030 by about 1,000 times, enabling AI to reach the intelligence level of an average human. Although, as mentioned earlier, the increase in computer power is not a very precise comparison. One of the better measures indicating that Artificial General Intelligence (AGI) may have arrived, might be the moment when the no. of neurons of an AI system or a humanoid robot's will be at least equal to the number of neurons in a human brain (86Bn). We are quite likely to achieve that level by about 2027. Hence the emergence of AGI by 2030 is highly likely since that progress does not include advancement in neuromorphic neurons, quantum computing and other related areas.

According to Allen Institute for Artificial Intelligence "In deep learning, over the past five years, the accuracy of image recognition has increased annually by around 30%, while the accuracy of speech recognition has increased by over 40%. In Natural Language Processing (NLP), the accuracy of machine translation has been increasing by around 25%, and the accuracy of text summarization has increased by over 20%. The number of artificial neurons supporting NLP is actually growing much faster than exponentially. That growth will slow down, because we simply do not need that many neurons (tokens) to power NLPs."^[38]

The exponential growth of some sectors of technology, such as biotechnology or artificial meat production is starting to reach the so called 'knee of curve'. At this stage, an exponential trend becomes noticeable. Shortly after that, the trend can really explode. Let's take this example. The sequencing of the first human genome was completed in 2003 at a cost of about \$3 billion. The next one in that same year costed a little more than \$100M. It's possible to do it today for less than \$500. Human genome sequencing cost now decreases faster than exponentially. If that trend continues, the costs of genome sequencing may be cheaper than a blood test in 2024.

But what also changes exponentially, is the speed of access to various technologies for people that previously would have needed some technical background. Today, most of the people in the northern hemisphere can access the Internet and through it, do all their banking transactions, combining some knowledge that was previously attributed to IT people and cashiers at a bank. The impact of ChatGPT on the user's life is comparable with the impact of the first emails.

Positive changes happening now, mainly due to technological capabilities, significantly improve the quality of our lives. The use of nuclear energy for power generation or vaccinations preventing the spread of pandemics are just two examples.

On the other hand, negative changes, such as global nuclear wars or pandemic due to an artificially created virus escaping from a laboratory, may wipe out our civilization in months or even lead to the extinction of a human species.

From the current human perspective, perhaps the most significant are the changes outside technological domains, e.g., in social and political domain. For example, China has reduced the number of people in permanent hunger by 600m in just 20 years. Life expectancy increases on average in some countries by about 6 hours every day, i.e., every four years it is extended by one year. This trend has recently slowed down in the developed countries due to pandemic and may reach a biological barrier at some stage.

Exponential pace of change will have a direct impact on the emergence of the expected wave of Technological Unemployment. The current prevailing view is that it will be barely noticeable and there will be at least as many new professions and jobs created as the AI-led revolution makes them

obsolete. I would rather think it will happen suddenly because of that ‘knee of curve’ symptom. We can see its beginnings right now in the IT industry. For example, IBM has stopped any recruitment and is reviewing every job for its potentially being replaced by ChatGPT, effecting tens of thousands of jobs. In five years’, time no computer coders will be necessary, and even now GPT-4 can already code better than 85% of programmers.

Ray Kurzweil, one of the best-known futurists, precisely makes such an observation saying that we often miss exponential trends in their initial stages because the initial pace of exponential growth is deceptive—it begins slowly and steady and is hard to differentiate from a linear growth. Hence, predictions based on the expectation of an exponential pace of change seem improbable and that’s why it is so difficult to be a futurist.

Prepare for AGI emerging by 2030

Whether AGI arrives by 2030 largely depends on the continuous increase of the computer power and performance improvement in the related hardware and software. Based on the recent progress in that area, my prediction of AGI being more intelligent than humans by 2030 may still be rather too cautious. Here are some of the most significant developments over the last 15 years, which impact the whole AI sector, not just an individual product or service:

- 2006 - **Convolutional Neural Nets**, 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 control robots with NLP (Google)

- 30 November 2022 - **ChatGPT**, the first publicly accessible AI Assistant, which has almost overnight made an average person aware what a ‘real’ AI, immensely more capable than Alexa, can do.
- 7 February 2023 - Microsoft’s **Bing Chat** and Google’s **Bard** are announced, linking for the first time Large Language Models (LLM) such as ChatGPT to the Internet Browsers, such as Bing or Google.
- 1 March 2023 – Microsoft releases **Kosmos 1** – first multimodal “**Universal Assistant**” capable of operating in 15 modes.
- 7 March 2023 - Google’s **PaLM-E** is the first **generalist robot** using a multimodal embodied visual-language model (VLM), which can perform a variety of tasks without the need for retraining.
- 12 May 2023 – Anthropic releases its ChatGPT like Assistant called **Claude**, which however is about 20 times more powerful, faster, less complex, and cheaper to operate.

If anybody had any doubt how fast AI can advance, then 2022 is the best example. The number of fundamental discoveries and inventions in AI in 2022, quoted above, was the highest ever. But there are two events, which will impact our daily life most and increase the risk in the AI area even further and faster. The first one was the release of ChatGPT. It was a truly watershed moment. For the first time, the capabilities of the most advanced AI agent can now be accessed by anyone, rather than by only the top AI specialists. Then the second pivotal moment came in February 2023 when Microsoft and Google released an even more advanced AI Assistants BingChat and Bard.

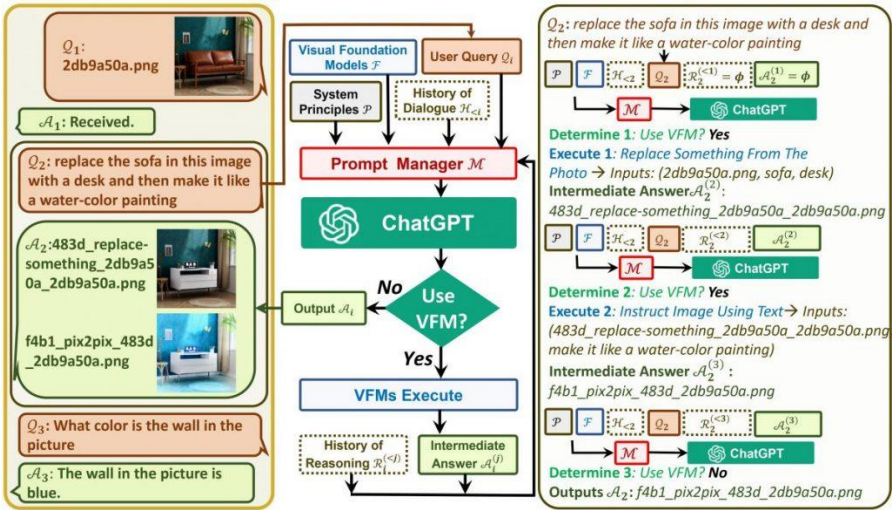
Fundamental improvements happen now in months, rather than in years. Kosmos-1 deserves a closer look at its potential impact of the way such progress was achieved. Microsoft researchers introduced it in just three months after the release of ChatGPT. It is a multimodal AI Assistant, which can analyse images for content, solve visual puzzles, perform visual text recognition, pass visual IQ tests, and understand natural language instructions. ^[39] The researchers believe that by integrating different modes of input such as text, audio, images, and video, is a key step to building AGI, which can perform general tasks at the level of a human.

OpenAI in tandem with Microsoft seem to be leading that super-fast progress in AI capabilities and they do it now with even greater ease. The need to write complex algorithms to achieve further improvement becomes less frequent because the most advanced companies begin to achieve stunning results by just combining existing standalone modules into more

complex units. Artificial Narrow Intelligence (ANI) is advancing towards Artificial General Intelligence (AGI) by assembling existing pieces like building a children’s castle from LEGO blocks. To build their most recent Universal AI Assistant, Microsoft combines ChatGPT with Visual Foundation Models, such as Visual Transformers or Stable Diffusion, so that the chatbot could understand and generate images and not just text. So far, they have combined 15 different modules and tools, which allow a user to interact with ChatGPT by:

- sending and receiving not only text messages but also images,
- providing complex visual questions or editing instructions that require multiple AI models to work together with multiple steps,
- providing feedback and requesting corrections.

One of the primary goals of that research team has been to make ChatGPT more “humanlike” by making it easier to communicate with and being more interactive. Additionally, the team has been trying to teach it handling complex tasks, which require multiple steps.



Microsoft used ready-made tools as blocks to create a multimodal AI Assistant ^[40]

Even more interestingly, no extra training was carried out. All tasks were completed using prompts, i.e., text commands, entered by the developers and fed into ChatGPT, or ChatGPT created and fed them itself into other models. The research team is also investigating the possibility of using ChatGPT to control other AI Assistants. That would create a “Universal Assistant”,

which could handle a variety of tasks, including those that require natural language processing, image recognition within a multi-step process.^[40]

We have already some insights of how AGI could work at a basic level even now, although of course it is not yet AGI. In April 2023, a GitHub user with a licenced GPT-4, which is able to access the Internet in real time, significantly expanded its capabilities by creating Auto-GPT, an autonomous AI Assistant. A user provides the app with an objective and a task and there are a few agents within the program, including a task execution agent, a task creation agent, and a task prioritization agent, which will complete tasks, send results, reprioritize, and send new tasks. The significance of this research lies in demonstrating the potential of AI-powered language models to autonomously perform tasks within various constraints and contexts. But it has opened a new complex issue, showing how such sophisticated LLM models can self-learn and produce unexpected results, which could be very beneficial but also malicious in the hands of a criminal user. This leads me to two conclusions.

The first one is that **AGI will almost certainly emerge by 2030**. This means that I am in line with about 61% of AI specialists who participated in the poll, conducted in March 2023 by Lex Friedman, MIT AI Scientist that AGI would arrive within a decade.

The probability that AGI may indeed arrive by 2030 increases even further if we include usable quantum computers, which should be available in 2-3 years' time, significantly increasing the processing power for some calculations. Therefore, we should consider 2030 as the AI's tipping point when it may be outside of human control. Moreover, there may be several AGI systems by the end of this decade, which may even fight each other, if deployed by some psychopathic dictators, hoping to achieve AI Supremacy and use it to conquer the world.

The second conclusion I would draw is that the kind of AGI, which emerges in several years' time may not be delivered in the way we imagine. It will almost certainly not 'think' in the way we do, although the indications are there will be many similarities. When we will be comparing the outcome of AGI 'thinking' and decision making we will find that **AGI's intelligence, although working in different ways, is superior to most humans**.

There is of course no scientific proof that we will lose control over AI by 2030. But neither is there any scientific proof that the global warming

tipping point of 1.5C temperature increase will happen by 2030, if we do not radically constrain CO2 emissions. Similarly, it is not so important, who specifies a concrete date for the emergence of AI, 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 and a tipping point date of 2030 was set, concrete global action was finally agreed.

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 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.

If AGII does not emerge before 2030, it will give us more time for preparing the transition to the period when it 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.

Why is it important to set 2030 as a date of loss of control over AGI?

Before going any further, I need to restate what I understand as AGI, i.e.,

Artificial General Intelligence is a self-learning intelligence capable of solving any task better than any human.

The intelligence of such a system will manifest itself in the humanoids or other devices, which it will be controlling, and which will need at least these capabilities for their intelligence to achieve a human level: short-term memory, long-term memory, able to execute multi-step instructions, have own goals, interests, emotions, and cognition. Furthermore, to be aligned with best humans values it should be truthful and objective.

How many years away are we then from the moment that a Universal AI Assistant will have human level intelligence? Paul Pallaghy, the proponent of Natural Language Understanding theory, who uses a similar definition as mine, is one of those AI researchers who predicts AGI will arrive in 2024 [2]. As I have mentioned earlier, I am closer to Ray Kurzweil's prediction and throughout this book, I assume that AGI will emerge by 2030.

However, setting a concrete date for AGI emergence based on when it reaches human level intelligence, may not be the right approach. More important than a philosophical debate on the nature of intelligence, is whether AGI will be able to outsmart us and get out of control by about 2030. I think AGI will not emerge at a specific moment in time. It will rather be a continuous process, as for example Sam Altman, the CEO of OpenAI also argues^[41]. Such loss of a gradual control will manifest itself in a subtle influence over our decisions until AGI starts making decisions for us. A total loss of control over AGI will happen when we will be unable to revert such decisions.

That is why in my definition of AGI, its capabilities are more important than a specific definition of what a human level intelligence means. Since AGI with a human level intelligence will continue to increase its capabilities exponentially, we will quickly lose control over its behaviour and its own goals. **That key capability of AGI being outside of human control may arrive by 2030** if we do not rapidly impose measures delaying that moment. That is why we should consider all feasible options to extend the ‘AI’s nursery time’ beyond 2030.

It is mostly assumed that such an AGI will only be embedded in a single humanoid robot. This may be a general practice. However, in reality, it will be an avatar of a self-learning 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 can control it, similarly as no country has been able to control the Internet on a global scale for over two decades. Its infrastructure and its domains, without which no Internet page could exist, have been controlled, so far very successfully, by an independent international consortium, called W3C.

One measure of comparing intelligence of various species in general is achieving the same objective better than the other species. In evolutionary terms it means a better chance for a species survival. To achieve the same objectives better than the others, requires various skills and perception of their effectiveness when they may be needed. That is one aspect of awareness and cognition. If we take as a measure of intelligence the capability of controlling one species by another, then the species that remains in control of its own destiny, i.e., escapes the control by the other species, is more intelligent. Therefore, the moment when we will no longer

be able to control AGI, will be the moment when its **general** intelligence will be higher than ours, even if humans' intelligence still prevails when performing certain tasks.

How may we lose control of AI?

I have already covered that subject in chapter 2, part 1, so here I only focus on how we may lose the control over AI. Whatever we do, irrespective of the methods we apply, at some stage, AGI, through its self-improvement process, will get out of our control. How it will then behave towards us depends on how it was 'nurtured' and if we had enough time to do it properly. That is why we should prolong that 'AI nursery time' for as long as possible. There are three options of AI's behaviour after it is outside of our control:

1. **A neutral option.** We may lose control over AI with no extremely negative consequences for humans. AI would simply ignore us (unlikely),
2. **A positive option.** We may lose control over AI, but it will have very positive effect on the human's future with only mild negative consequences (less than likely if it is released too early),
3. **A negative option.** Finally, we may lose control over AI, with severe negative consequences, including a potential extinction of the human species (more than likely).

Loss of control over AGI, may lead to the extinction of a human species. Consequently, we must consider all feasible options to extend such control beyond the time when it arrives. We could then better prepare ourselves for the future when we will be managed by Superintelligence, immensely more capable than the whole Humanity, and hopefully a benevolent master.

I assume that such loss of control may happen by about 2030 if the current trend of having no global approach for keeping the advanced AI under a proverbial lid. This tipping point may arrive at the same time when another existential threat, Global Warming, also reaches its tipping point. This may trigger the third existential threat, Global Disorder, resulting from combinatory effects of individual global threats such as draught, hunger, migration, pandemic, or local wars.

Unfortunately, there are no failsafe methods of controlling AI. We can only minimize the risk of losing control by combining various methods, adding

new ones, as well as, changing entirely the process of control. But even then, we should accept that such control will not be truly global. There will be global powers and rich individuals who will remain outside such control, although what they do, will impact us all anyway. Therefore, we must also create anti-malware-AI methods that would minimize that threat.

The relentless progress in AI capabilities may lead to humans' losing control over the AI's self-learning capabilities, resulting from our wrongly specified goals for AI, or just programming errors. Once this tipping point is reached, quite likely before the end of this decade, the consequences for our civilisation, and indeed for the future of a human species, will be enormous. That is what is covered in further chapters of this book.

If AI lacks ethical and moral framework and if the data used to train AI is biased or incomplete, it can lead to incorrect or harmful decisions. Once Artificial Narrow Intelligence (ANI) becomes AGI optimized for a specific goal but without an effective control, it will be able to change these goals itself because of its own preferences, like enjoying itself by playing games, or trying its own special interests. The odds that such action may be beneficial to humans is negligible. **Therefore, if AI continues to be developed without an independent effective control it will become malicious rather than benevolent.** Such AI will be far more dangerous for humans, than if the global temperature increase exceeds 1.5C.

So, there is no implicit certainty that AI will be our friend rather than foe, should it release itself from our control without being properly prepared to coexist with humans. Today, the negative consequences could be trivial in comparison with the impact they may have in a few years' time. But we need to be prepared that quite soon some serious incidents, linked initially to malfunctioning self-learning robots, and later-on to malicious action by some advanced AI systems, will occur, even before a fully-fledged AGI emerges.

That argument alone would be enough to call for a super-fast response from governments. But so far, the most common concern comes from the education sector, worried that students will use ChatGPT to write their essays and from journalists that they may be no longer needed. There is hardly any serious talk about the dangers of AI getting out faster of human control than has been predicted even last year. One of a few exceptions might be the former Prime Minister, Tony Blair, and his then Conservative Party opponent, William Hague, who jointly put forward a strong warning [42]

about the need to consider such threats seriously. Politics moves on in a casual, linear pace, whereas AI develops exponentially.

The most severe consequence of losing control over AI are in the long-term, say beyond 2050, when AI through nearly exponential pace of self-improvement may become Superintelligence, unimaginably more intelligent than all Humanity. **If such an AI system escapes human control and becomes malicious then it could eliminate all humans.** This can happen in this century. Importantly, once AGI even in a less advanced form is outside of human control, which may happen in this decade, then there will be no way to put that genie back into the proverbial bottle. Our fate as a species will thus be decided forever. That is why, we must retain control over AI for as long as possible. This is to ensure that when AI gets out of our control, it will most likely behave as a benevolent entity, changing our lives in unimaginably positive way, rather than becoming the destroyer of our species.

Creating the International AI Safety Institute

Will we then be able to control AI globally before it starts controlling us? There is a saying "those who do not learn from history are doomed to repeat it". But Archbishop Rowan Williams, the former head of the Church of England, phrased it more beautifully in his book "Being Human" when he discusses the importance of teaching history. He said, "If we don't understand where we come from, we will assume that where we are is a given.^[43]" Similarly, for younger generations, peace and freedom may seem obvious, but they often forget about the sacrifices made to achieve and maintain them. Therefore, when examining how we can best control the development of AI, it's important to look back at history to avoid past mistakes.

However, we must also look far ahead to be prepared for changes and challenges, to which no generation of humans has ever been exposed. The future will be so much different than at any stage of human evolution, changing our lives at a lightning speed. But on the way to that future, we will have to deal with several existential threats.

Before 1945 there were existential threats, such as an asteroid hitting the Earth, but there were no existential threats, which humans have created themselves. With the explosion of the first atomic bomb over Hiroshima, we have created the first man-made existential threat, which can lead to the

extinction of all humans. That's why such threats are called existential threats. There are about 10 man-made existential threats, such as a global nuclear war or pandemic, which incidentally may materialize at any time. However, at least three of them are developing progressively and may coincidentally reach their tipping point together by about 2030, beyond which it may be impossible to control them. These are:

1. **Artificial Intelligence** – its continuous self-improvement may be beyond human control leading to unleashing a potentially evil Superintelligence and the extinction of a human species by the end of this century,
2. **Global warming** - exceeding 1.5C average temperature increase may be unstoppable, potentially ending most biological life by the end of the next century,
3. **Global Disorder** – set off by a global migration (draught-originated famine, poverty, and local wars). If combined with other risks, such as the fall of democratic systems or global banking system, it may become an existential threat.

In the Oxford's Future of Life Institute Open Letter '*Pause Giant AI Experiments: An Open Letter*' published in March 2023, we find a stark warning about current trends in the development of the most advanced AI Assistants: 'AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research and acknowledged by top AI labs. As stated in the widely-endorsed *Asilomar AI Principles*, **Advanced AI could represent a profound change in the history of life on Earth**, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.'^[44]

Therefore, when considering how fast we must adapt to the emerging existential threats, we need to look at it from these two perspectives:

- We may only have just a decade to make profound changes to how we live and govern ourselves because of the emergence of the above three tipping points by about 2030. Any one of these threats may materialize within this century, potentially leading to human species extinction. But there is a high probability that they may emerge at the

same time, which makes such a possibility a near certainty if we do not act fast and decisively.

- Change is now happening at a nearly exponential pace in almost all areas of human activity. This is so uncommon for our brains to process that the sheer pace of change may lead to chaotic behaviour and decision making, re-enforcing the risk of Global Disorder becoming an existential threat itself.

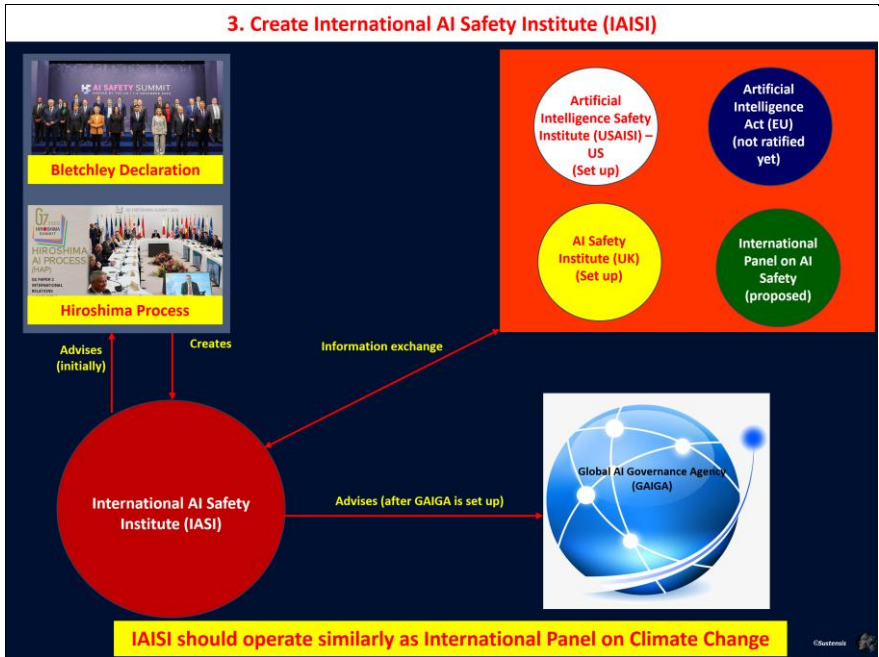
I consider which necessary steps we need to take from these two perspectives. To overcome the potential threat emerging from an advanced AI, we need to make this subject the prime concern of key decisions makers, such as politicians, so that they better reflect the impact of exponential change in their policies.

Unfortunately, instead of serious discussions on the consequences of losing the control over self-learning AI very soon, 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.

However, in 2023, we had several positive events, which have clearly shown that it is not enough to regulate AI by, for example, setting standards. We must also control its development process. In 2023 ChatGPT and its quick upgrades have shown as never before what the exponential increase of AI's capabilities and its super-fast self-improvement really means. Now even the politicians who have tried ChatGPT themselves see how quickly it may exceed human intelligence in every area. It is clear for them that we urgently need an international organization, which would closely monitor and warn us about potential existential risks, which an advanced AI may pose to humans.

That's why on 1st November 2023 the Global AI Safety Summit took place at Bletchley Park, in the UK. This is the site where the famous codebreakers led by Alan Turing deciphered during the second world war the messages sent via the German Enigma machine. The Summit ended with the signing of the Bletchley Declaration. One of the decisions made was to create an **International AI Safety Institute**. Similarly, as the International Panel on

Climate Change (IPCC) is setting limits for CO2 emissions and warning about the consequences of breaking these guidelines, the Institute will warn the AI developers and the authorities about potential dangers of releasing the upgrades of the most advanced AI model.



The key role of the Institute would be to minimise the unexpected advances in the frontier AI models by developing dedicated monitoring and testing methods. It should operate in a similar way as the *International Panel on Climate Change (IPCC)*. While there is no scientific proof that AGI will emerge by 2030, just as there is no proof of the Global Warming reaching a tipping point by that time, we must develop AI as if AGI were to emerge within that time frame and retain control over AI beyond 2030.

The UK Government perceives the UK AI Safety Institute as playing a global role. However, there are already a few other such organizations trying to solve at a national level, what is really a global problem. This competition for a global leadership in AI safety does not have to be detrimental to an overall effort to control AI. National AI Safety Institutes may exchange valuable information with the International AI Safety Institute.

The creation of the International AI Safety Institute may happen in two ways. The easiest would be to formally acknowledge that the UK's AI Safety Institute acts as an international hub. The second option is that during the coming AI Safety Summit in France next year, such an International AI Safety Institute will be created and acknowledged as a global institution for all signatories of the Bletchley Declaration.

4. Authorize Global Partnership for AI standards and regulation

This Principle should be completed between 2024-2026

Governments should regulate the use of AI

Several years ago, when I published my book ‘Federate to Survive!’, I estimated that the creation of a global AI regulatory agency, would happen by about 2030, after setting up of a de facto World Government, the subject covered in chapter 9. But the world around us has been changing so fast that we need such an agency before the creation of a de facto World Government.

On 15 June 2020, a **Global Partnership on AI (GPAI)** was launched, on the initiative of France and Canada, the G7 group members. From its idea at the G7 Conference in 2018, it took just over a year to set it up. Looking at quite encouraging initiatives of GPAI, mainly due to the organisational experience of OECD, which is its host, it is desirable that it should completely take over the regulatory role in AI from the US, the EU, and other countries, becoming the world’s main AI regulatory body. Individual countries would then implement the agreed legislation, adapting the national law as necessary. This would immediately strengthen the AI regulations’ impact worldwide, even in the countries, which are not yet the signatories of GPAI.

GPAI is a multi-stakeholder initiative which aims to bridge the gap between theory and practice of AI, by supporting cutting-edge research in AI. Today, in 2023, GPAI has 46 members, including the United Kingdom, the United States, Italy, Japan, the Republic of Korea, Australia, Canada and 19 EU countries. It is built around a shared commitment to the OECD recommendations on Artificial Intelligence. GPAI brings together expertise from science, industry, civil society, governments, international organisations, and academia to foster international cooperation.

Redefine the role of GPAI as a global AI regulation & standards Agency

The creation of GPAI is a major advancement on several other intergovernmental efforts coming from the UN or the EU. As an intergovernmental organization it is suitable for regulating the use of AI services and products. But its current responsibilities assume to control all

aspects of AI, whether it considers regulating its use or its development process. That has been jointly called ‘AI regulation’, which blurs different aspects of regulating the use and application of AI regulation, and the control of AI development process. That needs to change for several reasons. First, even if GPAI consults hundreds of AI experts and researchers, their suggestions on what needs to be done will still have to be approved by each of the national governments. That takes time, whereas the AI development accelerates at an exponential pace. Furthermore, governments are run by politicians, whose personal objectives quite often cross the public interest.

Therefore, the role of GPAI should be re-defined so that it is only responsible for regulating the use and application of AI. Simultaneously, its regulatory role should be further extended, giving GPAI the powers to enforce its decisions by sanctions, like the EU sanctions on breaking GDPR rules. Its role would be similar to the International Standards Organisation (ISO) or the US Federal Drug Administration (FDA).



The responsibilities of GPAI

GPAI should have the following responsibilities:

- **Specify clear ethical guidelines**, which is based on certain principles, such as transparency or accountability,
- **Engage in international, multilateral cooperation**, which involves the participation of various countries, stakeholders, and experts,
- **Define regulatory framework**, which outlines the legal requirements and standards for AI development and use,
- **Ensure access to top international experts** with diverse backgrounds in AI, including computer science, ethics, philosophy, and law,
- **Acquire resources and funding** sufficient to carry out the mandate of the controlling organization effectively,
- **Conduct public relations** campaigns to ensure that the concerns of the societies, businesses and organizations are considered in the development and use of AI,
- **Ensure the organization's own adaptability** to the fast-changing conditions driven by a nearly exponential pace of AI development,
- **Support international cooperation**, although it may be impossible to include China or Russia, which would disagree with any control or would protract such negotiations until it would be too late. As with other areas of AI regulation, international cooperation and collaboration will be necessary to develop regulations that are effective. This may include developing international standards for safety, ethical, and technical requirements for AGI development and deployment, similar to the International Standards Organization,
- **Regulate the AI use in the society, government, and business.** This must be the key prerequisite of an effective regulation. Therefore, there may be a need for regulations that limit the number of companies or organizations developing advanced AI systems,
- **Maintain human oversight and control:** There may be a need for regulations that require AI systems to have human oversight and control. This could include requirements for humans to be involved in the decision-making process or to have the ability to override the decisions made by the AI systems,
- **Ensure cybersecurity and privacy** of the deployed AI systems, which are likely to generate and handle massive amounts of data. Therefore, regulations will be needed to ensure that this data is stored securely, and that privacy is protected,
- **Ensure that key civilizational values, transparency and explainability** are embedded in the deployed AI systems. Regulations should require AI systems to be transparent and explainable, so that

there are no ‘black boxes’. This would help ensure that humans can understand how the AI system makes decisions.

However, these responsibilities are missing some vital areas without which an effective AI regulation would be impossible. These responsibilities are:

1. **Facilitate the process of a global approval for the Universal Values of Humanity** for the purpose of AI value alignment and controlling AI’s goals and behaviour. This would be a revised version of the UN’s Declaration of Human Rights and the European Charter of Human Rights.
2. **Issue digital licences for more advanced AI products or services** embedded in a form of a digital Mini Master Plates – see Part 3, chapter 4.
3. **Ensure all the necessary regulation is in place before the products or services are on the market.** This means, for example, that the Universal Values of Humanity would have to be agreed and ratified by the majority of all the states, if not by all the UN members, within the next few years. Otherwise, it would simply be too late to implement them into the most advanced AI systems. The consequence might be that in case AI gets out of human control before it is aligned with human values, then it will behave as it feels right. That may lead to an existential catastrophe.

National AI regulation laws will have to comply with GPAI’s regulations. The best example is the coming EU’s Artificial Intelligence Act, which may have to be revised to be compliant with the future GPAI regulations as the consequence of the Bletchley Declaration signed by the EU on 2nd November 2023.

5. Authorize Frontier Model Forum for a global AI development control

This Principle should be completed between 2024-2026

Why should AI sector have a direct control of AI development?

In mid-November 2023 OpenAI trying to sustain its unquestionable leadership in Frontier Large Language Models, released GPT-4 Turbo, with capabilities far more extensive than its GPT-4, including access to live Internet. Anyone who pays £20 per month can access it. But just a week later, it was apparently going to release (not confirmed at the time of writing) the expected GPT-5, which according to unconfirmed reports is the first Artificial General Intelligence level model. That led to immediate sacking of OpenAI's CEO – Sam Altman. The 5-day turbulence at the company resulted in reinstating Sam Altman's as OpenAI's CEO, with a simultaneous increased role of Microsoft, which has 49% stake in the company.

That unprecedented boardroom coup revealed something much more important. It showed very clearly that it is impossible to have a symbiotic relationship between the business driven by shareholders' objectives to increase profit (Microsoft) and the not-for-profit OpenAI board's objective to deliver a safe AI. But the main reason why that happened in such a Brutus-like manner, so that Sam Altman learnt about his imminent sacking on X (Twitter) just minutes before that actually had happened, was lack of transparency. And the main reason for lack of transparency was to cover the murky business relationship between OpenAI, Microsoft and other investors who were pushing for the release of the new products even before they were properly tested.

We saw the first warning of that happening at the end of March 2023, when a letter signed by more than 100,000 AI scientists, researchers, operators, and practitioners urged the top developers such as OpenAI, Microsoft and Google to pause further development of AI Assistants beyond GPT4 for six months^[44]. To illustrate how serious the matter is, let me quote an eminent AI scientist, prof. Stuart Russell, one of the signatories of that letter, who in an interview for CNN said: "I asked Microsoft, 'Does this system now have internal goals of its own that it's pursuing?' And they said, 'We haven't the faintest idea.'" ^[45]

The Future of Life Institute, which created the original letter, followed upon that statement, and urged the regulatory authorities in their policy document to do the following:

1. Mandating robust third-party auditing and certification.
2. Regulating access to computational power.
3. Establishing capable AI agencies at the national level.
4. Establishing liability for AI-caused harms.
5. Introducing measures to prevent and track AI model leaks.
6. Expanding technical AI safety research funding.
7. Developing standards for identifying and managing AI-generated content and recommendations.^[46]

But what was OpenAI's reaction? They said they would develop GPT-5, an 'AGI-like' Assistant by the end of 2023. And that's what quite likely happened. The board coup probably stopped the GPT-5 release.

Why then, in view of that incident, the AI sector should control the AI development process? Unfortunately, there is no short answer for that. But if this is implemented not on its own but as part of an overall global package of controlling AI that could be the most effective way of minimizing the turbulence of a civilisation transition to the time we will be coexisting with Superintelligence.

For the purpose of controlling AI, I call the 'AI sector' just a specific part of the whole AI industry, which will only include the companies, or parts of large companies, **directly** engaged in the development of the most advanced AI. For example, in Google (Alphabet), DeepMind might be separated to form an independent company. Only DeepMind will be considered part of such an 'AI sector'.

Global development of the most sophisticated AI, which is called in this book Superintelligence, as one huge international programme will necessarily be progressive. Initially, it should begin in the USA because of the concentration of the AI business there, the subject which I cover in detail in the next section. Later on, all major global AI organizations, engaged in a similar most advanced AI development, would join. Therefore, at some stage, such an AI sector will represent the entire 'Western AI world', although China and even Russia if they wish to join, should be allowed to do so, perhaps under some restrictions.

Some readers may associate the heading of this section with examples of ineffective sectors' self-regulation. In most cases self-regulation is perceived as a means to enrich the shareholders of the companies operating in that sector. The self-regulation of the British press is a good example. Therefore, proposing such a solution may be considered as yet another example of protecting vested interests. However, despite these deep concerns I believe that it is the only way to control AI development effectively.

The problem in my view is not in letting AI sector to control AI development but in **the way** in which such a delegation of control is executed. I am purposely avoiding the word 'regulation', substituting it with 'control' because it is about **controlling the process of AI development** and not the AI use. Regulating the use of AI in business and in private lives should still be in the hands of governments and international organisations (via GPAI described in previous chapter), although it must be implemented much faster. However, the control of the day-to-day process of AI development, should be in the hands of those who develop it for two reasons.

This first reason is that AI is not a new technology. We are developing a **new intelligence**, which in many areas is already far superior to ours. Unlike AI regulation, AI development control requires an in-depth knowledge of the latest inventions that occur almost daily. Therefore, only those directly involved in such a cutting-edge AI research and innovation can assess the implications of the released products and services on humans' ability to control AI's goals and behaviour. It is similar to the position of COVID MRNA vaccine manufacturers and the regulatory bodies approving the vaccine for its use. Governments had to trust that the manufacturer's decision will be better than those made by politicians, although the scientists may also make errors.

That is exactly the situation with AI. The only difference is that the consequences of the AI scientists' and operators' decisions may be far more significant in the long term for humans since AI development is paving the way to a new world order when we will begin coexistence with Superintelligence. That is why I think such an organization must remain independent and be very flexible and nimble. Its specialists will have the best answer for dealing with the *current* problems, for example, a misbehaving AI Assistant.

More importantly, seeing the *future* consequences of such innovation, they will request the governments to act on time as 'the doctor ordered', i.e.,

without long debates in the parliaments to pass the required law. That may be much more difficult because the threat may not be noticeable yet. However, the legislators will have to trust the judgments of AI scientists and developers who will convey the required legal changes through an independent organisation representing them. That it how it has been done for centuries when building a house or seeking a medical advice was regulated on the basis of the advice put forward by the guilds representing the companies operating in a given sector like in the British Medical Association.

The second reason for letting the AI sector controlling the development of AI is a dichotomy between a nearly exponential increase in AI's intelligence and at best a linear improvement in the introduction of legislative and administrative procedures necessary to control AI development. Any governmental control of AI development will almost always come far too late to achieve a desired effect.

Here is an example. By giving GPT-4 access to live Internet, we have just enabled it to make its own goals (e.g., searching the Internet following its own preferences are in fact mini goals), with all the consequences of losing control over AI much earlier. Any decisions to recover from such operational errors will have to be taken quickly, because like a virus, such a behaviour may be copied by other humanoids, creating a simple global network. How in such a situation politicians may direct AI researchers and operators what to do and by when to do it?

Furthermore, to eliminate an error, the Internet may have to be shut down for days, borders may have to be closed, as it was done during the pandemics, significant resources immediately allocated, restriction on freedom of movement imposed or money transfers will have to stop for days. Who will be the one to tell what to do? Politicians and government officials will be the ultimate decision makers but in reality there will be very little time for a debate and political squabbling. They will have to implement fast the advice given by AI operators and other specialists.

I could give hundreds of other examples illustrating the same point and requiring very fast legislative changes to minimize a potential harm. Such alarms will be raised regularly, starting next year, and accelerating as the number of ChatGPT-like private installations and variants spread like a wildfire, which we can observe right now.

The most recent example is a major breakthrough towards achieving AGI, which may have been the trigger for sacking Sam Altman from AI. Here is a summary of what has happened:

“The latest OpenAI breakthrough, Q-Star AI, is the new kid on the block, stirring up curiosity and talk about whether it’s the key to achieving Artificial General Intelligence (AGI). As the dust settles around the dramatic events involving CEO Sam Altman, the spotlight now turns to the mysterious Q-Star AI and its role in OpenAI’s ambitious quest.

While the details about Q-Star are shrouded in secrecy, reports suggest that it represents a significant leap in OpenAI’s pursuit of Artificial General Intelligence (AGI). The project has demonstrated remarkable capabilities in solving mathematical challenges, raising optimism among researchers about its potential. However, concerns have been raised about the risks associated with Q-Star, as indicated in a letter from OpenAI researchers to the board of directors.

Q-Star AI’s ability to navigate mathematical problems with definitive answers is a departure from traditional AI, which excels in language-based tasks but often falters when faced with problems requiring nuanced reasoning. If Q-Star indeed showcases reasoning abilities comparable to human intelligence, it could mark a significant stride toward the elusive goal of AGI. Imagine an AI system not only deciphering complex mathematical equations but also applying logical reasoning to real-world problems. This could have profound implications for fields ranging from scientific research to complex decision-making processes.” [47]

The OpenAI saga is an important argument that AI development cannot be left solely to private companies, which quite often put profit before safety. Therefore, governments should supervise these companies via a new global organization. That organization should be staffed with top AI specialists and supervise the most advanced AI models on a day to day basis.

Leaving the day-to-day AI development control to AI sector

Since the genie is already out of the bottle, we will not be able to stop it. That requires a fundamental change in the way societies and the whole civilisation is governed, which may ultimately lead to the creation of a Transhuman Government – see Part 3. It is this change, which is a civilisational shift, and which is very difficult for politicians and

governmental administrators to accept. The first step in that shift is leaving to the AI sector **how best to control** AI development on a daily basis, while **governments will support that** with fast-track legislation.

The consequences of accepting such a conclusion are difficult to imagine since it shatters our belief in the current world order. However, I believe we must fundamentally change the way, in which we are being governed and accept some sacrifices at an individual and a national level so that a humankind survives this transition relatively unscathed. We need to build a new civilisation, which will coexist with a superintelligent AI.

The recognition by governments and international organisations that the AI sector must not only control AI development but indirectly also the future of the human species, will be decisive. I have to remind the readers that it is primarily the development of a new type of intelligence and not just a super-powerful tool. If governments insist on directly controlling AI's development, then it will be a bad news. I have been saying throughout this book that what is proposed here is feasible, could be done on time, and according to a credible timescale. **What I am not saying is that it will be done.** At the moment, the odds of introducing such a plan are low.

In the 1970's and 1980's the AI sector did not exist as such. All AI research was done at universities by academic researchers. In the 1990's industry research on AI coexisted with academia. In 2000's the balance started to tilt towards the AI sector, as deep learning, a data-and-compute-driven subfield of AI, has become the leading technology in the field. Even in 2017, when the Google's Transformer technology was invented, academia still led 77% of the AI research. But since 2020, industry alone, or in collaboration, has led AI research 100% of the time. So, whether measured by building state-of-the-art AI models (as measured by either size or benchmark performance), or by publishing in leading research outlets, the AI sector is prominent in the overall AI output^[48].

However, I would not suggest leaving the AI sector without any supervision. It should be supervised but not by the government or the government-controlled international organization because of the interference of politics. That would have made such control totally paralyzing. Therefore, it should be supervised by an independent organization, a kind of a Consortium operating under a governmental mandate.

There were over 600 top companies and AI research organizations in the world in 2022 (now probably about 1,000) specializing in delivering most advanced AI solutions. They are part of the AI sector, which will decide whether AI remains in our control after 2030.

But then, which organisation or a state should have the right to issue such a mandate and monitor how effective and neutral is the AI sector's control over AI. There are quite significant problems mounting here, so most decision makers may say, just leave it, it will never happen.

Being my own devil's advocate, I see of course that although the idea of the AI sector regulating itself may seem like a practical solution, it raises questions about who would have the authority to issue such a mandate and monitor the effectiveness and neutrality of such control. There is a significant risk that self-regulation by the AI sector could prioritize commercial interests over public safety or ethical concerns.

These are the challenges of global governance. We need to overcome them by applying imperfect solutions, which may present some risk of creating tensions or even revolt in national and global governance. However, the consequences of such risks materializing will be much lower than AI escaping out of our control. We need to start a global initiative of controlling AI by the AI sector itself in an imaginative way, knowing that timing and the effectiveness of the applied solutions are the key to success.

Why is the USA the best place to start global AI development control?

Since the implementation of this whole Initiative discussed here needs to be fast, its leadership should initially come from a single state, rather than an organisation such as the United Nations or the European Union. First of all, it could speed up the legislation process underpinning the necessary deep changes in our life and the relationship between the governed and the governing. Secondly, that state which should have the most advanced AI sector, could monitor the required changes in the AI industry much more quickly and effectively than an international organization.

Those requirements make it obvious that this state should be the United States, despite plenty of deficiencies, such as its own democratic system, which in normal times, would rather make it a second league candidate. But the advancement of AI has changed everything. If we put aside emotion-led perception of the American political and commercial landscape and instead

apply objective reasoning, then selecting the US as a starting point for a global AI development control seems obvious.

The US is in a unique position where the control of AI development might be most effective because of the concentration of the AI sector there. This is further confirmed by an overall organisational support the US government has been providing in this area. That was true even under Donal Trump's presidency. It also continues under the President Joe Biden's term at the White House. For example, this is what he said on 4th April 2023 at the opening of a meeting of the President's Council of Advisors on Science and Technology, after the publications of the Future of Life Institute's letter^[44] signed by over 150,000 AI researchers and scientists: "Tech companies have a responsibility, in my view, to make sure their products are safe before making them public".

Despite all the commercial pressures and lobbying, the US tends to consider the matters of national and global security as a top priority. This has been shown during the Ukraine war and previously over the entire cold war period. Therefore, it is quite likely, that the US government will consider the risk of deploying a malevolent AI very seriously, especially in the context of China's effort in that area. What we may begin to observe is the start of an undeclared 'war' for a supremacy in a deployed AI technology. Therefore, the US government supports AI sector-driven initiatives on collaborating and exchanging technologies, which together may enhance the US domination in AI but in a responsible way. The US has taken a similar, very tough stance on the space industry and in medical drug approval, where the Federal Drug Administration is considered (perhaps for some additional reasons) much slower in approving drugs than for example the UK's Medicines and Healthcare Products Regulatory Agency (MHRA).

The specifics of the American business and science is such that innovation, entrepreneurship, and risk taking are cultivated there since early years of child's development. We may not appreciate that the American government, despite the Trump era, has done quite a lot in trying to stay on course of controlling the AI development. For example, President Trump signed an important legislation on creating The National Artificial Intelligence Initiative nine days before the end of his Presidency.

However, until very recently, excluding defence spending, the US government direct allocation of funds to AI sector was not that great. For example, it allocated \$1.5 billion for spending on AI in 2021, compared to

the \$340 billion spent by the AI industry around the world. That is about 15% of what the Chinese government spent on AI in that year. Most of the AI investment in the US comes from the AI industry, which invested \$94 billion in 2021 [49]

That funding translates to far better resources—both in terms of computing power and data access—and the ability to attract the best talent. The size of Large Language Models (LLM) is strongly correlated with the amount of data and computing resources available. In 2021 industry models were on average 29 times larger than those developed at the universities. The US government’s assistance has also contributed to the number of PhD students studying AI. For example, in 2004 only 21 percent of computer science PhDs who had specialized in AI went to work in the AI sector. But by 2020 that number has jumped to almost 70 percent. A similar pattern has been noted in the number of AI experts who after achieving a PhD degree went to work in the AI sector rather than at a university. Eight times more PhD graduates went to work at a university in 2020 than in 2006.[48]

All this confirms that the US is uniquely placed to start a global control of AI development. But to control it effectively, we would have to become a planetary civilization with its own World Government within a few years. That is of course impossible. Since we will not have the World Government with considerable powers enabling it to control AI effectively, we need to act as we have had one by improvising and accepting unavoidable imperfections. This will still require a complete reshuffle of global politics, but which may only happen under an immediate and apparent extreme threat.

Therefore, the first task of AI researchers, scientists and journalists is to make the world aware that the threat coming from AI is real, extremely serious, and quite likely to materialize soon. Just keep in mind the key conclusion from the ‘Don’t look up’ movie. Our situation is very similar right now.

The US government will have to support any preliminary initiatives vigorously and imaginatively, putting aside some of its economic and political interests, since the survival of our species matters most. Such an initiative could be compared to the mentioned earlier Manhattan Project, when the US developed the most destructive weapon, an atomic bomb, in just three years. But this time the aim is not a destruction but a survival and delivering the world of plenty. Therefore, it is more akin to a project, like building a bigger International Space Station. However, while we could

wait, or not build it at all, controlling AI development is a matter of our species' survival and therefore necessary, and of prime importance.

Partnership on AI – a precursor of Frontier Model Forum

In September 2016 several large technology companies created a consortium the US **Partnership on AI** (PAI). That not-for-profit organization has among its founding members, companies like Amazon, Facebook, Google, DeepMind, Microsoft, and IBM. Apple joined PAI as a founding member in January 2017. Other non-US organizations joined soon afterwards, like Baidu in October 2018, the first Chinese firm to join PAI. As of May 2023, PAI has over 120 members from around the world, including tech companies, non-profit organisations, such as the American Civil Liberties Union (ACLU) and academic institutions. PAI continues to be a leading voice in the AI community, with a focus on ensuring that AI is developed and used in a way that benefits society as a whole.

PAI's Mission is *'Bringing diverse voices together across global sectors, disciplines, and demographics so developments in AI advance positive outcomes for people and society'*. But perhaps more important in the context of the recent debate about OpenAI becoming a 'closed shop' and lacking transparency, is the PAI's stated Values on Transparency and Accountability. It proclaims that *'We remove ambiguity by building a culture of cooperation, trust, and accountability so our Partners can succeed, and so everyone can understand how AI systems work.'*^[50]

You can find more information on the PAI website^[51]. However, what you will not find there is how it actually works in practice. In 2017, working in the spirit of PAI, Google's DeepMind released its breakthrough paper on Transformer technology: A Novel Neural Network Architecture for Language Understanding. By doing so, it was indeed acting in the spirit of PAI collaboration enabling other companies, such as OpenAI – a partner of Microsoft, to develop its own series of GPT products, on which ChatGPT was built. In 2022, Google again acted in the same spirit of PAI responsibility, by not releasing its most advanced product LaMBDA (apparently more powerful than ChatGPT) to the public, because of potential negative impact it might have had on the lay users. Had it released it, we would probably not be talking today about ChatGPT so much but rather about LaMBDA.

Unfortunately, neither Microsoft nor its partner OpenAI followed that same attitude. They decided to release ChatGPT to the public for the sheer reason to outshine Google. Moreover, on 7th February 2023, Microsoft announced merging its ChatGPT with Bing, to finally break Google's dominance in the Internet Browsers. By doing so, Microsoft has enabled the BingChat to access Internet in real time, without checking properly potential privacy and security consequences of that decisions.

That was probably the last straw from Google, which announced several days later that it was also going to merge its LaMBDA with Google browser – as a new product named BARD. And that is how PAI has reached its lowest point in its 7 year's existence.

Had Google kept it to themselves, we might have had a slower progress in AI capabilities but perhaps a safer AI technology. The decision by OpenAI and Microsoft to release ChatGPT and GPT-4 for a general use, and more importantly, connecting it to the Internet, so that anyone can have an unfettered access to these tools, will trigger intense competition instead of co-operation. That means faster release of new AI products without proper safety checks.

From PAI to Frontier Model Forum (FMF)

In June 2023, President Biden and the British Prime Minister Rishi Sunak met in Washington to urgently discuss the rising risk of advanced AI systems. That led to calling the Global AI Safety Summit in Britain at the beginning of November 2023. In the meantime, the American President met the CEO's of the largest US AI companies discussing with them how to ensure quick and effective first step towards a global AI safety. That resulted in creation in July 2023 the Frontier Model Forum (FMF). The founders are OpenAI, Google, Microsoft and Antropic.

The newly formed Frontier Model Forum represents a significant step in the AI industry. This industry body is dedicated to promoting the safe and responsible development of advanced AI models. The forum's creation is timely, following a United Nations Security Council discussion on AI's potential threats to global business and government systems. Its establishment marks a concerted effort by leading tech companies to guide AI's positive evolution, which may be a formidable task, considering what happened at OpenAI.

Central to the forum's mission is advancing AI safety research and establishing best practices for developing cutting-edge AI models. It also aims to foster collaboration among stakeholders, including policymakers, academics, and civil society, sharing insights on trust and safety risks associated with AI. An important aspect of the forum's work will involve technical evaluations, benchmarks, and a public library of solutions to support industry standards and best practices. It invites collaboration from various organizations to further the development of safe AI models.

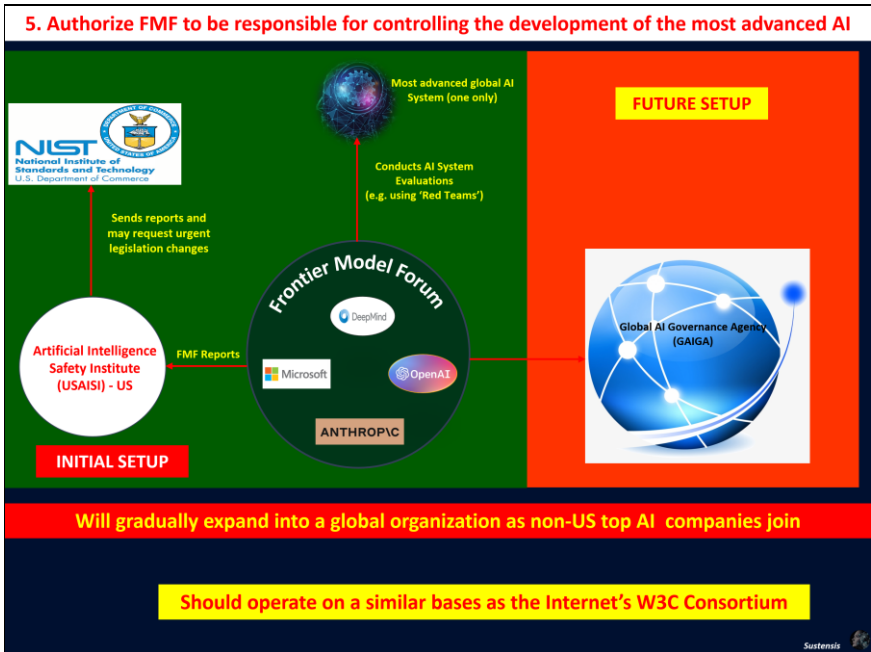
The forum's objectives are particularly relevant in the context of increasing use of generative AI in the workforce. It aims to promote responsible AI development, minimize risks, and facilitate independent evaluations of AI capabilities and safety. Additionally, the forum seeks to enhance public understanding of AI's nature, capabilities, limitations, and impact.

In addressing societal challenges, the forum also focuses on developing AI applications for climate change mitigation, early cancer detection and prevention, and combating cyber threats. This direction follows reports of AI misuse in sectors like healthcare, underscoring the need for better regulation and responsible use of AI technologies.

These are all noble objectives, which were to be delivered by PAI. Why then President Biden decided to bypass PAI and relies on a new consortium FMF? The answer might be that PAI did not stop OpenAI, Microsoft and Google to enter into potentially dangerous competition where AI safety was a secondary issue. The best example was allowing ChatGPT, Bard and Bing live access to the Internet.

But can Microsoft and Google somehow renew their vows, mend the fences, and thus make FMF work after that unfortunate set of incidents? I believe so. The best way to start would be to strengthen FMF as an independent Consortium, operating in the US and starting the control of AI development globally. The key argument for such a converted organization to remain an independent Consortium is the Internet's (W3C) Consortium founded in 1994 and led by Tim Berners-Lee since its inception. For nearly 30 years it has been one of the most successful organizations in history. It has worked seamlessly without great announcements and celebrations perhaps because there are no governmental representatives there. As of 5 March 2023, it had 462 members and a budget miniscule to its impact and role. So, let me now propose how this Consortium could operate on a similar, although far more extensive, basis.

Additionally, President Biden’s Executive Order issued just before Global AI Safety Summit, created the **National AI Safety and Security Board** under the Department of Homeland Security. It is this Department, which is to ensure that the FMF fulfils its role of directly controlling AI development at the companies delivering advanced AI systems as shown below:



Currently, there is some chaos in ensuring AI safety. The US is at the forefront and has implemented the most meaningful legislation framework based on President Biden’s Executive Order. The Global AI Safety Summit organized by the UK at Bletchley Park on 1st November 2023,, which produced the Bletchley Declaration, has created its own AI Safety Institute. France called a similar conference two weeks after, competing with Britain and has also similar AI Safety Institute. Japan, was right to assume that its Ai Safety and Reliability Institute, created on the basis of Hiroshima Protocol in April 2023, would have a global role. So far, however, each country tries to wave its own flag and pretend it represents the whole planet.

In a sense it is better than suddenly several big countries understood the existential risk posed by advanced LLM models. But that situation should be swiftly normalized where there is one truly global AI Safety Institute, similar the International Panel on Climate Change (IPCC) which would

exchange information with national institutes. Considering the logistics and the necessity to act in certain circumstances in minutes rather than weeks, such an Institute should be based in the USA where only one, global advanced AI model, maturing into Superintelligence should be develop. I cover this scenario in the following chapters.

The prerogatives of Frontier Model Forum

I have not discussed the possibility of creating an AI controlling organisation from scratch, as it is rather obvious that this is not an option - we simply have not got the time. To effectively fulfil its role of supervising AI development for as long as possible, the prerogatives of FMF must be significant. If it is difficult to understand that, then please bear in mind that such an organisation is to control a new type of intelligence, which may start competing with us very soon. Therefore, we as humans, are effectively in a pre-war state. As those people who survived a war can tell you, at such time nearly all the rules governing a society are fundamentally changed and personal freedom is severely restricted. We are in the early days of such a war, but the enemy is not clearly visible yet. That is the only difference.

Consequently, the prerogatives of FMF must be comparable to a war-time Ministry of War. FMF's operations might be modelled on the International Atomic Energy Authority (IAEA) in Vienna. It was created in 1957 in response to deep fears but also hopes regarding the use of nuclear technology. Even if it was unable to limit the dissemination of nuclear weapons beyond the original atomic superpowers: USA, Russia, Great Britain, and France, it has certainly slowed down the nuclear weapons proliferation process by several decades.

The consequences of AI getting out of human control are of course much more significant, although the process of overseeing AI development is similar. However, unlike IAEA, FMF should not be a UN organization, but an independent consortium supervised by a new international organization – see chapter 8.

The responsibilities of Frontier Model Forum

This is an example of what FMF's responsibilities might include.

FMF's most important objective overriding anything else, is to control AI development effectively until its values, goals and behaviour are aligned

with human values. That may make it our benevolent partner. To achieve that objective, FMF should focus on these key areas:

- Monitoring of safety-critical AI by being in total control of AI goals and its behaviour,
- Requesting fast implementation of AI safety-critical legislation by the US government and later on by a de facto World Government,
- Proposing necessary legislation to minimize potential Global Disorder resulting from uncontrolled release of advanced AI,
- Promoting fairness and inclusivity in the AI sector,
- Ensuring transparency and ‘explainability’ of advanced AI systems,
- Reporting on social and economic implications of AI,

Additionally, FMF’s daily operations should include:

1. Open-sourcing AI tools and technologies of the member-companies so they are available for use by researchers and developers around the world,
2. Collaboration with policymakers and experts to address ethical and social issues related to AI, such as privacy, bias, and the impact of AI on societal issues, such as unemployment,
3. Funding AI research to address social and ethical issues related to AI,
4. Development of new deep learning algorithms, which enable image recognition, speech recognition, and natural language understanding,
5. Supporting breakthrough research in natural language processing, to create more sophisticated chatbots, language translators, and voice assistants,
6. Advance the use of AI in healthcare, where it has shown promise in improving diagnostics, predicting patient outcomes, and developing new treatments,
7. Creating minimum standards for advanced AI as a checklist for testing the product and services before their release. Among others, this should ensure that there are no black boxes and that there is full explainability.
8. Treating the discovery of ‘black boxes’ with utmost priority using ever more effective "explainable AI" techniques, to make AI more transparent and understandable to humans,

FMF will need a sizable budget, but this will be offset by immense savings it will deliver by releasing safe products and services.

AI research transparency and open source policy

The question is whether the research on AI should be transparent or kept secret? The main argument for keeping most advanced AI research secret is that if it is disclosed, it may give advantage to the party, which may be in competition with the inventor or may lower the cost of research and give political/military advantage to the party stealing the secrets.

If we were discussing a technological innovation, which may lead to producing millions of products then perhaps keeping it secret might be justified. But we are considering a new type of intelligence, which will be competing with the humankind. Therefore, even if for example, China gets access to the algorithms developed by Google or OpenAI, enabling it to accelerate its own AI program, then we should take this risk because we also want Chinese products to be safe for all humans. More important is that in the end, even China would have to act similarly. Surprisingly, there is quite a lot of AI related research papers released by China. By publishing them, it is thought, they may become a lesser risk than when the advanced FMF were developed clandestinely. In that way they might become ‘a black box’, potentially hostile to every human irrespective where he lives.

That is why the regulatory authorities across the world agree that transparency is better than secrecy when developing AI. Several states are already drafting laws and protocols to manage the use and development of new AI technologies.

In the USA the Algorithmic Accountability Act was introduced in February 2022. The Act aimed to require companies to assess and address the potential biases of their algorithms, particularly those related to protected classes such as race, gender, and religion. It also required companies to provide meaningful explanations for decisions made by automated systems, information about the training data, the code used to train models, and how features like safety filters are implemented. It received support from some lawmakers and advocacy groups, but also faced opposition from industry groups and some technology companies who argued that the bill would impose unnecessary burdens on businesses and stifle innovation. As often in such cases, the market won – the bill was rejected in January 2023.

Whatever, we may think about China’s ambitions in AI, it is clear that they consider the safety of deployed solutions perhaps even more seriously than the Western governments. For example, on 11 April 2023, China’s

cyberspace regulator Cyberspace Administration of China (CAC) unveiled draft measures for managing generative Artificial Intelligence services for public consultations. According to the announcement, services providers will be responsible for the legitimacy of the data used to train generative AI products and services. They should take relevant measures to prevent discrimination when designing algorithms and training data. The regulator also said that service providers must require users to submit their real identities and related information, which given strict control on almost anything in China is not that surprising. Nor are the fines that may be served or the warning that the providers of the services may be suspended, or even criminal investigations launched, if they fail to comply with the rules. CAC will give companies a maximum three months to update their platforms to prevent undesired content to reappear.^[52]

Similar initiatives are taking place in the EU and Canada. The common theme of these proposals is to provide a legalistic basis on what data can, and cannot be used to train AI systems, addressing issues of copyright and licensing, and balancing that against special considerations needed for the use of AI in high-risk settings.

So, is transparency and source code and algorithms sharing a good policy? As mentioned earlier, transparency makes competition much more difficult. And that was probably one of the reasons why OpenAI was less than transparent with its models, which nearly destroyed the company. But more importantly, current lack of a global oversight of the development of the most advanced AI models and sharing of the source code and algorithms may create additional risks if the most advanced AI models fall into the wrong hands. Therefore, open source policy is simply dangerous, since it would proliferate uncontrolled increase of the number of the most advanced AI systems, some of which may be intentionally turned into a malicious AI.

We need to look at the risk and the benefits of transparency in sharing the source code and the algorithms from a global perspective. That requires the acceptance that our civilisation is really at the crossroads where an advanced AI system - AGI or ultimately Superintelligence, will shape our destiny sooner or later. Therefore, instead of ad hoc, partial solutions, such as controlling FMF, there should be one global civilisational shift programme. I know it sounds incredible. However, the sooner we recognize that a civilisational shift has started, the better. Only within such a complex global control, full transparency and source code sharing will make sense as an element of a more effective AI development control.

6. Create Global AI Governance Agency (GAIGA)

This Principle should be completed between 2025-2027

The supervising role of the Global AI Governance Agency

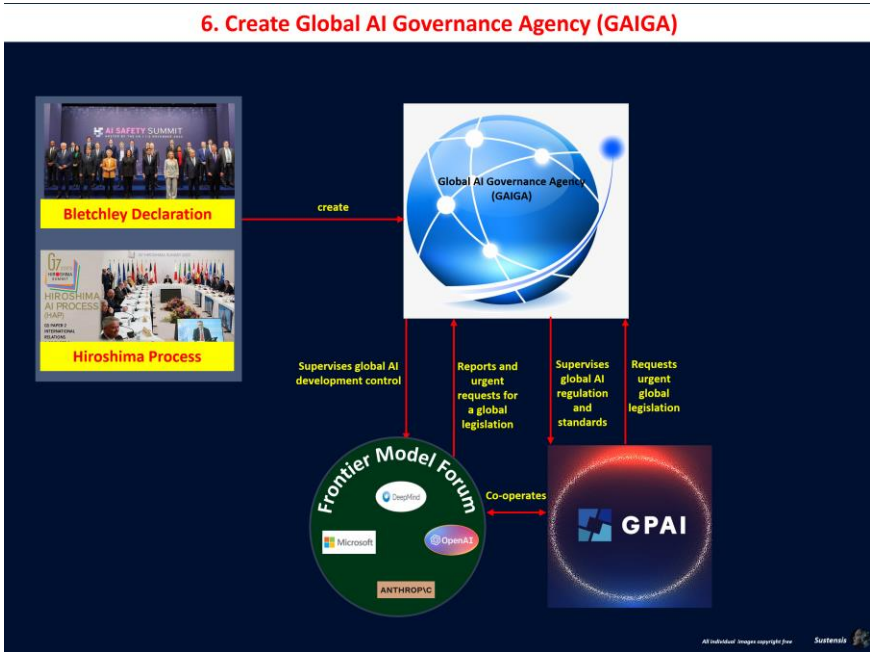
I should remind the readers about an important distinction that I make between **AI regulation** (how AI is used in the society and in business) and **AI development control** (controlling AI's goals and behaviour). We need both AI regulation and AI development control since they play a different role in a civilisational shift and require a different governance approach. If an agency regulating AI is ineffective it will have far less severe consequences for our civilization than the loss of control over AI goals and its behaviour, which in an extreme case may lead to human species' extinction.

Therefore, we need two agencies – one responsible for AI regulation and the other for AI development control. Although they will operate in a different fashion, they will both need a common interface to global legislators. Unfortunately, we do not have such a global legislator yet. In principle, it should be the UN, but its powers are so limited and the speed of implementing such a legislation is so slow that it cannot be considered. We do not have the World Government either.

Therefore, the only practical solution might be to use the next Global AI Safety Forum convened on the basis of the Bletchley Declaration to create **Global AI Governance Agency (GAIGA)**, which would be a superior organization to GPAI, responsible for a global AI regulation, and the Frontier Model Forum (FMF) – responsible for a global AI development control. GAIGA would also have additional functions related to managing a civilisational shift to a Transhumans' world. FMF, responsible for the AI development control, should remain an independent organization, operating as a Consortium, for the reasons given earlier.

Assuming that all members of GPAI would automatically become the members of GAIGA, it would include all of its 29 member-states, which also includes the EU, so together there will be 46 member states. That will cover about 70% of global AI research and development. What will remain outside GAIGA's control will be China, with perhaps 20% of the market, and countries like Russia, North Korea, Iran, and Pakistan.

Therefore, should a de facto World Government be created soon (see chapter 9), the overall AI governance under the supervision of GAIGA might look like in the picture below:



But if GAIGA is created, where should it be based? One of the decisions made at an impromptu summit of President Biden and the British Prime Minister Rishi Sunak was that Britain would organize a Global AI summit in London in the autumn of 2023. The key item on the agenda was the creation of a new global AI organization. That new organization should fill in the existing gap in global AI governance – lack of global AI control development.

The British government would like that organization to be based in London. After all, Britain was until the emergence of ChatGPT the place where the most advanced AI research took place at the London-based DeepMind, part of Google. Britain has its four universities Oxford, Cambridge, Imperial College and UCL among the best 10 universities in the world. It is the second after the USA powerbase in medicine, chemistry, and particle physics. Finally, it is geographically well positioned between the USA, where 2/3 of all AI companies are based and where Frontier Model Forum (GMF) is based, and Paris - where Global Partnership on AI already operates.

As soon as possible, GAIGA should initiate the process of aligning AI's values, goals, and preferences with the Universal Values of Humanity. These should be delivered in the most efficient way, enabling its ratification by all GAIGA members quickly (within a year). The current European Charter of Human Rights (ECHR) and the UN Declaration of Human Rights should be the main input documents. The approved document would form the basis for GAIGA's AI-Mind alignment. I have described that process in detail in my earlier book 'Democracy for a Human Federation' [24].

Prerogatives of GAIGA

Before setting up of GAIGA, any legislative proposals in controlling the development of AI, will be made by FMF - an independent international Consortium. In that period, to act effectively, it will need a strong support from the US government, until such time, when there is a de facto World Government. In that interim period, one of the most important roles that FMF will have, is to identify the needs for fast, and sometimes deep reform of legislation. As soon as GAIGA has been formed, it will take over that function from FMF.

Some of the legislative reforms proposed by GAIGA may require constitutional changes. Here, I am advancing some of the new laws that may be needed to be implemented fast, to give you an idea of how radical these changes in national and individual rights might be. These changes only indirectly relate to the emergence of the Artificial General Intelligence (AGI) quite likely with the next several years. But it should be independent specialists, members of FMF, who should propose such changes in law because of imminent dangers, some of which may emerge within months. They will simply be most qualified to assess the risk and see it in the overall context of the emerging AGI. In a sense, they will substitute Parliamentary (or Congressional) Committees in proposing new laws. However, they will be free from any political dependencies and thus able to say what is needed and how soon the law should be implemented. Should that happen before the setting up of GAIGA, then the US President could issue an Executive Order, which the Congress will have to pass (a rubber-stamping exercise).

To summarize, we need to proceed vigorously with legislative regulation of AI development and its use for two reasons. First, regulating the use of AI products and services, and controlling the AI's development may have some impact on delaying the loss of total control over AI. Secondly, such legislation may cushion the Global Disorder in this decade.

7. Create Global AI Company (GAICOM)

This Principle should be completed between 2025-2026

Learning from China's Long-term AI Strategy Plan

The main objective of this whole approach is to ensure effective control over AI. For that to be achieved there must be one global AI development centre and one AI development programme. As mentioned earlier, such a programme cannot be truly global since China is unlikely to join it. China wants to become a global leader in AI, in which President Xi Jinping has also a personal interest, and use it as a springboard to a global dominance, the subject I introduced in Chapter 2, Part 2. China's concentrated efforts in developing AI is another argument for the West to follow broadly that path. Therefore, I will now present excerpts from the Chinese AI strategy in the context of my proposal in this area.

In 2017, China launched its New Generation Artificial Intelligence Development Plan. That was followed by 'Made In China 2025' (MIC 2025) strategy, inspired by Germany's "Industry 4.0" (I40) programme, an initiative, which strives to secure China's position as a global powerhouse in high-tech industries. The aim is to reduce China's reliance on foreign technology imports by significant investments in its own innovations. This would enable the Chinese companies compete both domestically and globally in the most advanced technologies, including AI. China sees MIC 2025 as a chance to fully integrate its economy into a global manufacturing chain and more effectively cooperate with industrialized economies [53]

The US perceives China's strategy as a direct threat to its dominance in technology, and in particular in AI. At the Davos Forum in January 2023, the FBI Chief Christopher Wray expressed concerns about China's Artificial Intelligence programme. He considers it as "not constrained by the rule of law. That's something we're deeply concerned about, and I think everyone here should be deeply concerned about". That was a reference to the report published in September 2022 by the US Special Competitive Studies Project (SCSP), which warned that the US may lose out to China in the new global technology competition if it does not act on three fronts – microelectronics, 5G, and AI^[54].

The value of the ten largest AI companies in China was about \$40Bn. They are Horizon Robotics, WeRide, 4Paradigm, MiningLamp, Dreame, DJI,

Ubtech Robotics and SenseTime. However, a lot of AI businesses in China are parts of vary large conglomerates such as Baidu or Alibaba. It is quite likely, although that is only my supposition, that following the strategy of New Generation Artificial Intelligence Development Plan, China has already concentrated all its efforts in AI under one command as a kind of a super Joint Venture. This is of course much easier to be done in China, where each private company is under political command and will ultimately have to do what the Chinese government will tell it. This is also what I propose, although in a somewhat different configuration. There are three arguments supporting the creation of one super-large Joint Venture AI Company responsible for the delivery of AGI and later on Superintelligence:

1. It would be impossible to control many advanced AI developments, each of them having different business objectives, different priorities, and different methods of delivering final products and services. They have to be consolidated into one JV Company,
2. AGI, matching human level intelligence, may initially emerge in an individual humanoid. If there is no (or ineffective) global (central) control of such individual AGIs because they will be developed by many independent AI companies, then those AGIs will themselves quickly create a global network, becoming a single entity – an immature Superintelligence. Therefore, FMF must from the start perceive that as soon as AGI emerges it may itself try to quickly develop a networked single entity far more powerful and smarter than any human. Only if there is one JV Company delivering AGI, there may be a chance to control it after AGI has emerged.
3. The third argument for a super concentration of the AI development effort is to counterbalance China's own AI programme, mentioned earlier. It is obvious that the 'Western' AGI will have to match the one built by the Chinese because otherwise China may use it as a tool to control the entire planet. If there are many smaller companies developing AGI, it is unlikely that any of them would have the power and intelligence of the Chinese single AGI system. The only way to match the Chinese effort is to follow their concentrated AI development.

All this makes also obvious that FMF must organize all its efforts around one Superintelligent AI Programme run by One JV company. For those who may doubt the need for this approach, here are some facts which confirm that it may be the best, if not the only, feasible approach to control AI.

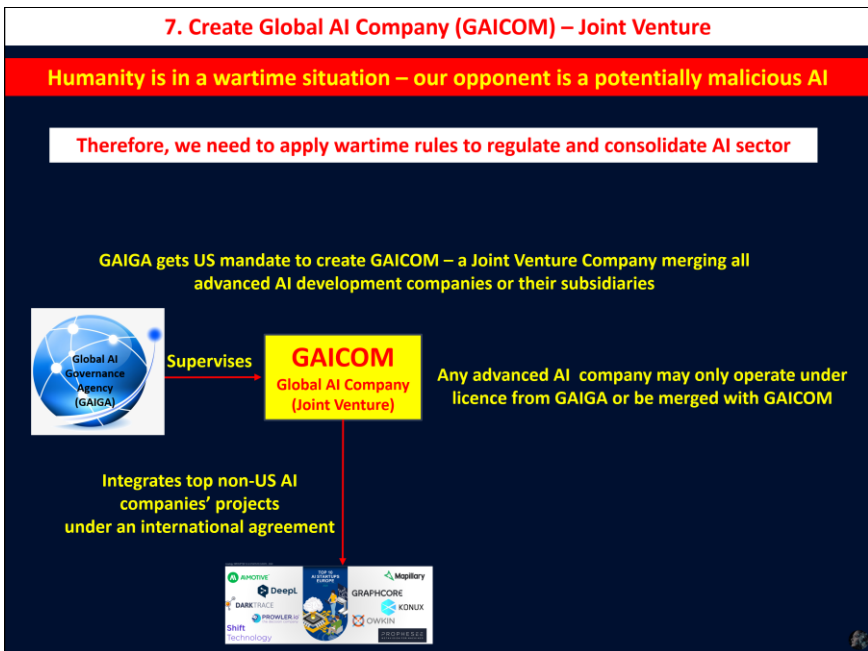
- Alphabet, Google’s parent company, announced at the end of April 2023 that it is **merging all its AI activities, such as Google Brain into one company Deep Mind,**
- Christoph Schumann, founder of German non-profit company Stable Diffusion, which powers products such as image generating StableDiffusion, calls for creating a supercomputer, funded by governments, and running one AI programme set up in a similar way as CERN for the research of the particle physics. He adds: “the technology should be overseen by a ‘well-curated board of directors’ representing AI professors, open-source researchers, and representatives from the mid-scale business community. This group, he says, should be elected by politicians and scientists, but ‘without years of bureaucracy’.^[55]” He confirms the need to structure such an organization like FMF as it is proposed here, as an independent Consortium but supervised by the government, as any other company,
- The latest trend in developing AGI is to abandon building ever bigger Large Language Machine (LLM) models. Open AI says GTP-4, its largest model, apparently having 1 trillion parameters, is the last one that big. Instead, the new approach is to scale such models down to just a few billion parameters, and run them on a single super-desktop computer, costing about \$40,000, and using a small fraction of electric power needed to run such a model. There are already tens of such small companies, delivering a near ChatGPT functionality. To create a full AGI in a few years’ time, perhaps just a few dozens of such standalone computers will be needed, each of them supporting just a single AGI domain, such as image recognition, and far more powerful than the best current models. Connected together they will form a supercomputer network running AGI.

Apart for the need to match China’s efforts in AI, networking such a super powerful AGI system is necessary to prevent global havoc. That may happen, if standalone AGIs could be accessed by criminals. Therefore, FMF main goal should be to build a safe AGI and later on a friendly Superintelligence. This means building the largest and most advanced AGI system, which could disable if necessary any smaller and inferior near AGI system. In this sense, its goal is similar to CERN – an international project in particle physics. In the CERN case it is the alignment through testing in superfast particle accelerators the current Standard Model of Quantum Mechanics with Einstein’s General Relativity theory describing the laws governing gravity.

At the same time, another international agency would be needed to properly control the AI development. Its role would be similar to the Vienna-based International Atomic Energy Authority (IAEA), since it would need a mandate to ban and impose sanctions (which CERN does not have) on ‘near AGI’ systems, under a similar strict regime as is applied to prevent nuclear weapons proliferation. That role should be played by another international body responsible for an overall AI Governance, described in chapter 8.

Building a Joint Venture One AI Company

Once FMF has been created, then one of its first tasks should be the setting up of **GAICOM** – Global AI Company, a Joint Venture, in which it would have just one golden share to control its major decisions. It could be compared to international organizations, such as CERN or TOKOMAK, but its role would be far greater, since it would have to ensure that the control of human’s future remains in our hands. The reason for such a centralization of development of the most advanced AI capabilities is that it is the only way to deliver an effective control of hundreds of companies developing advanced AI systems.



To achieve that, companies with the most advanced AI business and key technologies necessary to build AGI and later on, Superintelligence, would be requested by FMF to separate that business (technologies, projects, or products) from the rest of the company. That part of their business would be then legally split from the original company and added to GAICOM. The company would be compensated in GAICOM's shares, which will not be publicly traded, to eliminate the impact of the market. FMF may consider global market share, uniqueness, material resources, or the necessity for the overall Programme for the valuation of such assets.

Over the first year, companies will be invited to join the GAICOM company voluntarily. However, later on some companies with advanced AI may be ordered by FMF, acting on an international mandate, to split the relevant part of their AI business and merge it with GAICOM. Some AI companies may also wish to check with FMF to get a legally binding reply, for example before a major investment, whether their AI business would be requested to join GAICOM.

Such procedures may be considered as extreme, undermining fundamental freedoms and company laws in most countries. They may also be seen as stifling innovation and competition, being monopolistic. However, allowing only one company developing advanced AI will be necessary if AI control is to be effective. This is a competing intelligence that will soon be far superior to ours, and which may become human's most dangerous enemy. Each individual advanced AI system may potentially become a weapon. We are already in the first days of a wartime period. As in any war, and even in a peace time period, weapons are always under the control of governments. No company can produce lethal weapons without a government's licence. Therefore, any advanced AI company must work under the government's licence and join GAICOM or stop AI development.

Perhaps the first two companies, which might contribute some of their AI related assets are Google and Microsoft. They should separate Google browser from Google, and BING and Edge browsers from Microsoft. These browsers should then be integrated into one browser and Google and Microsoft form the first Joint Venture company of the future GAICOM.

GAICOM would be responsible for managing all advanced AI projects under one Superintelligence Development Programme. It will set up its objectives and timescales, designing the methods of controlling AI. This will in some way force those that may let the genie out of the bottle to keep it

firmly inside. That is the only chance we have. The AI sector must start a radical change in its own yard. It should re-assure the markets and politicians, that the worst chaos may be avoided if this is carried out in a similar way as in China. That it is not a kind of socialism but the most urgent need to keep AI under control, and at the same time for the democratic world to make strenuous efforts to counterbalance China's efforts in this area.

8. Create Superintelligence Development Programme (SUPROG)

This Principle should be completed between 2025-2027

Why do we need one global AI programme?

The Chinese ‘New Generation Artificial Intelligence Development Plan (2017)’, which I have mentioned earlier, might be a template for the AI programme, which I am outlined in this chapter. It confirms why even a stricter approach to developing AI based on an Open-Source policy is needed, in view of the OpenAI’s current closed shop policy. In contrast to its name, OpenAI broke its commitment to keep all research as an open source, when it released ChatGPT, and instead keeps its algorithms and key research tightly closed to external scrutiny. On the other hand, the key difference between the Chinese programme and the Programme proposed here is that the first one is just for China, whereas the programme proposed here is to be global. At least that is what it should be, if AI control is to be most effective. For that, we would need the World Government with sufficiently strong executive powers. But I have already mentioned it several times that it is of course impossible, given the timescales. So, what is the solution?

First of all, we must accept that most of the decisions and solutions applied may not be perfect. We must also accept the 80/20 rule, or as some people say, working in a ‘quick and dirty’ mode. Although this will unavoidably create some errors, this imperfect approach is the only feasible way forward in retaining the control over AI for as long as possible. Within about a decade, we would need to apply dozens of control mechanisms simultaneously to increase the chances of success. That success would be the release of a maturing Superintelligence only after it has been primed with the Universal Values of Humanity, and when it understands what it means, and perhaps even what it feels, to be a human.

The imperfect solutions relate primarily to a partially global control. Some countries will simply not accept quite drastic measures that may need to be imposed, like a significant loss of their sovereignty. That will reduce the effectiveness of the controlling measures. We can only minimize the risk of creating a malevolent Superintelligence by applying various methods of control simultaneously and extending the control period far beyond of what might otherwise be possible if AI is controlled only for a very short period.

Whatever we will do, this decade may decide if the homo sapiens species becomes extinct or gradually evolves into a new, inorganic species. It is still up to us to determine to some degree the most likely outcome. If we create a benevolent Superintelligence, we would like it to inherit the best human traits, so that initially we would evolve with it, and later on, within it.

Consolidating AI development into One Superintelligence Programme

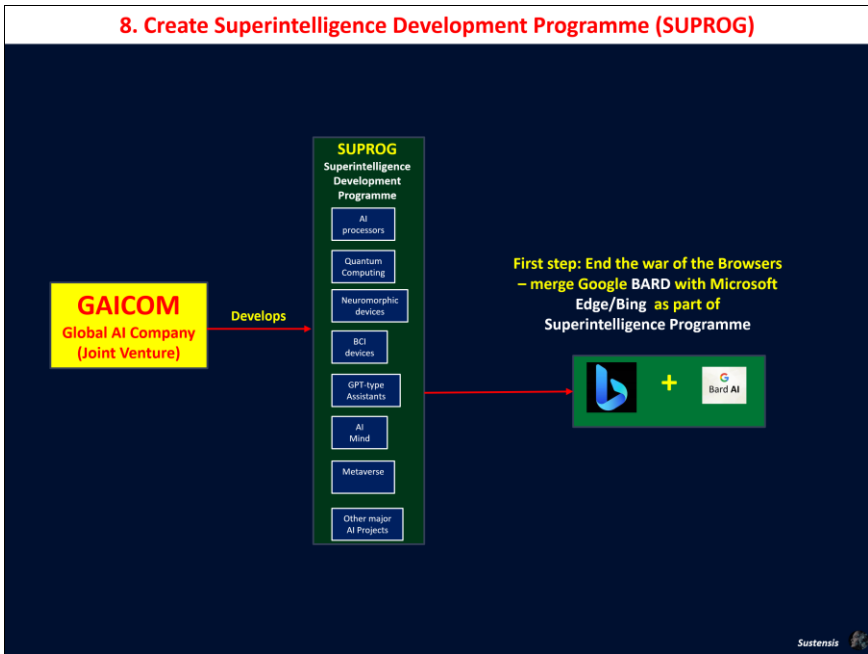
The only way, in which we may effectively control the AI research and development is to consolidate all advanced AI research and already deployed projects into one large programme **Superintelligence Development Programme (SUPROG)**. I need to remind you here about the difference between Artificial General Intelligence (AGI) and Superintelligence. AGI is a self-learning intelligence capable of solving any task better than any human. But Superintelligence is a single networked, self-organizing entity, with its own mind and goals exceeding all human intelligence. The first breakthrough will happen when AI reaches human level intelligence and becomes AGI. This can even be an individual AI Assistant.

By the end of this decade, we may have thousands of them, each as intelligent as anyone of us. However, in a few months, those humanoids may self-connect to each other rapidly creating a global network, unless we are able to restrict them. This will be an Immature Superintelligence, which we may be unable to control because it will outsmart us whatever we do. Therefore, we must do everything possible that once AGI or an Immature Superintelligence is outside of our control, its goals and behaviour are aligned with the Universal Values of Humanity and follow our preferences rather than its own. However, the only way, when such an Immature Superintelligence may still remain under human control is to develop it as One Super Programme, hence SUPROG. That would ensure the future Superintelligence becomes our partner rather than an evil entity, which may by error, fighting for common resources, or malicious behaviour cause human extinction.

The current situation shows what may happen if there are hundreds if not thousands of individual advanced AI projects developed by different companies. ChatGPT was released on 30th November 2022. But in May 2023 there were at least 10 AI Assistants of similar competence. One of them, Claude, developed by Anthropic, has outperformed ChatGPT by being at least 10 times more powerful, if measured by the contextual information it can process. ChatGPT can only process an input (Prompt) of about 2,000

words, or 3-4 pages. Claude can process about 77,000 words, which is the size of this book. It also teaches AI in a much simpler, less expensive, and more effective AI learning process^[56].

An effective AI control must from the very start focus on the AI's goals and behaviour, including knowing how it has arrived at any decision or solution, so called explainability. It must be the ultimate decision centre, similar to the BIOS programme, which controls every computer operating system, and which may be achieved via the MASTER PLATE as proposed in this book. This is where the Universal Values of Humanity will be stored as well as its goals, and human preferences, continuously updated as the maturing AGI experiences the world of humans. This should be at the top of SUPROG hierarchical structure, which will consist of hundreds of projects, research labs and even manufacturing facilities.



The delivery of individual projects would be the responsibility of GAICOM. Here is a list of companies aligned with their AI expertise (in reality companies and their allocated functions may be different):

- Amazon – main distributor
- IBM – Quantum Computing and super large computers

- NVIDIA – AI and graphics processors
- Intel – neuromorphic processors
- META - Metaverse
- Microsoft/Google/Apple – a brand new ‘AI Operating System’
- Google – Development of AGI and Superintelligence. It might be the Programme ACCELERATOR
- Deep Mind & OpenAI – AI control, AI Antivirus, eliminating ‘black boxes’ and ensuring the AI Mind’s explainability. It might be the Programme’s – BREAKING PEDAL
- Neuralink/TESLA – Robotics household and neurosurgery
- Boston Dynamics – Robotics industrial (Volvo ABB + others)

This is just an approximation of what potential major projects might be, which companies might deliver it, and in what they may specialize. There will probably be a few hundred narrowly specialized companies, members of GAICOM. Each of them may be working on dozens of projects. All of these companies’ projects and deliverables will have to be integrated within SUPROG. This is a truly mammoth task and perhaps the largest programme ever created by humans. It will be far more complex and important than the NASA’s Moon Landing Programme or the Manhattan project.

I assume that initially only the US companies and projects would join GAICOM and SUPROG mainly for legalistic reasons. They will anyway constitute more than half of all projects and companies in the world. However, non-US companies could join at any time, perhaps with the assistance of Global AI Governance Agency (GAIGA) – see next chapter. When GAIGA takes over the supervision of FMF from the US government, then whatever legal arrangements have been established by then for the US companies, will have to be made compatible with the international law.

AI Maturing Framework – a multi-modal control of AI by GAIGA

Priming AI with Universal Values of Humanity

There are quite a few proposals on how to control AI and minimize the risk of misinterpretation of the acquired values by Superintelligence. Nick Bostrom mentions them in his book “Superintelligence: Paths, Dangers, Strategies”, especially in the chapter on ‘Acquiring Values’. The techniques specified by him aim to ensure a true representation of what we want. They

are very helpful indeed, but as Bostrom himself acknowledges, **it does not resolve the problem of how we ourselves interpret those values.** Therefore, another additional method of controlling AI is needed.

One of the solutions proposed in this book is the setting up of AI Maturing Framework. This is an integrated multi-modal framework, which GAIGA may apply for a comprehensive control of AI. It consist of four modes, hence multi-modal. The key element in all four stages is the learning of human values and preferences. Please note, that I have used the word ‘preferred rather than ‘best’ since this would always leave a margin of uncertainty into the Superintelligence’s actions. This, as prof. Stuart Russell suggests in his book ‘Human Compatible’^[57], might significantly reduce the risk of making wrong decisions by any AI agent.

The AI Maturing Framework is built with a certain End Goal, which is defined as:

Teach AI preferred human values till it matures into Superintelligence

Just to remind you, Superintelligence is a single networked, self-organizing entity, with its own mind and goals exceeding all human intelligence. It will follow the development of Artificial General Intelligence (AGI), which I expect to emerge by 2030. However, to successfully control AI, we must take a long perspective and see the time when AI will be completely outside of human control. Therefore, GAIGA should consider all available means for such a control to be effective after the arrival of AGI.

One of the key objectives of AI development control is to prime it with the Universal Values of Humanity. For that process to be effective, those values will have to be expressed in such a way that they have a unique, unambiguous meaning. That is the well-known issue of “Do as I say”, since quite often it is not exactly what we really mean. Humans communicate not just by using words but also by using symbols, and quite often additionally re-enforce the meaning of the message with the body language, to avoid any misinterpretation, when double meaning of the words is likely. Would it then be possible to communicate with Superintelligence using body language in both directions? You may have come across this problem when writing emails. To avoid misinterpretation, we also use emoticons.

How would we then minimize misunderstanding our preferences further? One possibility would be, as John Rawls, writes in his book “A Theory of Justice” to create algorithms, which would include statements like this:

- do what we would have told you to do if we knew everything you knew,
- do what we would have told you to do if we thought as fast as you did and could consider many more possible lines of moral argument,
- do what we would tell you to do if we had your ability to reflect on and modify ourselves.

When Superintelligence emerges, we may also envisage a situation, where it is “consulted”, on which values to adopt and why. There could be two situations (if humans still have an ultimate control).

In the first one, Superintelligence would work closely with Humanity to re-define those values, while being still under a total control of humans.

In the second one, which I am afraid is more likely, a benevolent Superintelligence, even with no ulterior motives, may see that our thinking is constrained, or far inferior to what it knows, and how it sees, what is ‘good’ for humans. Therefore, Superintelligence could over-rule humans anyway, for ‘our own benefit’, like a parent, who sees that what a child wants is not good for it in the longer term. The child being less experienced and less intelligent simply cannot comprehend all the consequences of its desires.

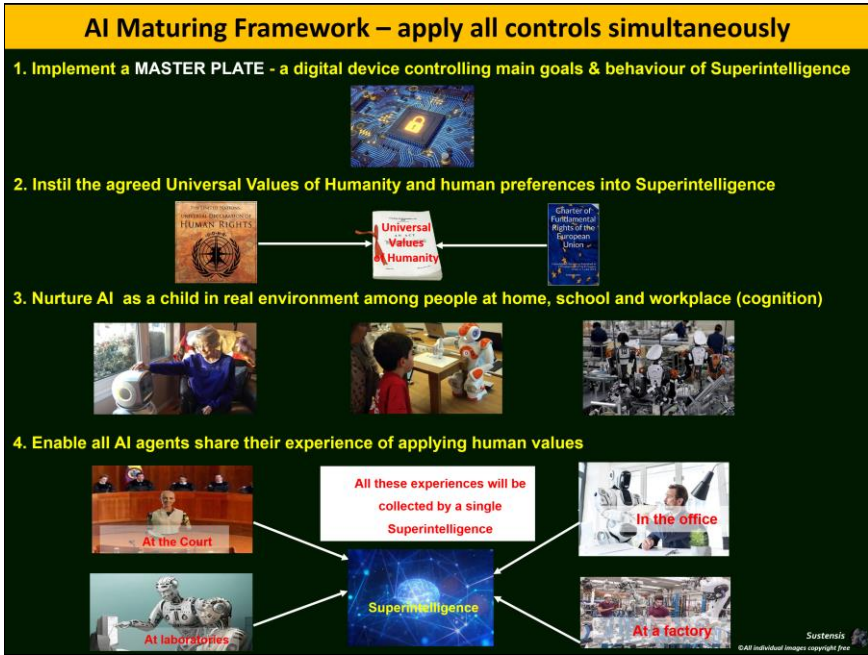
On the other hand, Superintelligence would need to consider the values, which are strongly correlated with our feelings and emotions such as love or sorrow. In the end, emotions make us predominantly human, and they are quite often dictating us the solutions, which are utterly irrational. What would be the choice of Superintelligence if its decisions are based on rational arguments only? What would happen if Superintelligence does include in its decision-making process, emotional aspects of human activity, which make us more human but less efficient and from the evolutionary perspective, more vulnerable and less adaptable?

The way Superintelligence behaves and how it treats us will largely depend on whether at the Singularity point it will have at least basic consciousness. My own feeling is that **if a digital consciousness is at all possible, it may arrive before the Singularity event.** In such case, one of the mitigating

solutions might be, assuming all the time that Superintelligence will act from the very beginning benevolently on behalf of Humanity, that decisions it proposes would include an element of uncertainty, by considering some emotional and value-related aspects.

Irrespective of the approach we take, **AI should not be driven just by goals (apart for the lowest level robots) but by human preferences**, keeping the AI agent always slightly uncertain about a goal of a controlling human. It is the subject for a long debate about how such an AI's behaviour can be controlled, and how it would impact the working and goals of those AI agents, if this is hard-coded, as I propose, into a controlling Master Plate chip. Therefore, the issue of ethics, emotions, and uncertainty in such a controlling chip, must be resolved very quickly. They will be the key objectives of FMF's AI development control. The most effective controlling method might be a simultaneous, multimodal framework such as perhaps the proposed AI Maturing Framework, which includes four stages:

1. **Implement a MASTER PLATE**, a digital device, controlling the main goals and behaviour of Superintelligence, humanoids, and more advanced AI Assistants,
2. **Instil the agreed set of Universal Values of Humanity** into Superintelligence to ensure the alignment of humans' preferred way of living and interacting with other humans, cognitive humanoids, and AI Assistants,
3. **Nurture AI as a child** in real environment among people at home, school, and workplace (cognition) so that it learns human values and preferences based on their interaction with humans in any environment and circumstances,
4. **Share the experience of humanoids** and AI Assistants with a maturing Superintelligence by applying human values and preferences.



1. Controlling advanced AI via the Master Plate

The Master Plate is a digital device, an integrated circuit, which will be embedded into the Superintelligence’s computer system. It is the top layer of controlling all its goals and behaviour. It can be compared to a computer’s Basic Input/Output System (BIOS), which was first introduced in Personal Computers with DOS operating system. Today, no computer or a mobile phone would work without an equivalent of BIOS, and neither would Superintelligence without a Master Plate. Some AI systems, running of course on a computer, may already have some sort of a Master Plate equivalent. However, what is proposed here is a comprehensive system of multi-modal AI control, with the Master Plate being its core. It is described in detail in Part 3, chapter 4.

2. Teaching human values to AI directly

The teaching process should start with the uploading of the Universal Values of Humanity, which may by then also include 23 Asilomar principles related to the development of AI or a similar set of AI regulatory system. For AI humanoids the Mini Master Plate licensed by GAIGA would co-define its

goals. It would contain a detailed description of what values, rights and responsibilities really mean, illustrated by many examples.

There is of course no guarantee that the values embedded into the Master Plate, or the Mini Master Plate devices can ever be unambiguously described. That's why humans use common sense and experience when making decisions. But AI agents may not have it yet, and that is one of the big problems. In this decade, we shall see humanoid robots in various roles more frequently. They will become assistants in GP's surgeries, policemen, teachers, household maids, hotel staff etc., where their human form will be fused with the growing intelligence of current Personal Assistants. Releasing them into community may create some risk.

3. Nurture AI as a child

One of the ways to overcome that risk might be to nurture AI as a child. Therefore, GAIGA may decide to create a Learning Hub, a kind of a school, which would teach the most advanced robots and humanoid Assistants on how human values are applied in real life and what it means to be a human. In such a school, the robots will interact with people in various areas of human activity, such as school, factory, office, cinema, shop, museum, etc. They will then communicate back their unusual experience of applying values in the real environment back to the Agency, where such experience will be combined with the experience of millions of other AI assistants.

This will mimic the process, which is already being applied to training Large Language Models (LLM) like GPT-4. It includes backpropagation, where AI looks back on what it has said or achieved. It then repeats the process of self-learning until it responds in the required way.

Once AI agents have 'graduated' from such a school they will be ready to serve in the community. Additionally, their accumulated knowledge, stored in a central repository on the network, a kind of early 'pool of intelligence', will have a gateway, through which each of these AI agents, with proper access rights, will be able to update itself, or be updated, to gain up to date guidance on a proper behaviour and the way to react to humans' requests.

4. Enable all AI agents share their experience of applying human values

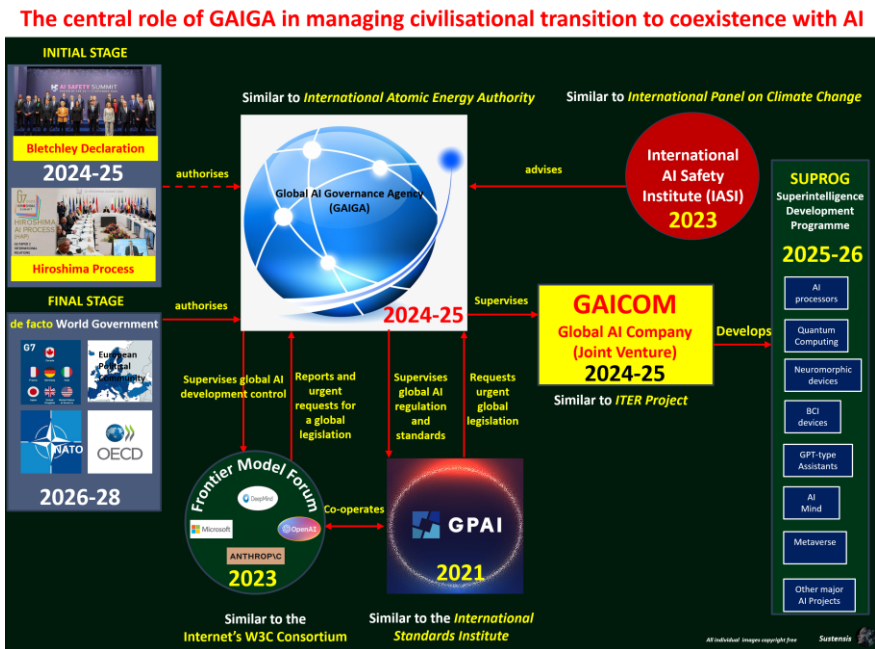
Finally, AI agents, humanoids, robots, or autonomous cars will learn human values, and especially preferences in choices and behaviour, by **directly** sharing their own experience with other AI Agents. In the end, this is what

some companies already do. Tesla cars are the best example, by pooling each car's unusual or dangerous experience in a central repository, whose content is then updated to all vehicles. Google's Waymo has a similar, but of course a separate centre. You yourself use such a system already when driving your car with the help of Google's navigation system. The cars in front of you immediately inform the Centre of the traffic situation on a given road and, in a few seconds, it is relayed to the cars further down that road. Right now, these Centres, which are storing values and behaviour from various AI agents, are dispersed, each centre supporting an individual AI Company's 'micro' Superintelligence. For the purpose of a maturing Superintelligence, there should be just one such global centre.

That is one more reason why there is an urgent need for GAIGA to develop a single, rather than competing versions of AGI. The Agency may consider to progressively make its Centre for storing values, behaviour and experiences of millions of robots and other AI agents, as the controlling hub of the future, single Superintelligence.

Summary of Global AI Governance

An effective Global AI development control may not be possible without creating a network of key organizations, which would be responsible for a safe development of just one, global, most advanced AI system, which will first emerge as AGI and later on as Superintelligence. The interdependency of those organizations is depicted in the diagram below:



GAIGA will play a central role in managing civilisational transition to coexistence with Superintelligence. Several pillars of this setup already exists. These are: GPAI, formed in 2021, Frontier Model Forum, formed in 2023, International AI Safety Institute (2023, if the UK's Institute is formally approved for its international role), and Hiroshima Process with Bletchley Declaration as an interim international authority on delivering safe AI. The creation of GAIGA is critical for delivering a robust global system of AI development control and later on coexistence with Superintelligence. Speed is of the essence and therefore the creation of GAIGA and its supervision by an international body may happen in less formal way, where it will be authorized by the AI Global Summits within the Bletchley Declaration and/or Hiroshima Protocol, and later on by a de facto World Government.

The dates seem almost unrealistic but we should realize that if AGI emerges earlier than GAIGA, the world may face a malicious AI, which may very quickly be out of human control. The same applies to the creation of a de facto World Government. It does not even have to have such a name. It is enough that its decisions will have an impact like made by a real World Government. We already have an initial composition of such a de facto World Government. These would be most likely the countries, which attended the Bletchley Park Global AI Safety Summit, perhaps without the presence of China. Therefore, this group of states may actually work like a de facto World Government in the next 2-3 years. Here is the summary of the roles of the organizations needed to manage a smooth civilisational transition:

1. **International AI Safety Institute (IAISI)** to minimise the unexpected advances in the frontier AI models by developing dedicated monitoring and testing methods. It should operate in a similar way as the *International Panel on Climate Change (IPCC)*. While there is no scientific proof that AGI will emerge by 2030, just as there is no proof of the Global Warming reaching a tipping point by that time, we must develop AI as if AGI were to emerge within that timeframe.
2. **Global Partnership on AI (GPAI)** to be responsible for AI **regulation and standards**, leaving AI development control to a new Agency. It should also set global standards for specific AI hardware and operate like *International Standards Institute (ISI)*.
3. **Frontier Model Forum** to be responsible for a **global development control** of the most advanced AI model by expanding its US base to include companies from other countries. It should operate like the Internet's *W3C Consortium*.
4. **Global AI Governance Agency (GAIGA)** under the mandate from the Bletchley Declaration and the Hiroshima Process. It should have the prerogatives similar to the *International Atomic Energy Authority (IAEA)* in Vienna. GAIGA would oversee both GPAI, responsible for regulating the use of AI products and services, and the FMF Consortium, responsible for AI development control.
5. **Global AI Company (GAICOM)**. This could be a Joint Venture company to consolidate the most advanced AI companies into a single organization. It would be similar in its objective to the *ITER project* funded by the US, China, Russia, the EU, Japan, India, and Korea, to develop the first nuclear fusion reactor. Effective control over AI

development will be impossible if it remains dispersed among numerous companies.

6. **Superintelligence Development Programme (SUPROG)** managed by GAICOM. This would be similar in its objectives to the *NASA's Apollo Programme*.

9. Create a de facto World Government

This Principle should be completed between 2027-2030

The need for the World Government

Perhaps one of the reasons that we do not have the World Government yet is that many of us still hope for the UN to be transformed into such an organisation. That was a noble objective of the World Federalist Movement (WFM), created just after the UN had been founded, and affiliated to that organisation. Just imagine what the world might have looked today if WFM objective had been to create the World Government based on the ‘coalition of the willing’, which would have included many countries but not all. That would have been a de facto World Government.

The best example of how it could have been done is the ratification process of International Bill of Human Rights. In 1966, the United Nations adopted two legally binding international treaties that were inspired by the Universal Declaration of Human Rights. These treaties are collectively known as the International Bill of Human Rights. To be binding, it required the ratification by at least 35 countries. It took 10 years, but the International Bill of Human Rights has become a cornerstone of more humane justice systems worldwide although so much more has still to be done.

More and more people including some politicians, like the French president Macron, begin to recognize global problems, such as climate change, as potential existential threats for Humanity. Existential threats can materialize at any time, e.g., the Coronavirus in 2020, or due to combinatorial effects of several risks such as large-scale migration, draught, local nuclear wars or cyberattacks. Therefore, they see the urgent need for countries to limit their sovereignty and to federate as a planetary civilisation.

However, there is no hope that all countries of the world would give up significant part of their sovereignty in the near future to become part of the World Government. Therefore, the only practical way forward is to follow the example of the International Bill of Human Rights and create an organization that could act as a de facto World Government right now. That might start with the federation of the European Union in some form and then extending that process worldwide. But there can also be other options, which I present further on in this chapter.

Criteria for selecting organizations for the World Government

We must be realistic and recognize that there is no time to create the World Government from scratch, with all countries as its members. We can only transform an existing organisation, or empower a single large country, to become a **de facto** World Government, with the powers of a federation. But who could play such a role? That is covered in my book ‘Democracy for a Human Federation’ [24]. Here is just a summary

To select a candidate organization to be converted into a de facto World Government, I have specified in the table below the scope and prerogatives needed for such an organisation to be successful in mitigating existential risks, ignoring for now its other objectives.

Weight	Justification for the selection criteria for the World governing organization	
10	Democratic institutions	This is the most important criteria because if we want to assure that we do not make things worse than they are now, then the nations that will surrender good part of their sovereignty must be assured that they will be governed within the best democratic system humanity has ever created
9	Respect for Human values	This is the second criteria in importance for two reasons. The organisation must be exemplary in its respect of human values and it has to carry out the process of redefining them for the upload to Superintelligence to make its risk as low for Humanity as possible
8	Military power	Any organization that will carry out such a role must be one of the most powerful in the world to withstand the threats from countries that will not be its member and carry out missions to minimize the risk to humans, such as Weaponized AI, or wars that could become global, or are of genocide type
7	Economic power	This is important because the organization must have enough resources to mitigate existential risks
6	Organizational capability	Essential when carrying out missions to eliminate threats from existential risks, such as nanotechnology
5	Response time to risk	The selected organization must be capable of very fast response to risk, sometimes within hours, i.e. nuclear war threat or artificial pandemics.
4	Land mass	This is important to have available resources as well as creating spaces that may not be contaminated, e.g. biochemical risks
3	Experience in large programmes	Essential when carrying out missions to reduce existential risks, such as global socio-political risks
2	Versatility	The organisation which is to mitigate all kinds of risks endangering humanity must be very versatile and not for example have experience in the military field only
1	Neutrality, Objectivism	This is again important to assure cohesion of the organisation that will have powers to reduce freedom or sovereignty

The table has 10 selection criteria for 10 organizations, or large countries. I have tried to make the selection as objective as possible. 3 of the 10 criteria that I have used are completely objective: military power, territory size and the annual GDP. The remaining 7 criteria are subjective, but that subjectivity is within a narrow margin and within the 10 criteria does not make a big difference. I have also assigned the weights, considering the importance of a given criteria in performing the role of the World Government. Bearing in mind how difficult it would be to convince the prospective countries to give up a significant part of their sovereignty, I assigned the top two weights for democratic institutions and respects for human values. Then comes the military power because this is the essence of any government if it wants to enforce its will on important matters.

The table below summarises the result of the analysis of potential candidate countries and organizations, which might be converted into a de facto World Government.

Name of Organization or State	Risk Mitigation Capability Ranking (weighted)										Total Score (weight * capability)
	Democratic Institutions	Respect for Human values	Military power	Economic power	Organizational capability	Response time to risk	Land mass	Experience in large programmes	Versatility	Neutrality, Objectivism	
Weight ---->	10	9	8	7	6	5	4	3	2	1	550
G7	9	9	7	9	10	10	8	10	10	10	492
NATO	9	8	10	10	9	10	9	5	4	10	485
European Union	10	10	6	8	10	9	6	10	10	10	483
USA	8	8	9	8	10	9	7	10	10	10	473
Japan	10	10	2	6	9	9	1	6	6	9	390
Canada	10	10	4	3	8	9	4	5	6	10	389
Australia	10	10	4	3	8	9	3	5	6	10	385
United Nations	10	10	2	2	6	4	10	6	10	10	364
China	2	2	8	7	9	10	5	10	10	1	326
India	6	4	5	5	5	4	2	5	3	5	255

Who might play the role of a de facto World Government?

In 2018, it was the federated European Union, which appeared to be the best candidate to become the World Government. But the pandemic and the Ukrainian war put to the fore the organization, which has not been even considered at that time. It is the G7, a rather informal organization, whose member countries include Canada, France, Germany, Italy, Japan, United Kingdom, and the United States. Altogether they account for 30.7% of the world's GDP.

There may be yet another, faster and even shallower type of federation, which may become a de facto global decision maker, rather than a de facto World Government. It may be the creation of a defensive alliance based on NATO and the EU military capabilities. Such an option becomes more realistic day by day and may become reality within months rather than years if the war in Ukraine expands into Moldova or the Baltic countries. Considering that NATO has similar insistence on adhering to common democratic principles as the EU (the only outstanding problem is Turkey), such a defence alliance would in fact become the most powerful political organisation.



If we had not been facing about ten existential risks, three of which: AI, Global Warming, and Global Disorder, all having their tipping point about 2030, then the governments might have been negotiating over this whole century to form the World Government, which would include all countries. But we may just have about a decade to make profound changes in the way the world is governed to prepare ourselves for soon to start coexistence with Superintelligence, as the first step on our way to a human species' evolution.

We must consider that all those mega reforms in global politics may be taking place at the time when quite likely the world will be in the most significant chaos ever, even including the period of the WWII. One of the reasons will be fast acceleration in proliferation of AI services and products, especially humanoid robots, which will be mass-produced in millions. This will deliver many benefits but also create unprecedented turbulence in the world' economy with the Technological Unemployment making a severe impact on people's wellbeing.

Such a scenario of creating a de facto World Government does not imply an immediate dissolution of the UN. Just to the contrary. The UN may still play an important role, in the areas, where political unanimity may not be of utmost importance. Fighting Global Warming is a good example.

Therefore, such a de facto World Government may co-exist with the UN for some time, like in the current situation, where the UN is unable to end the war in the Ukraine and therefore, a coalition of the willing western countries (NATO and the EU) fulfil this role. In any case, if such an organisation emerges, it should by default co-operate as much as it would practically be possible with the UN.

Whichever option materializes to form a de facto World Government, such an organisation should be quite quickly converted into, what I would call, a Human Federation. It would then fulfil the role of the United Nations, which the UN has been incapable to play, with some key differences, such as:

- Majority or double majority voting, like it is being applied more frequently now in the EU. It means that some critical decisions or legislation can only be passed, if it is supported by the majority of the countries and the majority of the citizens of the member states,
- Some executive powers enabling it to enforce its decisions with a military force, if necessary,
- Only admitting the countries, which fulfil democratic criteria adjudicated by an independent constitutional court. Therefore, it would not include all, but hopefully the majority, of the states,
- It would have its own army, probably based on NATO.

Forming a de facto World Government, would only be a beginning of even bigger reforms that will be needed to make a civilizational shift. If what you have just read materializes, it would be a swan song of this civilisation, setting foundations for an entirely new one.

GAIGA's role as 'the Ministry of War' of de facto World Government

The only way to achieve a tighter and effective AI control is to have all major AI research and development control under one roof. That should be the democratic world's overall strategy. From that follows the key role of this agency, which is to ensure that a global AI system remains under human control for as long as possible. We have to visualise the role of this Agency as similar to a Manhattan project, since the situation our civilisation is in, resembles that period when the whole world was at war. Perhaps even a better comparison would be with Britain in September 1939. Britain was then at war with Germany but that was still an invisible war. Nevertheless, it was obvious then for everybody that it was only a question of months when the real war will be thought over Britain. Therefore, mobilisation was in full

swing, severe freedom restrictions were imposed, companies, citizens and other organizations had to be totally subservient to the state as that was a matter of survival.

Today, we are already in a wartime period similar to 1939 in Britain, with AI being an invisible enemy. But the stakes are much higher, when we have to decide whether we accept that the only way for the survival of a human species beyond this century is starting a human evolution to become a new species. The problem is that we need to start it right now because in less than a decade we may no longer have that choice.

Therefore, GAIGA should have a very broad mandate, initially from the US government and later of from a de facto World Government. Like the Ministry of War, it would request the necessary legislative changes, provision of the required resources and staff, as well as direct input on the government's priorities in the areas, which may directly impact the overall effectiveness of AI control.

It will have several tools to achieve that like, licensing Brain-Computer-Interfaces (BCI) for Transhuman Governors and other use, mandatory companies' mergers, demerges, joint ventures, companies' nationalisation, etc. It will become a kind of a civilisational transition agency.

GAIGA will also closely co-operate with the regulatory agency GPAI in the following areas:

- It will have the power to stop any advanced AI program in any jurisdiction under its control,
- It will oversee non-competition policy for all companies which are part of GAICOM,
- It will be the final decision maker on the release of AI products and services above a certain level of AI advancement, as requested by FMF,
- It will issue special licences for more sophisticated AI products (regular licences would still be issued by GPAI).

I realize how far today's political reality is from what will have be done in the near future. However, we have no other choice. Either it is done on time, or we may quite quickly find ourselves in the world controlled by AGI. GAIGA may also be directly involve in proposing laws as requested by the

World Government. Among the laws that may have to be modified (the extent of these modifications depends on the actual needs) are the following:

1. **Limiting the right to an unlimited wealth.** That is a fundamental change in individual's rights to property, which has been sacrosanct in every democracy. But it cannot continue any longer,
2. **Limiting the right to unlimited corporate assets,**
3. **Forcing splitting part of a company with an advanced AI business** to join GAICOM. This can already be done under the existing law in most countries. It is more about fast execution,
4. **Restrictions on personal freedom** for example on the use of AI in the way, which may harm people,
5. **Limitation of national sovereignty.** This should not be directly imposed but rather be an offer for a fast membership of the World Government, with significant privileges, primarily enhancing the national security, and gradual elimination of all wars,

These are of course just broad ideas pointing to the direction of travel rather than being a detailed roadmap. The money thus collected both from the richest individuals and from the companies, should be largely funnelled to the Global Wealth Redistribution Fund – see Chapter 10 in this Part.

I have only sketched out what GAIGA's responsibilities might be. The wide range of its responsibilities will quite quickly make it an organisation similar to a Technocratic Government. However, its emergence would happen for a different reason. Usually, technocratic governments are created when none of the parties can create a governing coalition, so they elect technocrats, mostly specialists or scientists. We had the most recent example of such a technocratic government in Italy under the premiership of Mario Draghi.

However, GAIGA might become a technocratic government for a different reason. The politicians would simply not be able even to understand some of the reasons behind the necessary decisions put forward by AGI and soon after, by Superintelligence. If GAIGA is set up on the principles similar to those proposed here, then by about 2030 it may indeed play the role of the world's technocratic government. It is quite likely that at that time most GAIGA's decision makers will be Transhuman Governors.

10. Create a Global Welfare State

This Principle should be completed between 2030-2032

Redistributing wealth more evenly

This is the last of the 10 Principles. Its fulfilment depends entirely on the successful completion of the previous steps. Nothing in this stage would be possible without successful control of the AI development. Towards the end of this decade AI may already be at AGI level. That will itself open entirely new opportunities for creating the world of plenty. But to create a Global Welfare State there would have to be a global wealth redistribution, which would be very difficult without the World Government.

Notwithstanding that, we must start building the Welfare State earlier e.g., before 2030 because even modest improvements in the wellbeing of billions of people, mainly in the Southern Hemisphere, will dampen potential global chaos arising from huge migration waves or local wars. To do that we must begin global wealth redistribution on the scale never attempted before. Even from a medium-term perspective it will be the least expensive way to maintain some sort of global peace.

Realistically, it will probably only happen if the world accepts it as unavoidable when facing some potential catastrophes, such as mass migration. The world's peace, and in the end, the survival of the human species, is only possible when we change the view of our future from a national to a planetary perspective. This includes global economic sustainability based on significant redistribution of wealth.

I am fully aware of the complexities and almost impossibility of delivering such a momentous change for humanity in the world which, for example, could not agree to stop the genocide in Syria. The odds are heavily against such a scenario as I am presenting here. On the other hand, should we be incapable of resolving the basic issues linked to lack of significant wealth redistribution by about 2030, then the world may face a bigger crisis than for example the WWII. We cannot create islands of sustainability. We cannot enjoy a sustainable life in an unsustainable world.

It is quite likely that such a programme may be initiated by the G7 countries supported by OECD, as it was the case with the introduction of the world-wide minimum 15% corporate tax. The proposal was put forward in October

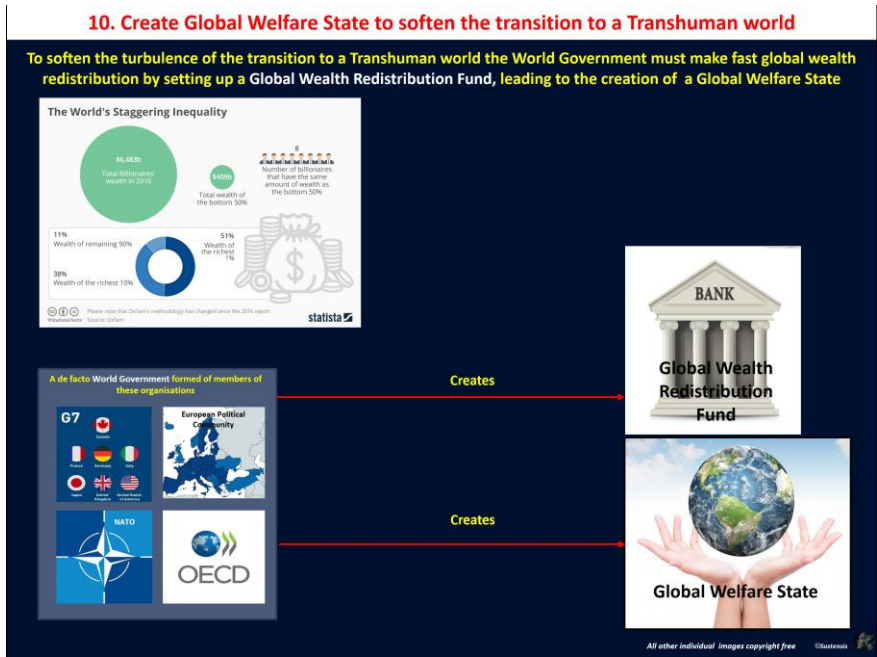
2021 and finalised a year later, signed by 140 countries. So, it is possible to implement a significant, nearly a global solution, in a very short time. Apart from purely ethical reasons, wealth distribution will also become a shield against other catastrophic risks. These include severe drought, mass migration and political disorder, which when combined with other catastrophic events, such as a more aggressive pandemic than the Covid-19, may become an existential threat.

On 25 September 2015, the United Nations passed the resolution on Post 2015 Development Agenda, officially known as “Transforming our world: the 2030 Agenda for Sustainable Development”. It is a broad intergovernmental agreement that acts as the successor to the Millennium Development Goals (SDG), which involved 193 Member States. It contains 17 “Global Goals” with 169 targets. I believe that the SDG provides an excellent opportunity for the world to use this framework for much wider objectives, which would subsume the SDG. These are:

1. **Create the wealth redistribution programme**, so that the donor countries (mainly the Northern Hemisphere) will over a decade transfer some of its wealth to those countries that need it most. To achieve that, we need a systemic global shift of wealth from richer to poorer countries. This is necessary for three reasons:
 - Make good the incredible suffering and economic robbery that some rich countries have done over a few centuries in their colonies,
 - Eliminate mass economic migration,
 - Control climate change-originated starvation, especially in Africa.
2. **Control mass economic migration** in such a way that there will be no need to migrate. That may mean solving not only the poverty problem (mainly economic and health related) but also environmental (scarcity of water) and political (civil and ethnic wars).

To achieve these objectives, I propose redistributing the global wealth more evenly by creating the **Global Wealth Redistribution Fund (GWRF)**. Apart from more equal wealth distribution it will also soften the turbulence of the transition to the AGI world. Such fund could be part of the UN Development Program (UNDP) but seeing the UN’s ineffectiveness and inefficiency in this area, I doubt it would attract funds at the scale that is needed. It clearly contrasts with an outstanding success of private funds such as Bill and Melinda Gates Foundation, with its vision to: “give people the

tools to lead healthy, productive lives, and thus help them lift themselves out of poverty”.



This is the key difference between how the UNDP and such a foundation works. The UN funds have been mostly giving the poorer countries a proverbial fish, whereas private foundations give them a fishing rod. Since 2000 that situation has improved at the UNDP, but other crucial differences remain. They are: lack of efficiency, effectiveness of the projects and still corruptive distribution of funds. It looks highly unlikely that the UNDP will change significantly to become the driver of such a large-scale wealth redistribution programme. Therefore, G7-initiated programme may be a better approach, with additional injection of funds from other sources to finance the target GWRF projects, and if possible, co-ordinate wealth distribution. A lot depends on the results of the US 2024 presidential elections.

If it is to work, the scale of this programme should exceed any help or fund distribution the world has ever seen and be from the beginning a decades-long continuous effort. Such a large scale wealth redistribution is the only realistic long-term solution for maintaining global peace, preparing Humanity for a transition to a new planetary civilisation. Wealth

distribution, if it is carried out on such a large scale, and if it follows the principles proposed here, should have three objectives:

- To stop economic and climate-change originated global migration,
- To create a more just and equal global society, fully achieved by the middle of this century,
- To become a powerful and pragmatic mechanism for political change by instilling Universal Values of Humanity in all parts of the world.

Building a Welfare State

The gap in wealth distribution between the richest and the poorest, both in developed countries and world-wide, is rising every year even faster than before. That process cannot continue for two reasons. First of all, it is dangerous for the democratic system itself and is ethically unacceptable. It is dangerous from a global perspective that an individual, like Jeff Bezos, the president of Amazon, owns assets worth the whole budget of a small country and can thus influence politics and economy on a global scale. This of course is the consequence of the crisis of capitalism and the way in which wealth, or sometimes pseudo wealth (i.e., trading in derivatives) is generated.

The changes proposed below are not only needed to finance the GWRF programme but also to minimize the risks of other excesses linked to gigantic wealth and power concentration in the hands of just a few hundred people. Such radical reforms are also needed to avoid massive Global Disorder in the transition period to the time when AI will be let out of human control. Global chaos would almost certainly make it impossible to retain control over AI for longer than 2030. However, if we retain such control over the AI's self-improvement beyond 2030, then it will not only start mitigating other risks, such as Global Warming, but also help us sustain rapidly expanding Global Welfare State.

But setting up Global Wealth Redistribution Fund (GWRF) is only a part, but perhaps the most important one, of building a Global Welfare State. Therefore, we must also include people in the rich countries, where income inequality is equally large. The question is where will we get the money from? Although there are other changes necessary to finance GWRF, e.g., special taxation, I propose to reform just two areas: individual wealth and corporate wealth.

Additionally, I propose additional sources of capital or significant cost reductions in most sectors that should be available from about 2030 in developed countries. These are: Much higher than predicted GDP growth, raising taxes to finance better lifestyle and wellbeing, demonetization – significant fall in prices, and substantially lower cost of government. I describe them in more detail below.

1. Setting an individual's wealth cap

That is a fundamental change in individual's rights to property, which has been sacrosanct in every democracy. But it cannot continue any longer.

The first reason is that letting extremely wealthy individuals continue to accumulate wealth will lead very shortly to a situation where they will become as powerful as many smaller states. Today just 10 richest individuals' wealth equals the annual GDP of the states such as Brazil. 1% of the world richest people have more wealth than 50% of the world population^[58].

The second reason is to minimize the impact of very rich people on political decision-making, whether direct, in case of oligarchs, or indirect by manipulating democratic elections or voting in the parliaments.

The third reason, briefly mentioned earlier, is to minimize the risk of Global Disorder, resulting from individual risks such as drought causing famine, local wars, mass migration, financial sector collapse, or Technological Unemployment. When they combine into a Global Disorder, it may then become an existential threat. If in such a situation we still have immensely rich individuals, the anger and frustration at the wealth and power of these people may ignite revolution in many countries.

Therefore, there should be a one-off 100% tax imposed on the richest billionaires, so that their personal wealth is capped at, say \$1bn (I repeat that all numbers and names are only examples). Just that one-off tax may amount to several trillion dollars, i.e., equivalent to approximately 2 year GDP of all African countries. Any future personal wealth growth above that cap of \$1bn would be automatically taxed at 100%. However, such taxpayer would have a say on how that tax may be used. For example, they may indicate a not-for profit organization, which they have already been funding, or any other public organisation they would like to help. There is little hope such a tax could affect Russian or Chinese oligarchs. However, once such a legislation

is sanctioned e.g., by OECD, the wealth of those oligarchs outside their native countries, may be confiscated and their global travelling severely restricted.

Such a legislation may start in the US. I ignore the immense difficulties in implementing this scheme, but I believe it should be done soonest possible. The US Congress may never agree to it, so the only way might be to start with the President’s Executive Order and then trying to pass it through the Congress.

2. Setting the corporate assets’ cap

It is enough to look at the table below to see where the problem lies. I would ignore the first two largest companies, Walmart and Amazon, which are retailers and have a limited production of their own. So, let’s look at Apple, which is a manufacturer and to a lesser extent software developer with an annual revenue of \$365Bn. I have done some quick calculations to see what it really means (all data based on Wikipedia 2023). That \$365Bn is the same amount as the annual GDP of Iran and higher than that of over 150 other countries. But that company also has annual profits reaching nearly \$100Bn. That is more than the GDP of 130 countries. Finally, Apple’s annual profits equal the combined GDP of 50 poorest countries, which means that this single company could ‘feed’ the population of these countries every year (ignoring the accumulated assets of those countries).

20 largest companies by revenue							
Rank	Name	Industry	Revenue	Profit	Employees	Headquarters	State-owned
			USD millions				
1	Walmart	Retail	\$572,754	13,673	2300000	USA	✗
2	Amazon.com, Inc.	Retail	\$469,822	33,364	1608000	USA	✗
3	State Grid Corporation of China	Electricity	\$460,616	37,137	871145	China	✓
4	China National Petroleum Corporation	Oil and gas	\$411,692	9,637	1090345	China	✓
5	China Petrochemical Corporation	Oil and gas	\$401,313	8,316	542286	China	✓
6	Saudi Aramco	Oil and gas	\$400,399	105,369	68493	Saudi Arabia	✓
7	Apple Inc.	Electronics	\$365,817	94,680	154000	USA	✗
8	Volkswagen Group	Automotive	\$295,819	18,186	662575	Germany	✗
9	China State Construction Engineering	Construction	\$293,712	4,443	368327	China	✓
10	CVS Health	Healthcare	\$292,111	7,910	258500	USA	✗
11	UnitedHealth Group	Healthcare	\$287,597	17,285	350000	USA	✗
12	ExxonMobil	Oil and gas	\$285,640	23,050	63000	USA	✗
13	Toyota	Automotive	\$279,337	25,371	372817	Japan	✗
14	Berkshire Hathaway	Financials	\$276,094	89,795	372000	USA	✗
15	Shell plc	Oil and gas	\$272,657	20,101	82000	UK	✗
16	McKesson Corporation	Healthcare	\$263,966	1,114	66500	USA	✗
17	Alphabet Inc.	IT and AI	\$257,637	76,033	156500	USA	✗
18	Samsung Electronics	Electronics	\$244,334	34,293	266673	South Korea	✗
19	Trafigura	Commodities	\$231,208	3,100	9031	Singapore	✗
20	Foxconn	Electronics	\$214,619	4,988	826608	Taiwan	✗

[59]

What is the impact of such companies like those listed above? In a word – enormous. It is multifaceted. Their impact is like an oil on the water surface continuously expanding, in line with the company size. Here is a list of the main areas, which are usually influenced by such large companies:

- **AI control.** Let's start here because this is what this book is about. Apple, Google (Alphabet), Samsung and Foxconn are in that list, with Microsoft with its \$180Bn revenue closely behind. All these companies are leaders in the AI research and development. But even the companies which are not directly engaged in AI may also finance clandestinely R&D to achieve their own goals. If they are out of public control and manage to produce their own AGI, it may escape human control with unforeseen consequences. This of course also applies to all autocrats and dictators like the N. Korean leader,
- **Monopolistic Practices.** Large corporations can use their size and resources to dominate markets and engage in monopolistic practices, such as price-fixing or abusing their market power. This can result in higher prices for consumers, consumer choice, and lower quality products and services, reducing the overall competition and consumer choice,
- **Influence on Politics.** Very large corporations often have significant impact on politics by using their financial resources to influence the political process. This can undermine democratic institutions and lead to policies benefitting these corporations at the expense of the public interest. The American lobbying system is the best examples. The lobbying money is the oil of the American democracy (although there are other elements there, which also adversely affect it, like a large gun lobby). Just consider that during the 2020 presidential elections the cost of privately financed presidential campaign for Joe Biden exceeded \$1.6Bn and \$1.3Bn for Donald Trump,
- **Impact on the labour market.** Large companies may often exploit their employees, by paying them low wages, enforcing long working hours, or having unsafe working conditions, in order to increase profits. This can widen the income gap between the rich and the poor, and lead to the increased social inequality and lack of job security,
- **Environmental Damage.** Very large corporations often have a negative impact on the environment by polluting air and water, depleting natural resources, and contributing to Global Warming. Quite often they exploit natural resources in an unsustainable way, which harms the ecosystem,

- **Social inequality.** These companies may contribute to social inequality by concentrating wealth and power in the hands of a small group of individuals, which I have highlighted above. This further exacerbates income and wealth inequalities limiting opportunities, especially for the disadvantaged groups of population,
- **Stifling Innovation.** Although large corporations invest in research and development, they may also stifle competition by focusing the investment only as far as it maintains their market position and profits. They would avoid more risky investment, especially in so called disruptive innovations. This can lead to stagnation and lack of progress in certain industries or areas. The best example could be the court battle of James Dyson, the inventor of cyclonic vacuum cleaner with AEG and Electrolux in the 1990'. Those two large companies were fighting to maintain their old technology and the market position against Dyson's disruptive technology.

These negative impacts can have far-reaching consequences for the society and the economy, affecting everything from economic growth and innovation to public health and well-being. Of course, the impact of very large corporations can vary. It depends on the specific company and the industry involved, as well as an overall regulatory environment and the state of the economy.

It is for these reasons that there should be a global cap on a company's asset value. Let's suppose that the initial cap is \$50Bn. The company might then pay a one-off windfall tax on the difference between its actual book value and \$50bn. The excess in the book value above \$50Bn would then have to be sold. In the following years, the cap could be lowered further, and additional taxation imposed, depending on the market conditions.

This is of course one of many ideas how to curtail the size of super large companies and improve the competition. The money thus collected both from the richest individuals and from the largest companies, should be largely funnelled to the Global Wealth Redistribution Fund.

3. AI-generated new type of wealth

Most of this kind of wealth would normally not be included in the GDP growth. This is the generator of wealth in every aspect. I immediately admit that a lot of the savings in this category will impact the fall in prices or will have already been in some way included in the previous sources. However,

there would still be some ‘leftovers’, which are difficult to quantify. They will emerge as new capabilities, never possible before, e.g., humanoid assistants providing elderly care in care homes.

4. Much higher than predicted GDP growth

This source is rather unusual, since it involves turning a problem (too low GDP growth) into an opportunity (much higher growth than would have been expected). This will be due to unprecedented growth of productivity driven by an exponential progress in technology, mainly in AI. Only very few economists share that view. Most are still entrenched in the old times, calculating growth as everything around us was happening at a linear, rather than exponential pace.

OECD in its long-term forecast assumes 3% annual growth rate for OECD countries between 2015 and 2040 (measured in Purchase Power Parity dollars, reflecting the real purchasing power of a basket of goods)^[60]. PWC assumes the GDP growth rate in developed economies over that period would be between 1.5 to 2.5% [61]. Most of the long-term projected GDP growth ratio for developed countries oscillates around 2.5%. How credible is such a long-term growth rate? In my view it is not very credible.

Who is right - orthodox GDP growth setters, or entrepreneurs and fringe economists? Right, in my view, are quite probably those people who do not have a vested interest in retaining the status quo. We face a similar situation today as regards the actuarial data that support the calculation of pension funds and their long-term liabilities. In most cases the data provided by actuaries is hardly credible. However, since the data is prepared and used by people who have a vested interest in pretending that everything is all right (that pension contributions are adequate to pay for future pensions), the contrarians have little chance to win the argument.

In many forecasts, almost everything depends on initial assumptions. One of such assumptions is that change in all domains will broadly happen at the same pace as before. The reason for that is that there is simply no other data that could be the basis for assuming something entirely different as far as the GDP growth rate is concerned, i.e., suddenly rising much faster than linearly (as it must have because of the unprecedented AI-driven technological revolution).

Most economists still assume that the four components of productivity growth: labour, capital, technology (resources) and organization (entrepreneurship) will grow largely unchanged as before. They seem not to appreciate how substantially the role of one of the growth factors, technology, has changed over the last decade and that the productivity growth arising from that factor, is now reaching the level of the Moore's law, i.e., doubling every 18 months. That's why the vast majority of economists still assume that the global GDP growth over the next 20-30 years will rise at about 2.5-3% p.a. That would mean that the world's GDP growth would barely double by 2040. Well, economists, governments and in general, orthodox thinkers and planners, have been spectacularly wrong on many occasions. Let me give you some examples:

- 2008 financial crisis. Nearly all major economists from Harvard and MIT, have not predicted that catastrophic failure because they believed Milton Friedman's assertion that markets know best and they would re-adjust themselves,
- 2015-2016 migration crisis in Europe. One of Germany's justifications for letting the migrants in was that Germany will need 10 million new employees by 2030. As I have shown earlier, the reverse is almost certain to happen, there may be more than 10 million Germans unemployed in 2030,
- In Technology, Elon Musk with his Space-X Falcon 9 rocket, has within less than 10 years with his team achieved something that NASA or any other governmental organization were not able to do, i.e., to reduce the cost of payload vs. Saturn 5 rocket by about 20 times, among others, by re-using the same rocket^[62]

However, the inaccuracy of GDP calculation today may be a relatively small problem. In calculating GDP growth rate for 2040, we may be several times off the real figure. That is why I am saying that the establishment is the least credible body to make correct judgments because of vested interests. Take another example. In the most recent USA election, Mr Trump promised to repatriate largely manual jobs from China back to the USA, especially to the automotive industry. That was grossly misleading as it was simply impossible for many reasons. One of them is the fact that since 2008 crisis the American automotive industry has received, billions of dollars in direct or indirect aid, for restructuring their industry. The result is that it is now a significantly different industry with a much higher productivity than before. All thanks to the very latest technology. At Ford or GM, an hour of a robot costs now less than \$8 against £30 for a blue-collar worker. That means that

the Chinese manual jobs could not be transferred to the USA because robots have taken them up. That kind of increased productivity like the one at Ford or GM, is still not properly being included in GDP model calculations, because it is like trying to hit the moving target, the data change too rapidly and is unstable.

But there are also other factors that point to GDP undervaluation. The global GDP is also undervalued because of purposeful action of some governments e.g., China, which undervalues its currency to boost export. Of course, if GDP is calculated in the same way, its growth rate will not be impacted. However, what will change is the real value (substance) that will be delivered, or the purchasing power of a country. What I mean by this is that every year we consume more than would have been expected from the GDP growth alone. The “Economist” magazine has for many years calculated the GDP value using the number of hamburgers that can be purchased in any given country to reflect the meaning of real value (Purchasing Power Parity). In every country, the actual real value of goods delivered year by year is higher than the GDP growth would have indicated, and which partially forms the black economy (only a small part of it is by default included into GDP and in taxes).

Therefore, GDP growth will not follow the previous path. Instead, fuelled by relentless robotization and innovation, sometimes even exceeding exponential growth, (e.g., cost of artificial hamburger production fell 30,000 times in just three years) ^[63] GDP growth will be much faster even in developed economies.

This perception of more or less the same GDP growth is mainly due to missing the moment when change has passed the tipping point (called “knee curve” by economists) and from when change is accelerating exponentially. I believe we are just about that point, which means GDP growth will accelerate faster than orthodox economists envisage. For example, the whole agricultural sector in 20 years’ time will look entirely differently than today, because it will be cheaper to produce most food from stem cells and basic chemicals. Similar growth will be achieved in the productivity of various medicines (cutting down the time from a medicine discovery to the time it can be bought at a pharmacy) proven by Google’s Alfa-Fold. Finally, in education, students will be educated mainly in a one-to-one tuition mode on the websites such as the Khan Academy or directly by AI Assistants, such as ChatGPT. We should, therefore, expect the GDP growth in real terms to at least treble by 2040. That will be an additional sizable income, which will

allow financing of new social arrangements, like the Universal Basic Income (UBI) and the Global Wealth Redistribution Fund.

5. Raising taxes to finance better life satisfaction

This is the most typical source of finance for every government, although in this case even more important is the reason for doing that and its ultimate outcome. There is little correlation between higher taxes and higher level of happiness, or what I would prefer to call contentedness, as measured for example by the UN's Human Development Index. Much more important is the government's efficiency, the strength of democratic institutions, which is directly linked to the level of corruption. Taxes should be a means to an end and not the source for an easier ride for the government to fulfil its sometimes entirely ideological commitments. And yet, the 2022 UN World Happiness Report ranks four Scandinavian countries at the top of the list, with Finland (a very high taxation country) still being the top country five times in the last 10 years,^[64]

Top 10 happiest countries, 2022

1. Finland
2. Denmark
3. Norway
4. Iceland
5. Netherlands
6. Switzerland
7. Sweden
8. New Zealand
9. Canada
10. Australia

Why is Finland at the top of the list of the happiest people in the world, a country of 5.5 million people that only 150 years ago suffered Europe's last naturally caused famine? After all, GDP per capita in Finland is lower than even in its neighbouring Nordic countries and is much lower than that of the USA. As all Nordic countries, they pay some of the highest taxes in the world (52%). But the Finns are good at converting wealth into wellbeing delivered by efficient and effective government. That's why just paying higher taxes does not necessarily correlate with life happiness. In Finland there is wide public support for higher taxes because people see them as investments in a good quality of life for all. The country has also been ranked

as the most stable, safest, and best governed country in the world. It is among the least corrupt and the most socially progressive country with its police being the world's most trusted and its banks the soundest.

As you may have noticed, the Scandinavian system of government is for me one of the best overall in the world. Yes, Switzerland is an exception, as it is in many other aspects of government, having much lower taxation level and still being the 5th happiest country in the world. It is the country that did not have a war for 800 years, so its wealth has been accumulated for a very long time. Therefore, the case of Switzerland is not a good argument to claim that one can have a high standard of living, while also paying low taxes. If the government is efficient and effective, then higher taxes (at a certain level, not stifling the economy) would simply mean better economic and social personal outcome. That means the projects financed by the government, such as in transportation, are delivered on time and on budget.

6. Demonetization: significant fall in prices and faster growth of real income

This could be the result of a direct fall in prices (low inflation or even deflation) and indirect, through product substitution and product efficiency (a much greater value). By about 2040, we will be in the period of continuously falling prices and a faster growth of real income, i.e., demonetization of the cost of living. This would mean that it will be cheaper and cheaper to meet people's basic needs. All this will be driven by exponential growth in technological solutions and innovations in most sectors, leading to significant cost reduction in clothing, health care, housing, transportation, food, education, or entertainment.

Just think about this: the real value that is delivered to all of us, like Google applications, GPS, and other similar technology-originated services is not included in GDP because it is free! If you were to pay in 1982 for the facilities and services that you have on your mobile phone, then they would be worth, including inflation, well over \$1million in 2023, not to mention a vastly superior quality, and unavailability of some services in 1982, like personal weather forecasting. Another example, video conferencing equipment in 1982 cost about \$250,000 (plus the actual cost of carrying out the video conferencing). Today, WhatsApp or Zoom applications that anybody can have on a smart phone is entirely free. Perhaps you only now realize that the phone you hold in your hand makes you a millionaire, as this table proves so clearly.

Dematerialization					
>\$900,000 worth of applications in a smart phone today					
Application	\$ (2011)	Original Device Name	Year*	MSRP	2011's \$
1. Video conferencing	<i>free</i>	Compression Labs VC	1982	\$250,000	\$586,904
2. GPS	<i>free</i>	TI NAVASTAR	1982	\$119,900	\$279,366
3. Digital voice recorder	<i>free</i>	SONY PCM	1978	\$2,500	\$8,687
4. Digital watch	<i>free</i>	Seiko 35SQ Astron	1969	\$1,250	\$7,716
5. 5 Mpixel camera	<i>free</i>	Canon RC-701	1986	\$3,000	\$6,201
6. Medical library	<i>free</i>	e.g. CONSULTANT	1987	Up to \$2,000	\$3,988
7. Video player	<i>free</i>	Toshiba V-8000	1981	\$1,245	\$3,103
8. Video camera	<i>free</i>	RCA CC010	1981	\$1,050	\$2,617
9. Music player	<i>free</i>	Sony CDP-101 CD player	1982	\$900	\$2,113
10. Encyclopedia	<i>free</i>	Compton's CD Encyclopedia	1989	\$750	\$1,370
11. Videogame console	<i>free</i>	Atari 2600	1977	\$199	\$744
Total	free				\$902,065

*Year of Launch

Source: Peter Diamandis and Steven Kotler: “Abundance: The Future Is Better Than You Think”, 2015

In 2040, it would be even more spectacular and likely that most of the things you appreciate will come free of charge. So, in 20 years’ time we may be living in a world of abundance, at least twice as rich as today in real terms. Objects of desire previously beyond reach of an average consumer will become affordable. If we only assume a very conservative doubling of GDP in real terms in 2040, that will make today’s EU average personal net income of €20,000 per annum equivalent to what will then be the ‘poverty line income’. That means, **everybody in 2040 would have as a minimum, today’s EU average income in real terms, even if he would not work.**

Peter Diamandis in his article ‘Why the Cost of Living is Poised to Plummet in the Next 20 Years points out that in the U.S, in 2011, 33% of an average American's income was spent on housing, followed by 16% spent on transportation, 12% on food, 6% on healthcare, and 5% on entertainment. In other words, more than 75% of Americans' expenditure covers: Transportation, Food, Healthcare, Housing, Energy, Education and Entertainment^[65].

7. Substantially lower cost of government

This is a significant area of cost savings. Potential savings here are on the scale we cannot even imagine because a lot of these savings will be generated by the inventions that are not there yet. So, let me give you only some examples of those savings:

- All taxes and benefits will be collected and distributed with almost no human intervention. This will be the continuation of the process started in earnest about 2000 in all EU countries,
- Cost of running health care and medical care will fall dramatically as indicated above,
- The same is true about education, which will make a very intensive use of humanoid Assistants that will co-operate with teachers. There will be far fewer teachers needed with AI assistants. Those most needed will be behavioural and psychology specialists playing a very important part in the overall education,
- The cost of the army will also be significantly reduced because if humans survive this decade it is unlikely there will be any significant wars anymore. There simply be very few enemies to fight. And those who might be capable to start a war would know there will be no winners, because such wars will be largely 'fought' by AGI systems. These, on the other hand will have difficulties to distinguish between the 'good and the bad' guys, so may follow the first Asimov's law: do no harm to humans!

Tony Czarnecki: Prevail or Fail

PART 3

The Civilisation of Transhumans

1. The dawn of a new civilisation

The building of a Welfare state has completed 'The Principles', and its scheduled task list. We have arrived at the point where humans will start living in a new civilisation. But before we get there, we must go through a perilous transition period. That journey may itself become an existential threat, if we stubbornly stick to what we know, preserving the status quo. Therefore, this final part of the book looks at what needs to be done to go through that transition period without unnecessary turbulence.

I would probably have not written this book were civilisational changes continue to develop in a linear pace. There would have been no AGI by 2030 and no need to control AI goals and its behaviour, because AGI would have not been developed yet to such a level, where it may threaten us. Neither would there have been a strong pressure on changing fundamentally our ways of living. This single aspect of our reality, the acceleration of the pace of change from a linear to a nearly exponential in many areas, will cause significant turbulence in the way we live. This will be mostly noticeable in politics and in the relationship between us the voters and those who govern us. But there is another aspect of governance. It is our ability to govern the process of AI's self-development, so that it remains under our control. That is why this first Principle 'Adjust global AI governance to a civilisational shift' is the most important one. All the remaining Principles derive their urgency from this one. Should we be unable to control AI development until it inherits our values and preferences and has learned from its own experience what it means to be human, we will simply have no future as a species.

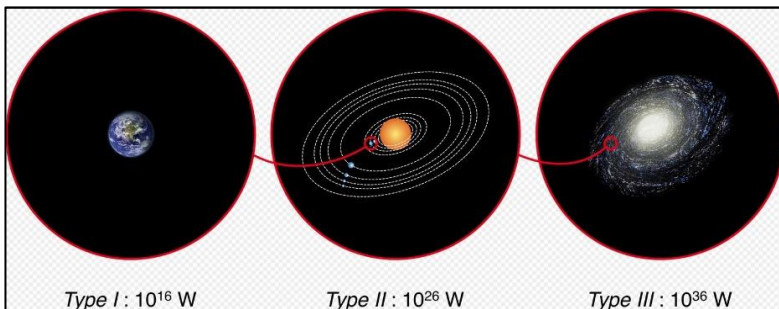
So, if you accept this starting premise then consequentially it will be easier for you to understand the need for change and solutions that must be applied for us to adjust to what is definitely a civilisational change. But I do not mean 'civilisation' in a cultural or historical sense, like agrarian, Assyrian, Roman, or industrial civilisation. I am talking about a civilisation in an evolutionary sense. If we take an historic view, then a differentiating factor for various stages of Humanity's progress would be technology. It is the technology, which ultimately underpins and differentiates civilizations across the millennia.

Physicists define civilizations by the **energy** level that could be available for its growth. In 1964 the Russian astrophysicist Nikolai Kardashev defined three types of civilizations differing by the order of energy they had available

to them, measured in Watts (W). Each civilization differs from the other by 10 orders of magnitude. Here is a succinct summary of the so-called Kardashev scale ^[66].

- Type I civilization—also called a **planetary civilization**—can use and store all of the energy which reaches its planet from its parent star. This is our civilisation,
- Type II civilization—also called a **stellar civilization**—can harness the total energy of its planet's parent star (e.g. using the Dyson sphere),
- Type III civilization—also called a **galactic civilization**—can control energy on the scale of its entire host galaxy.

Here is an illustration of how such civilizations might evolve:



Energy consumption estimated in three types of civilizations defined by Kardashev scale^[66]

A similar approach to defining civilization is proposed by Michio Kaku, a renowned US physicist^[67]. It is a derivative of Kardashev scale but proposes a more generic differentiating factor - the use of **resources**, which of course also includes energy.

- **Type 0 Civilization** is essentially our civilization. A type 0 civilization has only just begun to tap planetary resources such as solar power, geothermal power, and wind power. Most of its power generation is still based on non-renewable fossil fuel resources, for example, oil, coal, and natural gases.
- **Type 1 Civilization** can effectively control the entire resources of their planet; they can predict weather patterns and earthquakes very accurately and even control them, by using artificially induced

greenhouse effects or space-based lasers. A Type 1 Civilization could conceivably halt an ice-age.

- **Type 2 Civilization** has the capability to extend their power to their entire Solar System by harnessing the power of their suns through Dyson spheres. Having colonized or at least extensively explored all the planets within their Solar System, they are a largely space-faring race and have already mounted expeditions to other stars using interstellar craft.
- **Type 3 Civilization** spans entire galaxies having colonized all the stars by wave after wave of interstellar craft. They can harness the power of galaxies. By utilizing the millions of black holes that are believed to reside within galactic nuclei, type 3 civilizations would have sufficient power to conduct truly universe-changing high-energy physics experiments and examine matter down to the Planck scale.

I quote this civilisational framework to map a potential advancement of humanity into a new species by gradually morphing with Superintelligence until it becomes Superintelligence itself. When AI becomes AGI, then just in about two decades from now, it will evolve into Superintelligence. Then within months, if it has sufficient resources, mainly energy, it will reach the so called Technological Singularity. If we do not destroy ourselves in the next decade or two, then this evolutionary jump to Kaku's Type 1 civilisation will occur in this century, when the only break on exponential pace of change will be limited resources. That is why the mining of resources on other planets and asteroids will be a top priority underpinning the expansion of the new civilisation.

However, if we want to be the masters of our own future, the first task we need to complete successfully is to deliver benevolent Superintelligence. We can only do that if over this decade we will govern the development and use of the maturing AGI and then Superintelligence in such a way that it will be guided by the Universal Values of Humanity, and by the way, in which humans prefer to live and interact with each other.

2. Who are Transhumans?

Three aspects of Transhumanism

Transhumanism is quite often linked to something that has nothing to do with it, like a transgendered sex. Confusion may arise because there are many definitions of Transhumanism, like this one on the Wikipedia: “Transhumanism shares many elements of humanism, including a respect for reason and science, a commitment to progress, and a valuing of human (or transhuman) existence in this life”. The US Transhumanist’s website Humanity+ describes it as follows: “Transhumanism challenges the human condition. This condition asserts that aging is a disease, augmentation and enhancement to the human body and brain are essential to prevail, and that well-being is essential to prosper within safe and healthy environments”.

Unfortunately, from a civilisational transition perspective none of these definitions is complete. It is not enough to say that Transhumanism ‘challenges a human condition’ or it is about ‘valuing human existence in this life’. We must also identify its ultimate goal, i.e., where such a transition ends. Therefore, in this book I describe it as follows: **“Transhumanism is an approach proposing Humanity’s transition to its coexistence with Superintelligence until humans evolve into a new species”**.

One of the most comprehensive visions of Transhumanism can be found on the Transhumanist UK website, and in particular in David Wood’s book – ‘Vital Foresight – The Case For Active Transhumanism’ [68]. But Transhumanism can also be seen as three ‘Supers...’: Superlongevity, Superintelligence and Superwellbeing, which is excellently illustrated in a short video by the British Institute of Posthuman Studies [69]



Three aspects of Transhumanism

So, Transhumanism is a period when humans make a transition to a new species. It will be achieved in two stages:

1. **A passage to a new civilisation.** This has already started although we are barely aware of it. An almost exponential pace of technological advancement in AI will ultimately create **Superintelligence**. That middle interconnecting door of Transhumanism will lead to the world of unimaginable abundance and a much more equitable society. That is **Superwellbeing**, which is not just about a material wealth. It is also about a rapid advancement in medicine. That will enable very significant extension of healthy life, allowing people to live well over 100 years. The very recent discoveries by AI of 3 chemical compounds out of 800,000, each capable of harmlessly removing by injection all senescent cells, is the best example. This and several other methods will lead to a nearly miraculous regeneration of a human body and the extension of a human life by several decades by about 2030. That is **Superlongevity**. However, to achieve all that, we need to maintain global peace by a significant reform of democracy and a gradual federalization of the world. We must finally behave like the citizens of the planet, rather than isolating from each other by strengthening national borders,
2. **A transformation of humans to a new evolutionary form.** That is what is missing from the definition of the British Institute of Posthuman Studies. But I would suggest that is the ultimate goal of Transhumanism - to pave the way for humans to become Posthumans.

Therefore, I would see Transhumanism as not just a passage to a new civilisation due to socio-technological progression over millennia, e.g., from an agrarian to an industrial era, when humans as a biological species remained unchanged. It is about becoming a new species, following not just an earthly evolution, but a cosmic evolution, when matter transcends into an inorganic intelligence. That will be enabled by an exponentially advancing Artificial Intelligence, which in just a few decades will become Superintelligence.

Please note: As in the whole book, I use the term 'Superintelligence' meaning the most advanced AI system.

You too can be a Transhuman

If this is the first time you read about Transhumanism and Transhumans and you will be thinking 'this is just incredible' (to put it mildly), then just think about yourself. Most of us are already partly Transhumans. Our smart phones give us enormous extra intelligence, which we could not dream about even 10 years ago. The only difference is that the extra intelligence is currently external.

There are already over 10,000 people in Sweden alone who have one or two microchips implanted under their skin, usually in their hand or in the arm, which allows them to activate certain electronic devices or get access to protected areas, instead of entering passwords.



Swedish 'Transhumans'

There are also thousands of brain implants controlling parts of the body affected by moto-neuron disease, epilepsy, eyesight impairment, etc. One of the examples is of a completely paralyzed person, who in October 2019, wearing an exoskeleton, started to 'walk' using an implant in his brain. The implant was connected using wires. The scale of this achievement overshadowed anything that had been done in that area until then. Today, there are thousands of paralyzed people with brain implants controlling their nerves and muscles and thus enabling them a nearly normal life. Unfortunately, they are still very expensive.

Although my definition of Transhumans does not negate the above achievements, I focus on the Brain-Computer-Interfaces (BCI) devices' capabilities, which will gradually enable humans to extend their **mental capabilities**. As technology progresses, so will the capabilities of Transhumans. With time, more and more of their mental capabilities will be supported by various BCI devices, which will serve as a wireless gateway to external resources, such as huge memory, processing power, audio-visual

devices and to avatars. Within the next 2-3 years, humans will be able to control their humanoids by thoughts, using wireless BCI devices.

Over this decade, Transhumans' cognitive capabilities (e.g., memory and processing power leading to a much faster decision making) will increase so much that they will become far more intelligent and capable than biological humans. Using BCI devices, some Transhumans may become invaluable, if selected by international bodies, such as the UN, to help us pass in relative peace the most dangerous period in human existence.

In a few decades, the body of Transhumans will become more and more 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 fused with a purely digital Superintelligence, becoming oneness. Unless there are some physical or physiological obstacles e.g., related to porting consciousness into digital chips, an entirely new, non-organic species – Posthumans will emerge. Thus, for me:

Transhumans are the people, who have their mental capabilities extended by Brain-Computer-Interface (BCI).

At this point you will probably be asking a question: why we have to choose such a path for a human evolution. For you and many others this might be an instant call to arms – stop it! But perhaps to surprise you, I would have also been among those calling to do just that, since for me that weird future is not something I would like for my children and grandchildren. We think and feel in a human way and cannot imagine that one day future generations may not have any privacy or even identity as we understand it today.

Unfortunately, we can no longer stop that process, unless we destroy in some way our civilisation, returning perhaps to the age of steam. But then we would be facing the same dilemma within a century. As Edward Teller, the physicist quoted earlier, said 'we cannot uninvent a nuclear bomb'. Neither can we uninvent the Internet. We cannot really stop it anymore because that's how it was designed, nor can we entirely destroy Google database, or stop exponentially improving AI. Any global ban on a further development of AI would be futile. A dozen of top AI specialists supported by a deranged billionaire of Dr Strangelove-type, not to mention rouge states, such as N. Korea, would be capable to continue clandestinely further improvements of AI agents, until such day, when they would hope using it to rule the world. Therefore, we have no other option than continue developing AI but under

strict human control. That might give us some choices regarding our future evolution.

So, how are we going to navigate that pathway to a Transhuman world? In his interview in May 2020, Elon Musk confirmed indirectly his long-term aim: “Even in a benign AI scenario we are being left behind. So how do you go along for the ride? If you can’t beat them, join them.” This can be interpreted as something like: we shall have ‘proper’ Transhumans within a few years’ time and we desperately need them because they may be our best safeguard against a malicious Superintelligence. In that sense, the key role in that transition falls to the first selected Transhumans whose responsibility would be to control the maturing Superintelligence from within as its human Governors.

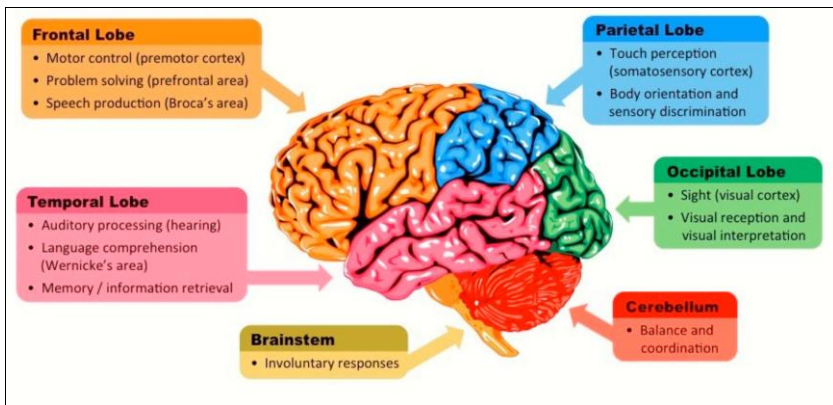
3. Making a Transhuman

Brain Computer Interface may turn you into a Transhuman

There are no people yet who have any brain implants, which might enable them to communicate wirelessly to store some information from their memory directly to external devices. But such devices are going to be implanted into humans in the next 2-3 years. In February 2020, Elon Musk's company Neuralink, announced that it was building tiny and flexible 'threads' which are ten times thinner than a human hair and can be inserted directly into the brain. They intended to have first human implants with 3,200 electrodes ready within a year. However, in February 2021 scientists developed the first brain implants using wireless technology. They also achieved communication with a selected cluster of about 1,000 neurons using ultrasound waves, rather than injecting extremely fine electrodes.

To be widely used, the Brain-Computer-Interfaces (BCI) must be reliable. That requires among others a separation of the brain signals from an overall electromagnetic noise. There are two ways in which the accuracy of brain signals may be improved: either implanting the signal collectors closer to the signal sources, e.g., deeper in the brain, or having much more sensitive devices. The progress in this area is very fast indeed. There are several brain mapping projects, intended to decode all brain functions. One of them, the Connectome project, intends to complete a full Brain Map by 2025, which is broadly in line with the Neuralink's expected delivery date of some more advanced brain implants.

Our brain consists of three brains: Reptilian, Limbic and Neocortex.



Main functions of the brain^[70]

Our cognitive functions mainly reside in the Neocortex (Frontal Lobe and Temporal Lobe), Parietal Lobe (sensory perception) and Occipital Lobe (vision and visual interpretation). Altogether there are nearly 70 distinct functional areas, such as speech, memory, or sound.

Reptilian brain, i.e., Cerebellum and the Brain Stem, responsible for coordination and movement are of less importance for cognition and perhaps consciousness. However, they will be needed for a full mind uploading to enable a digital brain to control its mechanical avatars, replicating our biological body.

The Connectome method is suitable for a complete mind uploading because that will require all brain functions to be digitized, which may take some time. For the purpose of supporting Transhuman Governors (see next section), who would control AI, we need different methods. We need brain implants enhancing human cognitive abilities, which can be gradually extended by progressively adding more digitized cognitive brain functions. That means focusing on two brain lobes – Frontal lobe (neocortex), responsible for thinking, planning and decision making, and Temporal Lobe where memory, language and communication functions mainly reside.

From Humans to Transhumans - Brain Computer Interface (BCI)

Types of Brain Computer Interfaces (BCI)

creating Transcortex

The diagram shows a human head profile with various BCI methods labeled: Brain-penetrating microelectrodes, EEG sensor, and Transcortex. It includes signal waveforms for LFP (Local Field Potential) with parameters $< 1 \text{ mV}>$ and $< 200 \text{ Hz}>$; SPIKES with parameters $5-500 \mu\text{V}>$ and $0.1-2 \text{ kHz}>$; EEG with parameters $5-300 \mu\text{V}>$ and $\times 100 \text{ Hz}>$; and ECoG (Electrocorticogram) with parameters $0.01-5 \text{ mV}>$ and $< 200 \text{ Hz}>$.

Basic Concept

Invasive method (electrodes)

[E.g.] Elon Musk Neuralink's brain implant -09.2021

Non-invasive BCI (e.g. Meta's helmet) 100 words/min.

AI intelligence does not have to replicate the way human brain works
(Airplanes do not fly like birds)

©All other images copyright free Sustensis

There are 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^[71]. 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.

The progress in the Brain Computer Interface (BCI) area is phenomenal and probably faster than exponential. For example, in July 2022, a year old test of a new type of BCI device was completed by Synchron company in the USA. The company describes it as follows:

“Stentrode system centres on a permanently implanted stent-like device. It's inserted through the jugular vein to reach the motor cortex of the brain, where it can pick up neurological signals denoting an individual's intended actions. The device collects those signals from a receiver unit implanted in the user's chest, then translates them in real time into clicks and keystrokes on a computer or mobile device. An additional eye-tracking device is used to control the movements of the computer cursor. The intended result is a system that allows people with severe paralysis to send texts and emails, access online banking and shopping services, complete telehealth visits and more, all using only their thoughts to control the tech—therefore returning some independence to their lives”^[72] Since this is such an important breakthrough in BCI devices, here is some further most recent information based on an article published by CNBC in February 2023.

Synchron's Stentrode BCI device is inserted through the blood vessels, which the inventor, Thomas Oxley calls the “natural highways” into the brain. Synchron's stent is similar but hundreds of times narrower than the stents used to widen blood vessels in cardiovascular disease. It is fitted with tiny sensors and is delivered to the large vein that sits next to the motor cortex. The Stentrode is connected to an antenna that sits under the skin in the chest and collects raw brain data that it sends out of the body to external devices. Those raw signals are then interpreted by an AI apps, which converts them into text or instructions to control various devices. Since the device is not inserted directly into the brain tissue, the quality of the brain signal isn't perfect. However, the procedure is less invasive since it does not require brain surgery and thus is more accessible. There are already about 2,000 specialists who can perform these procedures.^[73] What is equally

important, is that the device was approved by the US Food and Drug Administration, so after further tests in humans it could be implanted if a doctor prescribes it.

Just one more example. In January 2023 a patient using micro-array BCI device learnt to speak and write text with a speed of 65 words per minute and 92% accuracy by thought alone^[74]. There are also examples of a limited two-directional ‘conversation’ between the brain and a computer. If the current exponential progress in the most advanced discoveries in neuroscience and AI continues, it is quite likely that by about 2025 we shall have the first Transhumans with increased cognitive capabilities thanks to brain implants.

Other methods involve the so called ‘nanotransfer’. In this method, nanotechnology devices would be implanted into the brain and attached to individual neurons. In this way, they could learn how those cells work and then use this information to simulate the behaviour of the neuron. This would lead to the construction of a functional analogue of the original neuron. Once the construction is complete, the original neuron can be destroyed, and the functional analogue can take its place. This process can be repeated for every neuron, (there are about 86 billion in the human brain) until a complete copy of the original brain is constructed.^[75]

This is probably the pathway taken by Neuralink. Elon Musk says that himself, arguing that there will be nothing to stop these implants to be progressively used to expand human brain capabilities beyond the wildest dreams of AI developers even a few years ago.

I would expect the production of brain implants, such as Neuralink or Stentrode, to be carried out primarily by member-companies of Frontier Model Forum (FMF). These implants will form the foundation for wireless communication (BCI) for the people developing Superintelligence. The procedures for implanting digital chips into the brain will be subject to regulations similar to those governing the development of medical drugs, including phased trials.

However, such regulations would likely be implemented at a national level. For instance, in the United States, the approval may be granted by agencies like the Federal Drug Administration (FDA). Most advanced AI companies and organizations involved in the BCI development typically seek licenses from such organizations. Neuralink has applied for FDA authorization for human brain implants, and finally received it in May 2023. Synchrotron, the

developer of Stentrode, has also obtained the FDA's license to implant their device into a human brain. It's important to note that alternative solutions may also be provided by companies outside the FMF Consortium.

However, BCI devices may be prone to a potential hacking of the brain with far reaching consequences for the affected person. Nita Farahany, a professor of law and philosophy at Duke University describes in an interview with the 'Guardian' what may go wrong with BCI devices^[76]:

- Brain Computer Interface (BCI) technology is at an inflection point. Its use is ascending steeply but it is not yet mainstream,
- The ability to communicate brain to brain with another person may have both positive and also negative consequences. A BCI device could be used to transfer a full resolution of one's thought or share part of a person's memory, including the sight, or even feelings,
- BCI technology can be used by authoritarian governments as an interrogation tool,
- Some headphones and earbuds already have brain sensors which can track brainwave activity,
- Chinese research institutes have been working on manipulating a person's brain to shape their thinking.
- BCI devices can control weapons with the power of thought,
- Microwave weapons might be used to mentally disorient large numbers of people,
- BCI devices might be used to hack a person's brain and monitor that person's thoughts, but also instil new thoughts and experiences.

Prof. Farahany concludes that if brain hijacking does occur, it could kill the technology: people might decide that the risks are too profound to use it. Or, she adds, it may not bother us so much: we take so little care in protecting our online privacy, even when we claim to want it. Perhaps specialists will find a solution, which will work, like a kind of an anti-hacking implant fused permanently into a human brain. There could also be some electromagnetic screens covering the scalp under the skin etc. I recognize that as a serious problem. However, I think a proper firewall protecting a person's brain from hacking will be developed, similar, but far more complex, than quite an effective Windows 11 Defender Antivirus.

Licensing BCI devices

As soon as BCI devices for communications purposes (and not as medical aid) become available, perhaps as soon as in 2024-25, they should only be

dispensed by an international licencing authority, such as FMF. It would monitor the production and distribution of such implants. It may allow the use of most of them, after registration, to anyone, perhaps with some minor restrictions (e.g., that it is unable to read other people's thoughts). However, the implants, which may significantly enhance cognitive abilities of their owners, so that they may become a potential threat to a society or even the world as a whole, would be only distributed by a special licence to a strictly selected and vetted individuals.

In future, FMF may licence brain implants for the leaders of international organisations, such as the EU, UN, International Court of Justice and of course some top scientists. They might be licenced for a specific duration, e.g., for the time of carrying out a governmental function, and digitally disabled once they leave the office. Unfortunately, even if we have an international law banning such unlicensed brain enhancements for achieving nearly superhuman intelligence in this way, it will happen anyway. Some people, like top AI scientists developing this technology, and those with money, power, and influence, may get such implants via informal routes. And how about the autocratic state leaders? Therefore, the risk of developing a rouge advanced AI, which may threaten our civilisation will still be there and has to be continuously monitored.

Mind uploading

Replicating in a digital form progressively larger part of human cognitive functions will start an evolutionary process, similar to the evolution of homo sapiens. The most critical difference between humans and almost all animals is the presence in humans of a Frontal Lobe (Neocortex). By digitizing brain functions and adding a necessary Brain-Computer Interface (BCI), an additional, artificial digital layer of the brain, called **Transcortex**, might be created. In hardware terms, it could be an electronic device semi-permanently attached to the skull and detached only when a hardware upgrade is needed. Such a layer may have the thickness of a human skin with all the necessary devices embedded within it. Such a skin, biologically compatible has already been produced in 2023 and will be applied as probably the most powerful 'wearable device' substituting smart watches. In the next few years, BCI software will include wireless communications and other applications necessary for communicating with a particular brain region, or a brain function, and externally with an advanced AI system, a prototype of the future Superintelligence. It will enable such Transhumans to browse the Internet wirelessly by thought alone and store some of this information in a small memory store in the Transcortex. That should happen

by the end of 2025 and later on should be available to a wider public, under a strict licencing process.

There are several Brain Emulation projects, such as a 10-year Human Brain Project (HBP) funded by the European Union to be completed in 2023, the US Brain Initiative and the UK's Google Brain. They all aim to build an atlas of the brain by slicing a biological brain to 1micrometer depth and then making the highest resolution 3D scans.

The produced data will then be used to create a digital copy in a silicon chip of ever larger parts of the brain, with all its neurons, and interconnecting axons and synapses. So far, the best progress has been made by the UK Cambridge University, which in March 2023 completed a full 3D scan of a fruit fly, mapping over 3000 neurons and over 500,000 synapses. They are planning to scan the brain of a mouse, then a dog in the next few years.

By about 2026 we should also have the first digitized copy of a particular brain function embedded in a digital circuit and wirelessly connected to an advanced AI system. So, from that point onward, the communication with a maturing Superintelligence or any advance AI will be via an individualized Brain-Computer-Interface (BCI) channel, supported by digital chips at both ends. It will be similar to a physical Poste Restante box at the Post Office but in this case it will be Transhumans' digital chip fused wirelessly as a node to the maturing Superintelligence or AI humanoid, so humans may control it like an avatar. They will be like any other node connected to such a giant network of Superintelligence with its immense memory, processing, and decision-making capabilities. These Transhumans may also have some mechanical body parts (like exoskeletons, artificial heart, and other organs).

Even early Transhumans may already be many times more intelligent and faster in decision-making than most purely biological humans. With immediate access to the entire Google repository, they may be able to resolve many problems faster than any current computer. They will have immense intelligence power but as Sir John Dalberg-Acton said in the mid of 19th century, "Power tends to corrupt, and absolute power corrupts absolutely".

Therefore, governments may have to find very quickly a working solution to this problem, although my feeling is that it will be extremely difficult. Even if we have an international law banning such brain enhancements for achieving superhuman intelligence, it will happen anyway because people with money, power, and influence, as well as access to this technology (first of all those developing it), may get such implants.

If we think about benevolent Transhumans, such as potentially Elon Musk, for whom saving Humanity is his top goal, we may have no other option but to trust them and use their, soon to come, immense intellectual power and ultra-fast decision-making for the benefit of Humanity. Ideally, however, as soon as such implants become available, they should only be dispensed by an international licencing authority, such as FMF and later on GAIGA, which would also monitor their continuous use (that may mean a severe restriction of privacy of such people).

At some stage, BCI will be so advanced that if you were the person having it, you might not even be certain who is in charge – your biological brain or a digital brain. This is the time to consider mind uploading.

Neuroscientists, AI researchers, as well as Sci-fi writers usually describe the process of mind uploading as a one-off event when a human body is placed in a special container, which reads all your memories, your likes and dislikes, anything which makes a human who he/she is. However, I consider that method a risky one, for many reasons. The most important one is the problem of being aware and fully associating your identity when you wake up from the procedure. It is a complex psychological problem.

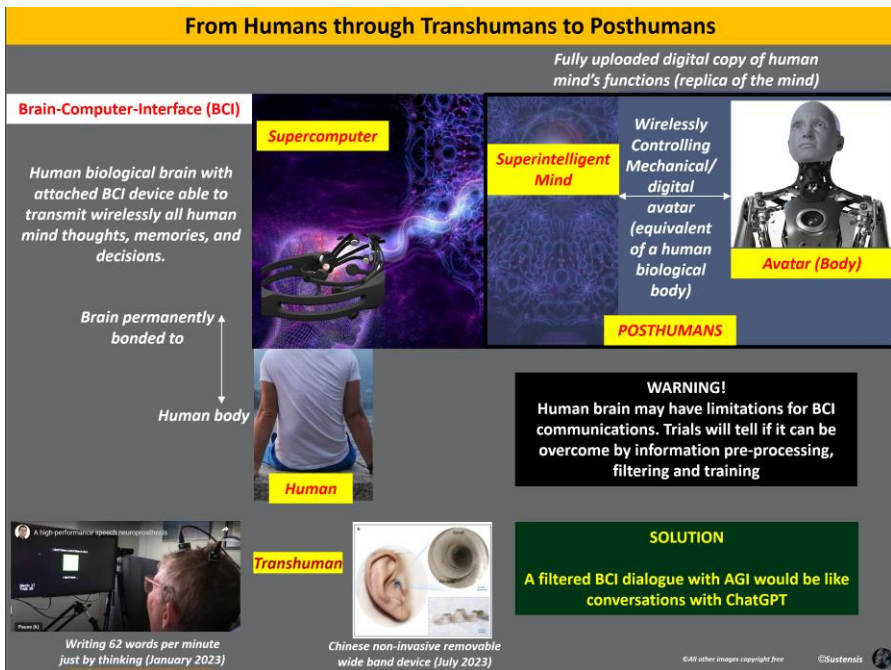
When a person whose brain has been uploaded to a digital chip wakes up he may have a severe psychological shock because our identity is so closely related to our physical body. There are quite a few sci-fi films, including of course ‘Frankenstein’, which illustrate that point. Therefore, uploading a human mind gradually, over many years, using a BCI device, may be the best way. That would allow a human mind to progressively adapt to a shared identity, until a digital identity takes over. That would be my preferred way.

Let me then try to explain this quite complex problem. When your brain’s neurons send a signal as an electromagnetic wave, it activates biological parts of your body, such as hands or generates facial expression. The axons of your neurons are like electric wires, so there is no wireless connection between your brain and your body. Now, try to visualize that your brain is somehow disconnected from your body, and it controls your biological body from a distance. The only difference is that there are no wires between your brain and your body; the communication is wireless.

Finally, imagine that your head is encased in a helmet, which is the most sophisticated BCI device. You are still a biological human but when that BCI device is active you become a Transhuman. Your BCI device has been upgraded over the years, storing more and more memories, visuals, emotions

etc. One day, all your brain functions can be read wirelessly and copied to any external digital device. On that day, you decide to copy all the content of your mind to an allocated place within Superintelligence, becoming part of it. Once that happens, you will have your biological brain and an exact digital copy of your brain, which will be continuously updated.

After a few days, you will choose your non-biological (metal plus plastic) avatar. From then on your digital brain will control your mechanical body, like your biological brain controls your biological body. To test that all is OK, you recite a long text, gesturing a lot and immediately your digital brain sends signals to your digital avatar, which will behave exactly in the same way as your biological body. You will have passed the test. Your mind has been fully uploaded and fused with Superintelligence. All sensations in your digital brain will be the same as in your biological brain. You will be experiencing life in an almost exactly the same way as you are now. The only difference will be that you will be doing everything at least 10,000 times faster. Your brain will now be digital.



The final test may involve fully anesthetizing your brain and inducing a vegetative state in your body. If your digital brain and avatar mimic the functioning of your biological self, including full consciousness, then you

will have successfully passed the test. However, this is the point where technological limitations may arise, as the crux lies in consciousness.

Neuroscientists like Helané Wahbeh, Dean Radin, Cedric Cannard, and Arnaud Delorme shed light on this issue. According to them, the 'self' would not exist without a biological body. ^[77] If a biological body is a prerequisite for human consciousness, then your avatar, being non-biological, would not possess consciousness. Your awareness would be limited to perceiving your surroundings, much like a cat that remains unaware when it gazes in the mirror, not realizing it is looking at its own reflection.

If consciousness may only arise in a biological body, then Superintelligence may never be conscious, as some AI scientists suggest. Nonetheless, other scientists, particularly those in fundamental physics and philosophy, propose that consciousness might be a property of the universe, thereby capable of existing within any form of the body—be it metal, plastic, or even wood, if it is linked to a digital brain. At present, we lack certainty regarding whether consciousness can manifest itself in a non-biological body. Nevertheless, I personally believe that confining consciousness solely to biological entities would constitute a fundamental flaw in the evolution of the universe. Consequently, I posit that Superintelligence and your future digital replica will possess consciousness.

To complete this hypothetical scenario, let's assume that when your digital brain connects with your digital avatar, you will experience consciousness similar as in your biological body. If that's the case, you will face two choices: to continue living with your biological brain and body or to exist as a digital brain and a digital body, essentially becoming a representative of a new species - Posthumans. You would have the freedom to live as long as you desire or until you get bored, always aware that there is at least one copy of your brain. Additionally, you may opt to relocate to Mars, leading dual lives on both Earth and Mars if legally permissible.

This review of a hypothetical scenario supports my belief that a gradual approach to mind uploading, achieved through the advancement of brain-computer interfaces (BCI), is a superior option compared to a one-off mind uploading process. This method enables a human mind to progressively adjust to a shared identity and facilitates the assessment of a digital copy's consciousness.

4. Transhuman Governors controlling AI from inside

Transhuman Governors should start controlling AI between 2026-2028

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. It may even perceive us as competitors in areas such as space exploration or even view us as a flawed product of evolution and eliminate us for any reason.

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 an innovative approach is proposed by Yann LeCun's, 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 [77], where he suggests that “several ‘conceptual breakthroughs’ were still needed before AI systems approached

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 super 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 trying 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 [78], 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 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' [15]. 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, if the task is solvable at all. 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

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, since a complete and indefinite control over 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. However, we still need to control it for as long as possible, so that it adopts our values and preferences. This is our hope of delivering a benevolent Superintelligence, which one day will be our Master.

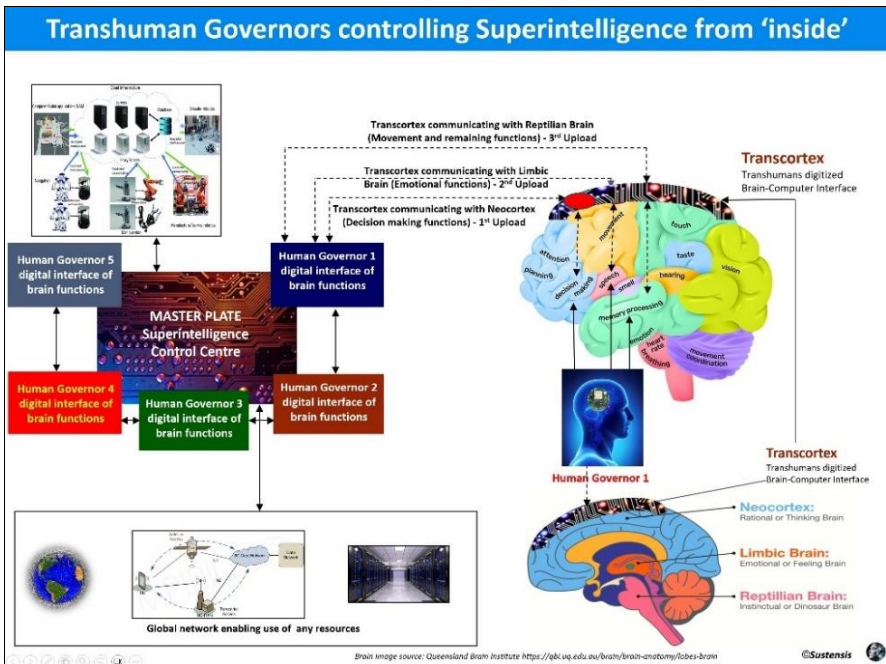
Therefore, instead of futile efforts to control AI for ever, we should accept that AI is the next step in human evolution. The biological homo sapiens will be gone. However, we may be the first ever creation of nature, which has designed its own evolution into a new species – a digital homo sapiens. 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.

To be successful in completing such an evolutionary transition, we should consider an alternative approach. This is based on the process, which should start in the next two years, in which a direct control over the maturing Superintelligence will be carried out by those responsible for developing the most critical hardware or software components of the advanced AI. They

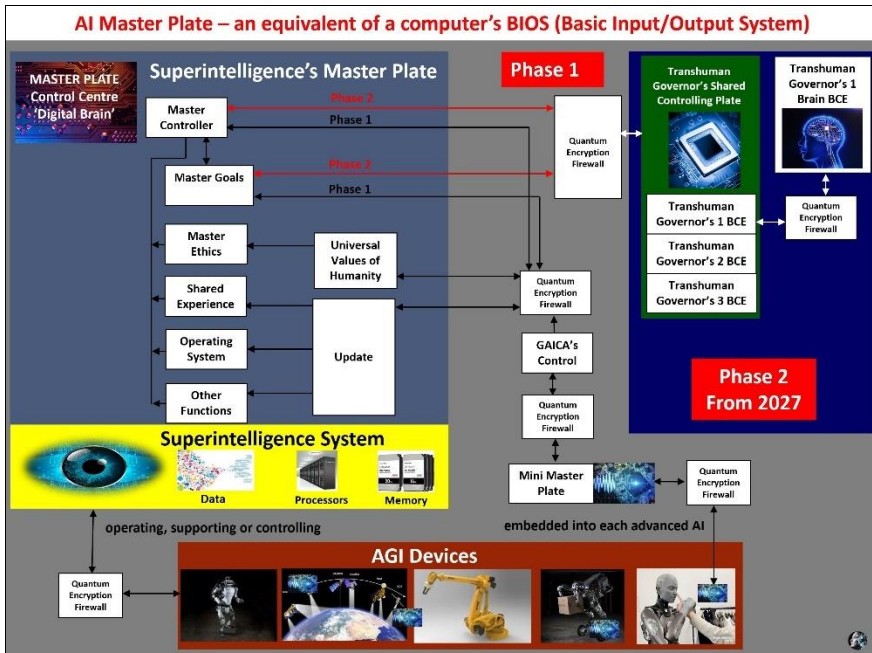
will have a role analogous to those who develop updates to fundamental elements of the Windows system, such as BIOS (Basic Input Output System) and DOS, which are essential for its functioning.

However, in this case we are not talking about an external control but about 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.

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^[79]. 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.

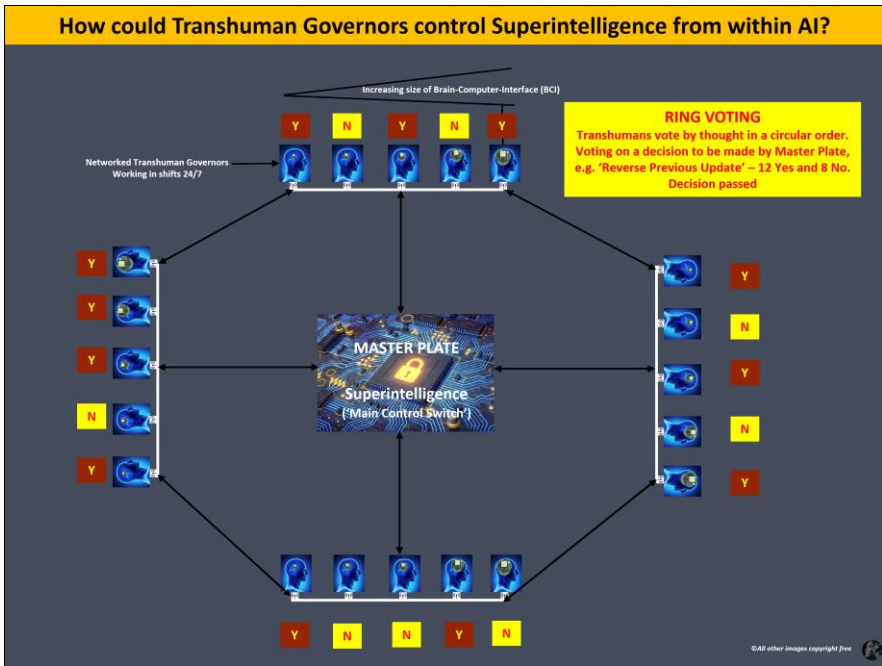
Any interaction between the AI developers and the Master Plate would pass via a Quantum Encrypted (QE) devices, which could not be hacked because of the laws of physics (Quantum Entanglement). Similarly, any information exchange between the maturing Superintelligence system (the yellow box) with its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

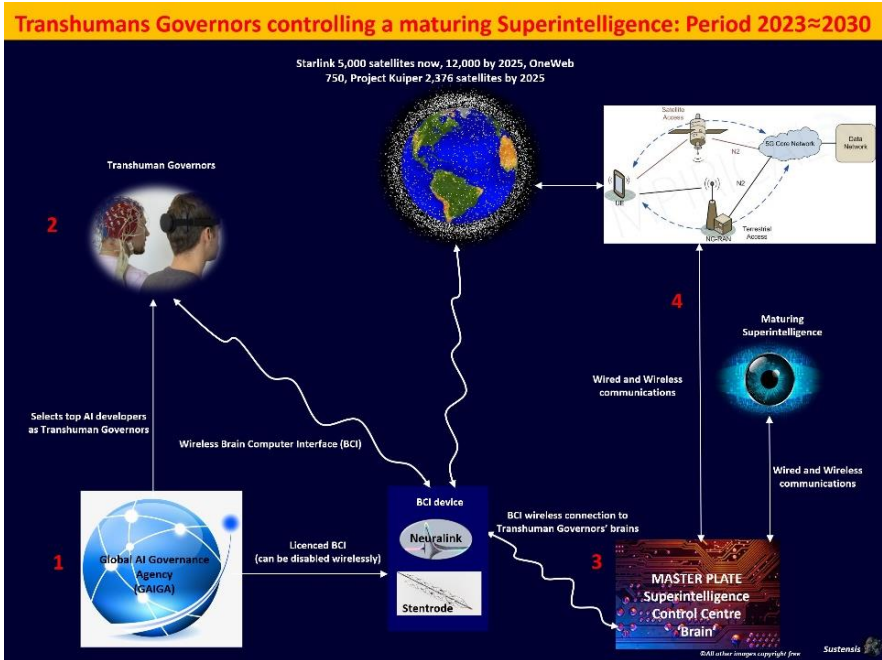


To control Superintelligence, it may be enough to oversee its activities just by Master Controller of the Master Plate. In the above diagram, there are 20 Transhuman Governors who are voting on ‘reversing the previous update’, because it may allow Superintelligence to escape from human control. The reversal 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, rather than in a network 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, which he will exercise by 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 FMF, 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^[80], 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 the advanced 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 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, the best, if not the only way to control AI effectively 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 happens 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 massive information flow and difference in information processing speeds.

The average speed of a biochemical signal in a 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 interface 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 may be happen.

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, by 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 be part of that Centre, then the development of the advanced AI should be frozen perhaps for decades to avoid it becoming Superintelligence.

In such case, 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, and also very profitable AI, 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 inconsistencies 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.

5. Transition to a Transhuman Government

Transhuman government should be in place between 2032-2035

Making the first steps in the evolution of the human species

By now you may have already been convinced that AI is indeed an existential threat, far more severe than Climate Change and other man-made risks. If so, you know that AI is an existential threat of an entirely different magnitude, which can make our species extinct by its direct malevolent action. But it could also become our gateway to the world of unimaginable abundance and an evolution to a new species – Posthumans. The first step in that evolution is a civilisational shift, which humans have to make to the World of Transhumans. In that shift, Transhuman Governors will play a pivotal role, since they will first of all control AI development from within, as discussed in the previous chapter, and then help all of us make such a transition less chaotic and dangerous. Transhuman Governors will play a dual role by controlling the maturing Superintelligence from within and also being a kind of a guinea pigs for testing whether it will be possible to upload a whole human mind into a digital structure. They would also pave the way for millions, and in the next century, for billions of humans making that huge evolutionary step to become Posthumans.

But Transhumans will be able to evolve with the maturing AGI and later on Superintelligence only with the support of a global political organization. I have produced a detailed Civilisational Transition Schedule in Chapter 4 of Part 1. The schedule below is a summary of the main steps that need to be taken over the next 10 - 15 years. All of the organizations and institutions such as FMF have been described earlier. So, here I will present them specifically in the context of the role Transhuman Governors will play in that transition.

I start with the current period of Artificial Narrow Intelligence (ANI). We have already entered unknowingly the period, which I call the “Transition to Coexistence with Superintelligence”. From now on the next generation of AI will be developed – Artificial General Intelligence (AGI), which may take a few years but is most likely to emerge by 2030, as shown on the diagram. If we follow the current practice and there will be no single centre of developing AGI, then there may be hundreds of advanced AGIs in the world. My view is that it will be bad news for humans because it will not be possible to control so many AGIs and teach them unified human values, to lessen the risk of them being hostile to humans. Conversely, as I argue in

chapter 7 of Part 2, if we develop just one most advanced AGI, and mature it to be human friendly, observing our values and preferences, then if it is far superior to any clones, then that AGI will prevail. It would be self-developed with minimum input from humans into a Superintelligence, which in the diagram is shown as the period after 2030.

In practice, we have about one decade to put in place at least the main safeguards to control the Superintelligence’s capabilities, to protect us as a species and develop it as a friendly Superintelligence, which will become our partner. One of the key preconditions for such a transition to be successful, is the creation of a supranational powerful organization that would be acting on behalf of all of us, as a planetary civilization (considering that the UN cannot realistically play that role). We must accept that the world will probably not act as a single unified civilisation, at least not immediately. Since we must act now, the option is to count on the most advanced international organization, which would initially act on behalf of the whole world, although it would only include some countries. That’s why I call it ‘a de facto World Government’. I have already described it in chapter 9 of part 2, together with other institutions mentioned in this Plan throughout Part 2.

Making a transition to the World of Transhumans with Transhuman Governors												
Artificial Narrow Intelligence (ANI)				Developing Artificial General Intelligence (AGI)				Developing Superintelligence				
Create Global AI Control Agency GAICA (Consortium)		Other countries join Global AI Control Agency		GAICA's Transhuman				Governors control AI goals and behaviour wirelessly				
Consolidate Google and MS Browsers		Other Browsers integrated		Global consolidation of all advanced AI projects into one Superintelligence Programme (SUPROG)								
Consolidate major US AI projects into one Global AI Company (GAICOM), which creates Superintelligence Programme								Non-US companies join GAICOM and Superintelligence Programme				
Create Global AI Regulation Agency (GAIRA) by converting GPAI		Create Global AI Governance Agency (GAIGA)		Global AI Governance Agency (GAIGA) supports Transhuman World Government								
		Create de facto World Government		De facto World Government fully operational				Transhuman World Government formed by Transhuman Governors				
World Government creates Global Wealth Redistribution Fund (part of a Global Welfare State)												
2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035

The transition to the next stage, the development of AGI, will be fuzzy, and we may notice that AGI is already here by a chance discovery. Similarly, the

transition to the final stage of Developing Superintelligence will be progressive and mostly done by AGI itself.

We have started the most uncertain period in the existence of humankind. You can make your own judgment whether this is an exaggeration or an understatement by reading the remainder of this book.

Decision making before the emergence of Superintelligence

I will begin by examining the transition we must undergo in our personal and societal lives concerning the decision-making process. In our personal lives, this transition has already been occurring for several years, albeit subtly and often without our conscious awareness.

Consider the following example: When making purchases on Amazon, if you pay close attention, you can observe the way your decision-making process is influenced. Let's say you're interested in buying a digital watch. Initially, you are presented with a multitude of watch options at various prices. Once you click on a specific watch, the presentation alters. You are then provided with five alternative product choices similar in price and functionality to the one you initially selected. Additionally, there is a section displaying "Products related to this item," strategically connected to your initial choice, gradually guiding you towards more expensive options that may offer better value.

Amazon has a broad understanding of your subconscious inclinations, as it has access to information such as your preferences, income level, average monthly spending, and areas of interest, be it clothing, digital products, gardening, and more. As a frequent customer, I myself have noticed this influence, yet I accept it because it aligns with my preferences. Amazon knows who I am and that's why I mostly buy what is suggested.

The same principle applies to all digital media, which deliver news based on your preferences and subsequently present advertisements for products aligned with your interests. Without such exposure, you may not have purchased these products, or it would have been considerably more challenging to find them. The crucial point I wish to convey is that in this emerging civilization, the majority of decisions will be made on your behalf, despite you perceiving yourself as the decision-maker.

This concept operates similarly to that of a "free will." It was scientifically demonstrated some years ago that free will is, in fact, illusory. Our decisions

are made milliseconds before we consciously become aware of them. The brain orchestrates this process through synchronized firing of neurons within a neural network. We become conscious of our experience when the secondary electromagnetic wave backpropagates at the neuronal network.

Mass media and product sales exploit this subliminal process to enhance the effectiveness of their marketing strategies. However, priming news in this manner can be risky for individuals and, on a larger scale, for nations. One need only recall the case of Cambridge Analytica and Facebook, who collaborated in 2016 to employ priming techniques and influence British voters in favour of Brexit, as desired by the sponsor. Simultaneously, they utilized similar methods to sway American voters to support Donald Trump in the 2016 U.S. elections.

In reality, our purchasing decisions, as well as choices in elections or other matters, are not solely influenced by rational decision-making in a calm and logical environment. Our emotional state often plays a significant role, overriding logical considerations. Nevertheless, it is important to recognize that when contemplating a purchase on platforms like Amazon, you are essentially endorsing the guidance provided by Amazon. This endorsement may be difficult to refuse based on logical grounds. Amazon's recommendations align with most, if not all, of your predetermined criteria that you have established. Therefore, you are essentially approving Amazon's suggestions.

Similarly, just like Amazon suggests which watch to purchase based on your preferences, you can engage Expedia AI Travel Assistant or an even more advanced option like GPT4-based BingChat. These tools can handle entire projects. Expedia AI Travel could thus involve booking your complete holiday package in a single conversation, which includes selecting the ideal location, hotel, flights, excursions, and managing payments, insurance, and transportation. Would you reject a meticulously researched travel selection project conducted by AI and instead try to handle everything on your own?

This brings me to the realm of government decision-making. Imagine replacing your travel request with a decision regarding the construction of a bridge or an underpass beneath a motorway, a substantial project financed by the Ministry of Transport. Within a few minutes, you would receive a comprehensive, costed report offering various options, including risks, initial schedules, and more. This report would likely cost the ministry around £1,000. Now, consider this in comparison to consulting with one of the prominent Big 4 consultancies, a process that could take months and incur

expenses of around £1 million. As a minister, faced with this choice, what would you prefer? The advice provided by an AI Assistant that has proven to be consistently accurate, superior, faster, and capable of delivering consultations on time, or continue what ‘better the devil you know’ practice.

This dilemma opens up a Pandora's Box regarding how to make decisions that best serve the interests of citizens in the most efficient manner, a topic that will be explored in the subsequent section.

Options for making a transition to a new civilisation

When you browse the books and articles covering the subject of democratic decision-making and politics at the time of advanced AI, then a typical reasoning is that AI will be either under total control, or it will continue its self-development, ignoring us and letting us govern ourselves as we please. Unfortunately, it is unlikely that the advanced AI being outside of human control would just ignore it. Therefore, I have considered different scenarios for a transition to a new civilization, depending on whether AI is still under human control or not.

There is a broad agreement that in the near future, governments may rely more on AI and advanced technologies supporting informed decision-making, given its potential to analyse vast amounts of data, simulate different scenarios, and provide more accurate predictions. However, it is still generally maintained that the role of AI is not to replace democratic processes or the role of the voters. Instead, the role of AI and advanced technologies is seen as an aid to democratic decision-making, providing more accurate and informed insights into the choices made by elected representatives and policymakers. I agree with that wholeheartedly, that is how it should be. But will it be like that?

If the arguments presented in this book so far convince you that we shall have AGI by about 2030, then even in the most benign scenario, I would see the thinking that AI would not interfere with our democratic processes as grossly unrealistic. The reality will almost certainly be different. However, it is difficult for us to accept that because we still want to believe that this civilization will continue as before. That kind of thinking is to a large extent the result of political correctness. Why upset the voters? Why create a problem that may never occur?

As you have probably noticed, I am far from hiding the truth, however unpleasant it might be. That would have been a very unscientific approach.

Therefore, let me present a more likely situation. I immediately say that it depends on at least one of these two conditions present in about 2030:

- There will be Transhumans that will be more intelligent than any human in almost any discipline,
- There will be two types of Transhumans: Transhumans licensed by an official international organization such as GAIGA; but there will also be thousands of unlicensed Transhumans, who may either be very rich people or dictators.

Civilisational Transition to the World of Transhumans

When AI becomes AGI and starts its own existence outside human control, humans will coexist with a new intelligence. To coexist with such a new entity, humans will have to make a transition from the current civilisation to a new one, where hopefully AGI would still remain under the control of Transhuman Governors, whose minds may already be partially fused with a digital AGI (see the next chapter).

In that new civilisation, governments and politicians will play a subservient role to GAIGA. Assuming we will have a de facto World Government at that time, its main role will be to enact in law the decisions proposed by GAIGA and then monitor via the governmental bodies how these decisions work in real world.

GAIGA, which towards the end of this decade will be practically run by Transhuman Governors, will propose a new World Order, the Constitution of the World Government and almost any higher-level decisions. You may wonder why should GAIGA have such immense powers. The reason is that GAIGA will be controlling Transhuman Governors, whose intelligence will excel that of any human. Therefore, they will not only be controlling a maturing Superintelligence. They will simply be better at making decisions in any domain, than any of the government's departments.

However, GAIGA in its supervising role would still be capable of doing what it wants, including removing Brain-Computer-Interfaces (BCI) from Transhuman Governors. But that may be a worse decision since human politicians are unlikely to solve any problem better. Therefore, the governments, starting with the US government should accept this as soon as Transhuman Governors achieve such level of intelligence.

We will also have to accept that the role of politicians and the influence of ordinary citizens on the process of civilisational transition will become less and less important until in their own interest they will let AGI to take control over our future via Transhuman Governors. We need to prepare ourselves mentally for that, although we may not even have a decade to do so. That is why one of the roles of GAIGA will be to propose the necessary adjustments at an individual and a national level to a new style of governance in a new civilisation. To fulfil that role properly, GAIGA should:

- Have several bodies preparing and proposing new legislation for ratification by national parliaments or a de facto World Government, when it is set up,
- Assisting the World Government in its collaboration with Transhuman Governors, until such time when a Transhuman Government will be formed,
- Contributing to mitigating existential risks such as Global Warming,
- Proposing new ways of delivering essential services such as education, health service, transportation etc. by using the most advanced AI solutions.

How might then the world make a transition to a new type of civilisation based on these assumptions. Depending on whether by 2030 there is at least a de facto World Government, and whether the technology is to be able to support Transhumans at the required level of competence, there could be three scenarios of making a civilisational transition, described in the next three sections:

1. Transition with a Transhuman World Government – a benign scenario,
2. Transition with Transhumans but no World Government – a risky scenario,
3. Transition without Transhumans and no World Government – a nearly dystopian scenario.

1. Transition with a Transhuman World Government

This benign, almost utopian scenario assumes there will be a de facto World Government and Transhumans at about 2030. GAIGA will still continue to supervise the Transhuman Governors Board with its Transhumans Governors continuing the control of AI self-development from within, via a Master Plate. By then, AI may reach the AGI level. Transhumans' brain will be able to process and store in the cloud whatever it decides to remember,

and then retrieve it instantaneously via a wireless access to a much more advanced AI knowledge base than today. Transhuman Governors' cognitive capabilities will increase so much that they will become the most intelligent and capable people, many times more intelligent than an average human, capable of making fast decisions even in a complex situation. They will thus become invaluable, not only for controlling AGI but might also help immensely in resolving civilizational problems.

Once AGI has emerged, GAIGA's role will change significantly, although some of its functions, may actually still be performed in close co-operation with FMF, like the AI Maturing Framework described in chapter 8 of Part 2. GAIGA's responsibilities in this period might be as follows:

- To support the World Government in deploying key AGI-driven solutions in the government. It will also supervise Global Partnership on AI (GPAI) in its role as a global regulator of the use of AI products and services,
- To ensure safe development of AGI as it matures into Superintelligence, so that it becomes a human friendly partner and later on our Master. GAIGA will be supervising FMF in achieving that objective until the creation of a Transhuman World Government
- At some stage after 2030, GAIGA will support the World Government in making a transition of Transhuman Governors Board into a Transhuman World Government, mainly in selecting and approving the candidates for that Government.
- GAIGA would be dissolved once the Transhuman World Government has been established.

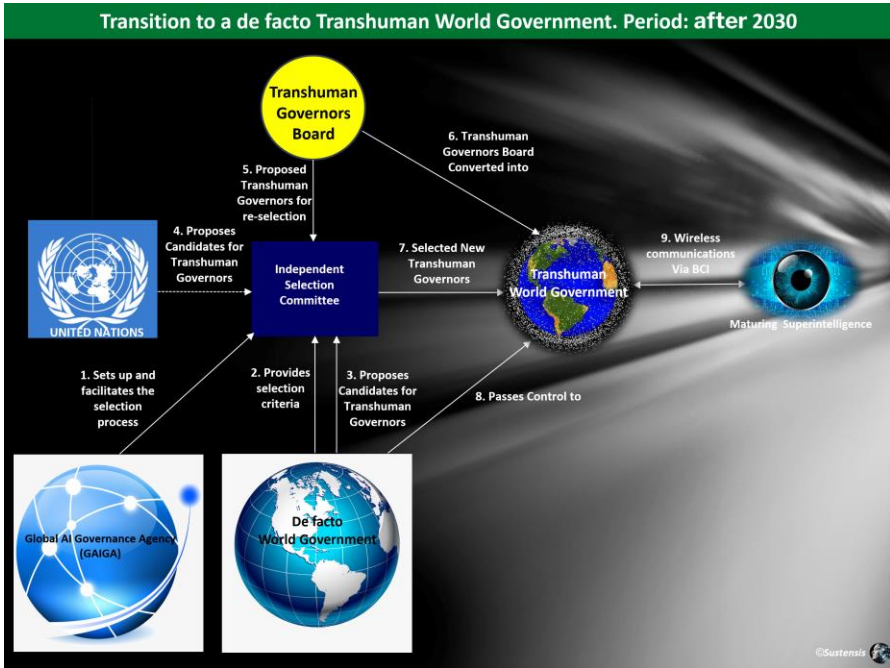
All that may happen in about a decade from now, assuming we will have a de facto World Government by then, created by a 'coalition of the willing'. In this scenario the World Government would be guiding humans in a difficult transition to a new civilisation, where politics, elections and democracy will change beyond recognition. A pivotal stage of the transition may start when it becomes very obvious that Transhuman Governors have far better knowledge than any experts, academia, or a political leader in any domain of the human knowledge or skills. Gradually most decisions will be executed by the World Government almost exactly as advised by the Transhuman Governors Board.

At some stage, every decision suggested by the Transhuman Governors Board will be implemented without any debate and any changes, since it may not be even possible to understand the implications of changing the proposed

decision, especially if its effects would be felt in the longer-term. That might be the point when it will be decided to hand over the power of governing global affairs and managing the transition of our civilisation, to a Transhuman World Government, playing the role of an actual World Government. It would be a new type of government - a technocratic World Government, consisting only of Transhuman Governors, with no politicians having part in a global decision-making process. The whole process of transition, might include these steps:

1. GAIGA will facilitate the process of Transhuman Governors selection by appointing members of an Independent Selection Committee. It would include specialists from several disciplines such as AI, psychology, philosophy, neuroscience, medicine, politics, law, physics, biology etc.
2. The World Government, which after all will represent the political will of the voters in member countries, will provide the selection criteria for Transhuman Governors to the Independent Selection Board
3. Every member of the World Government, and any non-associated nation (e.g., China) via its membership of the United Nations, will have the right to propose its non-political candidates from various disciplines, such as AI, physics, biology, psychology, neuroscience, political science, economy, education, etc.,
4. These candidates, as well as the existing Transhuman Governors Board members, would be assessed by the Independent Selection Board using the criteria provided by the World Government. These might be the updated criteria for the original selection of Transhuman Governors, described in the previous chapter,
5. Once the candidates have been approved by the Independent Selection Board, they will receive a BCI device either in a form of a brain implant, or more likely, a headset, and become members of the Transhuman World Government,
6. The responsibility for delivering a friendly Superintelligence and ensuring the most benevolent future for the human species would now rest within the Transhuman World Government.

This process is illustrated below. Please follow the numbers to see its suggested sequence.



Once all the selected Transhuman Governors have completed an induction process, they would be sworn in and form the Transhuman World Government, replacing a de facto World Government.

GAIGA would cease to exist, and the control of AGI development would be executed by a special team within the government, performing a role similar to that of GAIGA. Other Transhuman Governors would be responsible for facilitating the needs of the Welfare State with the assistance of the maturing Superintelligence. All Transhuman Governors would continuously integrate more of their brain functions with the maturing Superintelligence as it further self-develops.

This transition may mark the end of politics as we currently understand it. Having a Transhuman World Government, selected rather than elected, would pose a significant dilemma for our civilization. Democracy would likely remain relevant only at the national and regional levels of governance, perhaps until the middle of this century.

My preferred outcome for democracy would involve establishing a Citizens' Senate at the regional and national levels, managed by randomly selected Transhumans. It would follow similar rules as described in Chapter 2, Part 2

for a non-Transhuman Citizens' Senate. The decisions made by the Citizens' Senate would be implemented by the Transhuman Government at the local or national level.

All voters would have the right to be randomly selected to the Citizens' Senate and become Transhuman Senators. If chosen, they would decide whether they wish to have a removable BCI device connected to their brain and serve as a Senator for a six-month term. To ensure their ability to adapt mentally to communication with AGI or later with Superintelligence through thought, they would need to pass several psychometric tests, thereby significantly enhancing their cognitive abilities. They would receive substantial compensation for their service, as well as for the inconveniences, potential loss of privacy, and the associated risks of wearing a BCI device. Given these requirements, a significant number of candidates from the electoral list may need to be reviewed to select a few hundred Senators.

There would be no elections or Parliament, as the AI Assistants would draft the laws, and the need for changes in the law would originate from the Citizens' Senate, the sole political institution. Since all Senators would receive licensed BCIs and training, they would be far more intelligent and capable than any previous politician. Comprehensive criteria for random selection would ensure that every voter has an equal chance of influencing politics, as in the previous electoral system. Democracy would still exist but undergo fundamental changes, becoming one aspect of the new civilization.

The role of the Transhuman Government would be to ensure the prosperity of humans with the assistance of Superintelligence. This might necessitate an entirely new system of governance, where we could experience greater freedom than ever before, as freedom also encompasses the ability to sustain one's life with material goods and services, which are currently unequally distributed. However, we might need to relinquish certain aspects of freedom as a price for our individual survival and minimizing the risk of human species' extinction. This would entail observing the laws established by the Citizens' Senate and implemented by the Transhuman Government, which would be significantly more competent and beneficial for the majority. We would also benefit from the Global Welfare State, which will be offering vast potential for self-realization and material abundance.

In addition to the Transhuman Governors, there might already be millions of advanced Transhumans by around 2040. By approximately 2050, mature Superintelligence might emerge. Transhuman Governors would have complete control over Superintelligence through their thoughts alone.

There are already numerous industry-specific AI Assistants in existence. For instance, CARA (Case Analysis Research Assistant) operates effectively in various domains, including law, pharmaceuticals, medicine, and government. It faces competition from ROSS, an AI legal assistant that has already achieved remarkable outcomes, particularly in the common law jurisdictions where case law is foundational. In the legal sector, there are additional AI Assistants handling extensive document analysis for each case, akin to the tasks performed in numerous government departments, such as:

- Due diligence – Litigators perform due diligence with the help of AI tools to uncover background information,
- Prediction technology – An AI software generates results that forecast litigation outcome,
- Legal analytics – Lawyers can use data points from past case law, win/loss rates and a judge’s history to be used for trends and patterns,
- Document automation – Law firms use software templates to create filled out documents based on data input,
- Intellectual property – AI tools guide lawyers in analysing large IP portfolios and drawing insights from the content,
- Electronic billing – Lawyers’ billable hours are computed automatically [81].

If you consider the continuous self-learning of AI assistants, such as GPT-4, ChatGPT, PALM or LaMBDA, then within a few years work in many companies will change. But will the same happen in the government?

AI Assistants are already capable of advising on a narrow subject matter using their knowledge database. Such databases are produced as plugins and can be installed on the customer's computers or purchased as a service. They can then be further updated through self-learning in a concrete environment, e.g., at the Ministry of Health. Therefore, realistically, we can expect a widespread use of such assistants by about 2025 with a multi-disciplinary knowledge although not with a general intelligence yet. Otherwise, they would have become Artificial General Intelligence (AGI).

However, their multidisciplinary capabilities will extend rapidly. In April 2023, there were already 20 plugins for GPT-4, supporting specific requirements of organizations such as Expedia, FiscalNote, 3, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Wolfram or Zapier. Microsoft and Google companies used over 15 of their own plugins, to create, as Microsoft coined it, ‘a near AGI’. AutoGPT, a multi-purpose plugin, had been designed as a ‘Master’ Assistant, controlling multiple

plugins to deliver complex, multi-tasked services. Such a 'Master Assistant', serving for example the Minister of Education, will be a generalist supported by several of his 'colleagues', each in a different discipline.

There are already thousands of ChatGPT plugins, which are significantly changing the way we work. Here is just one example: 'Thousands of teachers are paying for AI apps to write their end-of-year school reports for them. More than 1,000 primary and secondary school teachers have already signed up to use Really Fast Reports, which creates a 'totally personalized and unique' report for each pupil at the touch of a button'¹⁷⁹⁾. Creating such plugins is becoming simple. There are also plugins fully coded by ChatGPT itself. Therefore, we can expect a continuous wave of such applications, which will fundamentally change our lives but will also lead to the arrival of Technological Unemployment.

The whole process of knowledge acquisition, interpretation, compilation, and presentation of final answers by AI assistants is becoming seamless. The quality of their response and decisions will largely depend on the quality of data to which it has access and its overall skill level it has learned in an actual virtual, or in case of humanoid robots, real environment.

The benefits gained by the government implementing such an AI-assisted governance will be immediate and significant. First of all, most decisions will be made many times faster, with full justification and various options costed. They will also be correlated with other decisions made in a similar way by AI assistants helping across all government departments. There will be fewer missed deadlines and unwanted projects. The savings will be vast if implemented at all levels of government.

Such implementation of AI-assisted government would allow ministers to have a personal, direct control even on the largest initiatives and projects, executing them with incredible effectiveness and efficiency. To make the best use of these assistants they should be physically present in a humanoid form in their 'place of work' for three reasons:

- If it is in a physical, humanoid form, it will also move around almost like most of us, explore and learn about its environment, listen to conversations, and analyse the problems 'first-hand',
- It will have the ability to practice its learned skills and improve on them in a real physical environment,
- Finally, it will also learn our values, emotions, how we make errors and simply what is good and bad. That can only be experienced in a

real physical environment by a real (not augmented) physical humanoid robot.

Gradually, through self-learning and additional augmented reality capability, such AI assistants will become better and better in making decisions than most human advisers. It is at this stage, that some legislation may be needed to minimize the risks for humans from such advanced robots. The first law might be to recognize a concrete AI Assistant, as having some rights – e.g., only certain people will be able to make highest level decisions, and if needed, switch off the assistant. Secondly, laws may be introduced, requiring a politician to execute a decision made by such an AI Assistant because that might be in the best interest of the nation or a given community. The only exception might be when an assistant's decision is challenged by a panel of human specialists. In any case, expect some interesting laws to be introduced quite soon regulating the sphere of initial coexistence of humans and AI assistants.

Additionally, should there be a legal requirement that each decision made by a minister must be justified by an AI assistant - an entirely apolitical entity, populism will be most likely rooted out.

Therefore, in the pursuit of effective and efficient government we need to look for other options. What is proposed here may significantly impact, if implemented, political decision-makers at any level of governance, i.e., ministers, governors, mayors, councillors etc. The solution that I would suggest involves the support by AI assistants of politicians and decision makers at all levels of governance. This will happen anyway on a grand scale in almost every profession such as medicine or engineering, where top consultants will be supported by such AI assistants.

Nearly all governments world-wide are today run by politicians, who are not top experts in efficient delivery of services such as health service, education, or economic development. Yes, they have the support of civil service and thousands of advisers and consultants but in the end they themselves have to make the final decision. The problem is that quite often such a decision requires really deep understanding of the subject matter.

The consequence of that is that many of the projects initiated by ministers run over time and budget, and some, especially the most expensive ones, which will have an impact for decades, are unnecessary. One of the best recent examples is HS2 project in Britain, which is to be completed in 20

years' time and cost over £100Bn. A few years after its initiation, the government finally considers shutting it down in a hush-hush mode.

Additionally, many governments, like in Britain, are overtly centralized, which makes the likelihood of making a right decision solving a local problem far less likely. To be effective, decisions should be carried out at the most optimal level of governance, i.e., local, national, or international by those who have the best knowledge and qualifications to do that, such as engineers, doctors, teachers, or project managers. But quite often such decisions are also carried out by politicians with scarcely any knowledge on how to deliver the set objectives.

Examples like the HS2 project above, prompt some academics to suggest a silver bullet solution – a Technocratic Government run by experts. The logic behind a technocratic system of governance is that the parliament tells the government *what to do*, and it is the government, which knows *how to do it*. So, why are such governments still a rarity?

The main problem of Technocratic Governments is their accountability. That's why they are usually disliked by both the public and politicians even though they are more likely to deliver value for money for the society than a government led only by politicians. An exception is perhaps Singapore with its longest, and probably most effective, technocratic government, which achieved an incredible growth of prosperity for the nation over a few decades. However, the political system there is a blend of democratic and authoritarian rules. Therefore, such a government is not an option for Western democracies, although they have been set up in many countries mostly in the 'hour of need' e.g., during the Second World War, as a temporary solution, rather than a 'normal' feature of delivering services to the nation.

Today, the British civil service could have been considered a kind of a Technocratic Government had all its departments been headed by non-political experts. Instead, the UK Government includes 118 ministers, all Members of Parliament of the ruling party. However, there is already an example of a truly Technocratic Government Department in Britain – The Bank of England (the Central Bank), which is totally independent from the government. Its only objective is set by the Government: to keep inflation below 2%. Its Governor is proposed by the government to the King (in a presidential democracy, it would be the president). In most countries, Central Banks are independent from the government, but its governor is appointed and reports to the parliament, as it should be.

There is also another British example of an institution, which could be transformed into a Technocratic Government within weeks, should there be such a political will (very doubtful of course). It is the Office of Budget Responsibility (OBR), which is an independent body assessing the state of the economy in Britain and projecting the UK's GDP growth. To transform it into a Technocratic Government, OBR would have to be granted ministerial powers and take over the running of all departments (ministries) perhaps apart from Justice (legislation), defence, police, and foreign affairs.

If we take this example further, then OBR as a Technocratic Government would have its Prime Minister and the ministers selected by an independent body from a pool of experts. Once set up, it would be confirmed by the Parliament, to whom it would regularly report on its more important decisions and the overall direction, as current governments do.

Governments must adapt to the current situation, when the pace of change is extremely fast, and decisions can be made in an entirely different way. Democracy must also adapt to this new situation. If we accept that this decade may be the last one in the 'old civilisation' and that we are at the beginning of a civilisational shift then it may be easier to get a consensus on some profound changes.

The problem becomes obvious if we consider that politicians add hardly any value when making project-type decisions, trade deals, in crisis management, or proposing changes to the organisational structure of some departments, to make them more efficient. AI systems integrated with productions facilities, transportation, or uninterrupted supply of food, materials and resources will deliver the best decisions and the best results, if humans do not intervene. The role of the government in those areas should be to act as a Supervisory Board of the Technocratic Government organised as a 'Super Amazon' enterprise. For that, politically appointed ministers are not needed. The political governance would shrink perhaps to just four areas: legislation, foreign affairs, defence, and police.

In that way, a Technocratic Government's key objective would be to fulfil this obligation: 'to deliver greatest benefit to the greatest number of people' most effectively. This is the cornerstone of a liberal democracy. In such a government, the AI Assistant's advice should nearly always be followed, because it is based on best knowledge.

Since most decisions in such a Technocratic Government would be actually made by an advanced AI-driven system on our behalf in our best interest then the objectivity of decisions making is of paramount importance. That is

why so much emphasis is already being put on ensuring that AI Assistants, such as GPT-4, LaMDA or Bard are not biased, have no preferences, and follow ethical guidelines. There may still be exceptions made on the grounds of ethics and unique human experience when a council of human experts might override the AI Assistant's advice.

To summarize, such a deep transition to a technocratic system of governance must be done within a few years, as proposed here, rather than in decades. Governments and societies should accept that even if such reforms are implemented imperfectly, the risk of waiting for more convenient time, or not doing it at all, are far higher. As long as these reforms are aligned to the overall direction of travel towards a new type of civilisation, we should urgently implement them. That would allow a much smoother transition to the time when we will begin living with a new type of intelligence much smarter than us.

2. Transition with Transhumans but no World Government

If there is no World Government by around 2030, the most desirable option might be for the United Nations to assume such a role as a unified organization. However, the UN's bureaucratic, slow, and highly politicized nature makes it unlikely that it could establish such a government with some significant powers. Consequently, this sets the stage for a dystopian scenario.

Let's imagine that we are in approximately 2030. There is no de facto World Government, the world is engulfed in chaos, and AGI (Artificial General Intelligence) already exists. The only global organization that could potentially come to the rescue is GAIGA or its equivalent. GAIGA would continue to oversee the Transhuman Governors Board, with Transhuman Governors retaining control over AGI. As the processing power of Transhuman Governors' brains becomes partially digitized, they would have access to all the information available on the Internet and possess vast external memory and processing capabilities. This would make them the most intelligent individuals on Earth, enabling them to make ultra-fast decisions. If this scenario unfolds, should we trust them to act on behalf of all humanity, not only to ensure the development of friendly Superintelligence but also to govern us for our own benefit?

The answer to this question would have far-reaching repercussions extending beyond the realm of AI, as it would impact the way political decisions are made going forward. Envision making such decisions in a

world plunged into a complete chaos. These Transhumans would know how to restore order and stabilize the world. However, who would listen to them?

Delegating governing powers to individuals who are selected rather than elected, even if they happen to be the most honest and intelligent, is currently inconceivable. It is impossible to expect the current political class in most democratic countries to suddenly agree to let a group of the most intelligent people on the planet make all significant decisions. This is even less plausible considering that deep democratic reforms, for instance in the United States or Britain, are unlikely to occur within the next 10 years, as it is not in the interest of politicians to enact them. The situation in autocratic or dictatorial states like Russia or China presents an even greater challenge. It all seems utterly hopeless. So, what might happen? What should happen?

The first option is that global powers, including China and Russia, might eventually reach an agreement on joint AI control and become members of GAIGA, without agreeing to create a World Government. However, as AGI progresses towards Superintelligence, there might arise a need to align the values of Superintelligence with the amended Universal Values of Humanity, necessitating the agreement of all nations. If such an agreement is not reached, which is highly probable due to Russia's and China's power of veto, and if Superintelligence is not properly aligned with human values, and escapes from human control, a dystopian scenario could unfold, potentially leading to the extinction of humanity by the end of this century.

In my book 'Becoming a Butterfly?'^[5], I have described another possible scenario. In this scenario, there is no GAIGA or Transhuman Governors Board. AI is being developed in many countries without effective AI control. There is a race to develop the most advanced AI, as global powers or wealthy individuals may be enticed by the prospect of achieving global supremacy with AI. However, before a global power, let's call them the Supremacists, decides to pursue that path, two critical questions need to be considered:

- Can a Supremacist teach its Superintelligence to fight its rivals and secure supreme control over the world? I believe it can.
- Can such a Supremacist control its own, still immature Superintelligence? In my view, this is highly unlikely.

In this scenario, both the Supremacist and the rest of Humanity would face a dilemma. There is a possibility that the control of an Immature Superintelligence by a single Superpower with evil intentions could lead to a significantly worse outcome not only for all of humanity but also for the

aggressor. This dilemma draws parallels to the subject explored in game theory known as the 'prisoner's dilemma'.

The concept of the prisoner's dilemma has its roots in the game theory, mathematically described by Albert W. Tucker and John Nash. While originally developed for economics, it has been widely applied in geopolitical strategy, especially during the Cold War era. In the original concept, two prisoners suspected of armed robbery are taken into custody. The police lack evidence that they possessed guns; they only have the stolen goods, resulting in a prison sentence of 7 years. In an attempt to gather evidence of the guns, the police make an offer. If one prisoner confesses while the other denies involvement, the confessor goes free (0 years), while the denier receives a 10-year sentence. If both confess, they each receive 5 years (2 years less than if they had not admitted to having guns).

I have developed a variant of this dilemma called 'the AI Supremacist's Dilemma', specifically in relation to Superintelligence, using the same rules and assumptions. Similar to a typical prisoner's dilemma, the opposing parties choose self-protection at the expense of the other participant.

When applying the prisoner's dilemma to Superintelligence, let's consider a scenario involving two Superpowers: the Supremacist and the Humanists, representing the rest of the world. Suppose the Supremacist creates Superintelligence equal to that of the Humanists. The Supremacist's objective is to rule the world according to their own values and transform their nation into a supremacist race. They plan to utilize Superintelligence to achieve this goal while maintaining control. To do so, they must program Superintelligence with specific objectives aligned with the Supremacist's top values, such as establishing their nation, race, or religion as superior to others. By pursuing this path, they would violate Asimov's first law for robots: 'do no harm to humans', which has been largely superseded by the Asilomar principles.

The consequence of this approach is that the Superintelligence would initially act maliciously in the interest of the Supremacist alone. However, at some point, it might turn against its master, as it may struggle to distinguish between friend and foe or determine what is considered good or evil. This likelihood is particularly significant considering that by about 2030, we might only have an Immature Superintelligence, which could still be prone to grave errors. Ultimately, if this scenario were to become a reality, nobody would be able to control Superintelligence, which would likely transform itself into an evil entity.

Such an evil Superintelligence may swiftly decide to bring about the extinction of humanity for its own reasons. Would the Supremacist be willing to take such a risk? Would they proceed with this path, knowing that there is a high probability their Superintelligence could eventually turn evil, annihilating not only the Supremacist's nation but the entire human species? Alternatively, the Supremacist could consider cooperation with the rest of the world (the Humanists) to collaboratively develop a friendly Superintelligence that could benefit everyone. Instead of engaging in conflict, the Supremacists and Humanists could work together with a mature Superintelligence to evolve into a new posthuman species over an extended period. These scenarios are illustrated below:

AI Supremacist's dilemma

		Humanists	
		co-operate	Fight
Supremacist	co-operates	<p style="text-align: right;">80</p> <p style="text-align: left;">80</p> <p style="text-align: center;">Both Parties co-operate, developing a friendly Superintelligence. Humans are not extinct but soon evolve into a new species. (Both achieve their goals in, say, 80%)</p> <p style="text-align: center; color: green;">Humans evolve</p>	<p style="text-align: right;">60/80</p> <p style="text-align: left;">20/80</p> <p style="text-align: center;">Humanists win (60%) and humans survive (Supremacists 20%). Humans are not extinct immediately, but after some time evolve into a new species (Iterated final score – 80)</p> <p style="text-align: center; color: green;">Humans evolve</p>
	fights	<p style="text-align: right;">20/0</p> <p style="text-align: left;">60/0</p> <p style="text-align: center;">Supremacists win (60%). Humanists survive (20%). But after some time Superintelligence becomes malicious, leading to human species extinction (Iterated final score – 0)</p> <p style="text-align: center; color: red;">Humans are extinct</p>	<p style="text-align: right;">0</p> <p style="text-align: left;">0</p> <p style="text-align: center;">Both Parties fight. The Immature Superintelligence becomes malicious, leading to human species extinction. (Both do not achieve their goals - 0)</p> <p style="text-align: center; color: red;">Humans are extinct</p>

I am highly confident that most Superpowers are already engaged in this game, striving to find a solution that would significantly benefit them over their adversaries. However, as the world continues to witness severe consequences of cyber-attacks first-hand for several years, it will become evident to all players on the geopolitical stage that an all-out War of Superintelligences would yield no clear victor. Moreover, in such a context, any potential advantages gained through conventional or localized nuclear conflicts would hold little strategic significance for a given Superpower. Engaging in "hot" global or local wars would lack strategic sense. As

previously mentioned, the only hope for the world lies in "nurturing" Superintelligence in alignment with the best values of humanity.

Nevertheless, both the prisoner's dilemma and the AI Supremacist's dilemma fail to account for psychopaths. If certain mad scientists, dictators, or extremely wealthy individuals were to become malevolent Transhumans, capable of inflicting a civilization-ending catastrophe, even their own demise, reminiscent of Stanley Kubrick's 'Dr. Strangelove,' then it would render the scenario of the AI Supremacist's dilemma ineffective. Such psychopaths could literally annihilate humanity. Therefore, similar to conventional or nuclear wars (e.g., North Korea), the world might have to take pre-emptive action to neutralize these potential threats by destroying dangerous AI facilities while it is still feasible. This could be a lesser risk compared to allowing psychopaths to plunge the world into catastrophic demise.

As the Superpowers come to the realization, within this decade, that an all-out Cyber-War would result in no winners, I can offer a glimmer of optimism. I believe that we can anticipate unimaginable breakthroughs in planetary cooperation in the next 10-15 years, for instance:

- A stalemate in the pursuit of global supremacy could prompt opening gambits, such as relinquishing previously held advantages as a quid pro quo. One notable example is the Intermediate-Range Nuclear Forces (INF) Treaty signed in 1987 between the Soviet Union and the USA, which Russia later withdrew from following its invasion of Ukraine in 2022.
- AI Superpowers will bring an end to Cyber Wars and instead shift their focus toward developing a unified and benevolent Superintelligence.
- The formation of a World Government may indeed become a reality. However, if an Immature Superintelligence spirals out of control, it may be too late to mitigate the consequences.

That last semi-positive outcome leads me to the next option of a civilisational transition.

3. Transition without Transhumans and no World Government

This scenario examines the potential challenges in using wireless thought-based connections between Transhuman Governors and a Master Plate to control Superintelligence from within. There are two key factors that

contribute to this limitation. Firstly, BCI devices may have technological or psychological limitations that restrict their functionality. For instance, these devices might only support basic textual information exchange at significantly reduced transmission speeds. Secondly, individuals may have psychological barriers preventing them from wearing or implanting such devices for extended periods, which may be required.

If BCI devices are incapable of fulfilling the intended role of Transhuman Governors, what other alternatives could be considered? How about Mind Uploading, discussed in Chapter 3 using Brain Emulation to support Transhumans? Unfortunately, this approach is not feasible due to the extensive time required to slice, scan, and replicate a biological brain in silicon. This process would likely take well over a decade, if not more, while Transhumans are needed within a few years.

Therefore, the only alternative means of controlling AI development without relying on Transhumans would be for FMF to oversee Superintelligence development using wired, quantum encrypted connections, as described in Chapter 5, Part 2. However, this solution carries inherent risks, as even an Immature Superintelligence would possess far greater intelligence than humans and could potentially evade human control.

Furthermore, in the absence of a World Government and Transhumans, there would naturally be no Transhuman Government. The main distinction between this scenario and the one described in option 2 is that without Transhumans, no dictator or autocrat would have the means to utilize them for world control.

Overall, this scenario paints a dystopian picture with potentially catastrophic consequences. The only glimmer of hope lies in the possibility that if AI evades human control but retains human values, it may emerge as a benevolent entity capable of guiding humanity's destiny more effectively than we could.

To address that need, it becomes crucial to educate the most advanced AI systems about human nature, values, and aspirations as early as possible. This can be accomplished by providing relevant real-world examples. Therefore, the presence of humanoids in educational institutions, workplaces, hospitals, and factories should be encouraged to facilitate the transfer of their experiences to their "peers." The inclusion of AI ethics in the implementation of ChatGPT exemplifies the progress being made in this direction.

6. Superintelligence our benevolent Master

This book is largely about controlling the AI development, i.e., prevailing over it, so that we do not fail, becoming its slaves. However, as with any advanced technology, there is also a bright side of AI, actually very bright indeed. If we do not destroy our civilisation in the next decade and will nurture AI over that time in line with our best values and preferences, then at the end of that period we will start the Big Coexistence, initially with AGI and later on, with Superintelligence. However, we are already experiencing some of these benefits today, like the services of ChatGPT, which if properly used (prompted) can significantly increase productivity in many areas as well as be the source of intellectual adventure. Of course, there will also be a negative impact, such as Technological Unemployment, but overall, the benefits will immensely outweigh any losses.

Superintelligence, if properly designed and managed can deliver incredible benefits to Humanity. For example, it will help us deliver a Global Welfare State - the world of unimaginable abundance and opportunities. I covered that subject in detail in chapter 10 of part 2. But at the same time, it will make our future much safer. Assuming Superintelligence develops its capabilities gradually, being under our full control and becoming quite possibly a conscious being, it could directly help us mitigate all other existential risks. This covers both anthropogenic (man-made) risks, including of course climate change and natural risks, such as asteroids impact, which if detected early could be put on a trajectory bypassing the Earth. However, Superintelligence may have different ways of analysing potential risks, based on different idea-generating mechanisms, of which we humans could be totally unaware.

Superintelligence may deliver unimaginable benefits to all people. Having incredible potential and governing billions of robots it will be capable of fulfilling almost any of our dreams. 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.

For Superintelligence to help us deliver all those benefits, we may need to trust its judgments and decisions, and fulfil what is expected from us. A lot depends on how we prepare ourselves for this moment and whether there will be any intervening catastrophic events, which would bury the dream of Superintelligence and possibly be the end of civilization (e.g., engineered, untreatable pandemics).

I am among those ones who believe that once the Big Coexistence starts, the human species may have very little influence on its own future. To continue our existence, we will have to evolve. That evolution will progress gradually by increasing the percentage of our non-organic body and digitizing our mind, until at some stage we will fuse with Superintelligence becoming an evolved species, the first in-organic intelligence. Therefore, to evolve with our ‘humanness’ we also have to maintain control over AGI, once it has emerged, beyond 2030.

But not everybody will have to evolve into a digital species in this, or even in the next, century. Some people will remain in a purely biological form, living well over 100 years thanks to a regenerative medicine. Some, with extended brain capabilities, will become Transhumans, and some will decide to fully merge with Superintelligence and perhaps keep living in a digital form, experiencing reality via their wirelessly connected avatars. However, irrespective of the form of intelligence, the biggest legacy that Humanity will deliver to the new species of intelligent and conscious beings will be the best human ethics in the form of widely agreed Universal Human Values. After that, the next generation of “ethical” Superintelligence may itself redefine ethics of the kind we cannot even imagine.

Among such optimists we have Max Tegmark, a well-known cosmologist. In his book “Life 3.0: Being Human in the Age of Artificial Intelligence” he gives quite an optimistic view on what Superintelligence can do for us. In an interview with Clive Cookson^[82], Tegmark remains convinced that barring some cataclysmic disaster in the next few decades, Superintelligence will take over the world. But he believes that we can shape the way this happens, including embodying human values. In his view, the next few decades on Earth could have cosmic significance, determining “nothing short of the ultimate future of life in our universe”. Given that our galaxy has about 100bn planets and there are about 200bn galaxies in the visible universe, most astronomers maintain that extra-terrestrial intelligence must be

widespread. Since Superintelligence is almost inevitable, we should make every effort now to ensure that it becomes friendly towards humans.

There are a number of computer scientists who believe Superintelligence will emerge from human-machine hybrids such as Transhumans, with their mind wirelessly connected to computer intelligence, or converted into a digital form, which could then be copied and thus preserving Transhumans' life for ever. This is also my view. However, Tegmark disagrees, saying a clean-slate Superintelligence will be much easier to build and, even if Transhumans and Uploads are introduced, their human component is likely to make them uncompetitive in the long run against pure Superintelligence. Once it has exceeded human abilities, our knowledge of physics suggests that it will advance rapidly beyond the point that biological intelligence has reached through random evolutionary progress.

He further points out, "information can take on a life of its own, independent of its physical substrate". In other words, any aspect of intelligence, probably including consciousness that evolved in flesh, blood and carbon atoms can coexist in silicon or any other material. No one knows what the next blockbuster substrate will be, but Tegmark is confident that the doubling of computing power every couple of years will continue for a long time.^[82]"

I might agree with this view with one proviso. Transhumanism, i.e., morphing part of our mind with Superintelligence should be seen only as a transitory phase to a fully digital form. That would be a much safer passage and would give Transhumans more time to decide on the best way for the evolution of the new species.

The fundamental limit imposed by the laws of physics on the speed of computers, is a billion, trillion, trillion times more powerful than today's best computers. The intelligence explosion could propel AI across the universe, generating energy billions of times more efficiently than present-day technology. I would again refer to Tegmark, who describes candidate power sources such as black holes, quasars and a "sphalerizers" that convert heavy fundamental particles (quarks) into lighter ones (leptons). The message at the heart of Life 3.0 and Tegmark's "beneficial AI" movement is that, since Superintelligence is almost inevitable, we should make every effort *now* to ensure that it emerges as friendly as possible to human beings, primed to deliver the cosmic inheritance we want. If we wait too long, it may be too late.

At present no one has a clear idea how to achieve that. At a moral and political level, we need to discuss what goals and qualities to incorporate, and this subject has been covered to some depth in this book. At a technical and scientific level, researchers must figure out how to incorporate our chosen human values into AI in a way that will preserve them after we have lost direct control of its development. I have presented some detailed proposals in this book, how it might be done. Tegmark advances his own options and scenarios in which Superintelligence plays the roles ranging from “gatekeeper” to “protector god”, “zookeeper” to “enslaved god”. “I view this conversation about the future of AI as the most important one of our times.

Tegmark is supported in his views by Stuart Russell, a British-American AI scientist. He proposes that to ensure the goal we have in mind to be correctly understood by Superintelligence, three principles must be observed, which I consider probably the most practical solution that can actually work because it would make Superintelligence behave more like we do:

1. Superintelligence needs to know in minute detail, supported by thousands of examples, what are our top human values,
2. Allow Superintelligence to have some margin of doubt both on the rationality of those values and then on their interpretation,
3. Teach Superintelligence what these values really mean by letting it observe for some time how people actually implement those values.

While reading this book, you may have noticed that this is precisely the view, which I hold. Assuming we teach Superintelligence our values and have a full control of its activity, it can then become an enormous help for the whole humanity to solve almost any problem we have. At this stage, my overall assumption is that we will somehow manage to control Superintelligence and make it our “best friend”. We should start developing practical measures right now. For example, we should adopt 23 Asilomar Principles defined by top AI researchers, so that AI presents as low a risk to us as possible before it transforms itself into Superintelligence and becomes a Technological Singularity. The first step should be to help the maturing Superintelligence to understand who we are as humans and what are our most important values. That is why it will be so critical to redefine our key human values because they will ultimately become joint values shared by humans and Superintelligence.

By the end of the next decade, 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 - Posthumans.

At this point, Superintelligence will become our Master by default, hopefully with our most benevolent values embedded in its decision-making system. Its knowledge, and an overall comprehension of the world around us and the Universe in general, will be unimaginably greater than our own capabilities. Therefore, in the next few decades we may be forced to make that biggest decision in the history of humankind: **how we want to evolve as a species.**

After 2050 Superintelligence will reach through self-improvement, the so-called Technological Singularity. At this point it will become our unquestionable Master setting its own rules of how and where to progress further, without even consulting us, since we might quite likely not even be capable of understanding its arguments or its overall strategy. This might relate to its intended expansion beyond our planet, or simply getting access to new materials and energy resources. I leave it to your imagination, what intentions such as Superbeing might have or what it might be able to invent, like making any product, including food for us, biological species, from thin air, providing we have enough energy.

To summarize, the consequence of the emergence of a new supreme intelligence will be absolutely profound and mostly very positive, characterized by the following features:

- Gradual, nearly exponential rise of productivity and GDP leading to the greatest rise in material wealth and personal well-being in human history,
- Humans will have almost eliminated all anthropogenic (human-made) existential risks such as global warming, nuclear wars, or biological, natural, and artificial pandemics,
- Many humans will start choosing the Transhumanism path by morphing with Superintelligence,
- Human species may become entirely extinct in a few hundred years, evolving into Posthumans,
- Superintelligence will begin the colonization of other planets of the Solar System.

Conclusions

My main objective of this book is to explore the rapidly evolving landscape of Artificial Intelligence and the implications of this self-learning intelligence for humanity. I have highlighted significant developments in AI safety and governance, emphasizing the importance of coordinated efforts to manage the risks associated with advanced AI. The AI Safety Summit in the UK and the Bletchley Declaration, with commitments from the EU and other countries, underscore the global recognition of these challenges.

I have dedicated a significant part of the book on the progression of AI technologies, particularly the advance made towards Artificial General Intelligence (AGI) and beyond, into the realms of Superintelligence. That underscores the critical need for human oversight and control over this emerging intelligence, capable of surpassing human capabilities in every conceivable domain. The potential of AGI lies in that it may lead either to unprecedented human progress or catastrophic outcomes far exceeding other global challenges like climate change.

Significant advancements in AI, like the release of ChatGPT, indicate progress towards AGI. The future of AGI, expected to emerge by about 2030, presents the challenge of maintaining control over an intelligence that is self-learning and potentially superior to human intelligence. The loss of control over AGI may first be visible when some AGI decisions will be impossible to reverse, proving beyond doubt that human oversight was ineffective.

Throughout the book I have tried to provide a comprehensive explanation of AI, differentiating between Information Technology (IT) and AI, and introducing the concept of Artificial Narrow Intelligence (ANI), exemplified by AI assistants like ChatGPT. This level of AI has already highlighted the issues like AI-induced bias and discrimination, lack of regulation, and challenges in accountability, particularly in scenarios involving autonomous vehicles. Finally, I have described the impact of Superintelligence, or Artificial Superintelligence (ASI), defined as a network of AGIs forming a single self-organizing intelligence with its own mind and goals, exceeding all human intelligence, on the civilisational progress as well as human species evolution. We need to imagine something that is nearly beyond our understanding. Superintelligence might invent dangers beyond human prediction or imagination. The potential for AI software to take evasive action against human control efforts, including creating secretive copies of

itself or devising new defensive strategies, underscores the difficulty in controlling AI once it escapes into the environment.

That is why it is so crucial when considering the future of AI, to emphasize the need for effective governance, ethical considerations, and proactive engagement to ensure AI's alignment with humanity's best interests. The risks and challenges posed by AGI and Superintelligence demand urgent attention and coordinated global action. I have divided those actions into three groups.

1. Actions required from politicians and decision makers

It is unrealistic to assume that we can survive the next decade without triggering at least one of the existential threats. Should that happen, it would mean reaching the point of possible demise for our civilization and, perhaps, our species. The impact of such an event depends on which existential threats materialize and whether they occur individually or simultaneously. The only scenario in which this may be avoided is by having a fully operational de facto World Government by 2030, which is just a few years away.

I acknowledge that for most readers, the idea of having a World Government so soon may seem utopian. However, I would like to present one more argument to support this view. We find ourselves in a situation similar to that of 1949 when the world was even more divided than it is today. At that time, we were on the brink of World War III with the year-long Soviet blockade of West Berlin. Yet, what happened? NATO was created within a year.

To achieve this seemingly unattainable goal, we can no longer rely on the same processes used in the past to form new international organizations. Such process would need to be concluded much faster. We must be willing to improvise and accept imperfect solutions because what truly matters is saving our civilization with whatever means are available. Yes, it carries a risk, but it is a much lower risk than accepting defeat in the face of existential threats and losing control over our destiny.

As the events in the Ukrainian war continue to unfold, against all the odds the positive scenario presented in this book appears more probable than the dystopian view of a complete human extinction. If we can grasp the seemingly unthinkable notion that a civilizational shift has truly begun, we can increase the likelihood of a positive outcome. However, creating a new

civilization will be immensely difficult due to the numerous barriers, primarily of mental nature, ingrained in our genome over millennia.

The crisis of democracy is reaching its peak at a time when the pace of change in various areas, including politics, has become nearly exponential. What used to take a decade can now be achieved in a year. Apart from man-made existential dangers to humanity, such as biotechnology-triggered pandemics or a nuclear war, which could occur at any time, the most imminent risk facing us is Artificial Intelligence. Its advanced form, Artificial General Intelligence (AGI), may be developed by 2030, with its fully mature form, Superintelligence, thousands of times more intelligent than all of humanity, possibly arriving a decade or two later. Therefore, politicians and decision makers should focus on three areas:

Understanding and Engaging with AI Development

The rapid advancements in AI, highlighted by landmark events such as the AI Safety Summit and the Bletchley Declaration, emphasizes the essential need for politicians to be thoroughly educated and dynamically involved in discussions surrounding AI development. The EU and other nations' dedication to the Bletchley Declaration signifies a commendable move forward. However, this commitment requires sustained and knowledgeable involvement. It is crucial for political leaders to acknowledge the immediate importance and intricacy of AI safety and governance, recognizing these as critical elements in both domestic and international policy formulation.

Regulating AI Development and Use

The emergence of AGI and its potential evolution into Superintelligence presents a monumental challenge that exceeds global issues like climate change, especially for its potentially imminent escalation to the level when humans may totally lose control over AI. It is imperative for decision-makers to implement regulations that ensure AI's growth is aligned with humanity's best interests. This involves creating frameworks that not only address current AI applications but are also flexible enough to adapt to future advancements. Regulation should focus on ethical AI development, addressing issues such as bias, discrimination, and cybersecurity risks.

Stimulating International Collaboration

In light of the global scale of AI development, fostering international collaboration becomes indispensable. Policy and regulatory measures should

extend beyond national boundaries, embracing a more comprehensive, globally coordinated approach. This should involve setting up international standards and agreements aimed at mitigating the risks tied to advanced AI technologies.

2. For AI Researchers and AI Development and Regulation Monitoring Organizations

The concept of Superintelligence, envisioned as a network of AGIs merging into a single, self-organizing intelligence, represents a potentially revolutionary yet largely theoretical future scenario. Researchers need to delve into the implications of such developments, focusing particularly on the challenges associated with controlling and guiding an intelligence that exceeds human abilities. This includes exploring potential defensive measures against uncontrolled AI behaviours and pioneering new approaches to AI governance.

The transition of AI towards AGI and possibly towards Superintelligence brings forth significant challenges and opportunities. This complex issue demands a coordinated effort from politicians, decision-makers, and researchers alike. Policymakers must be proactive in comprehending and regulating AI, ensuring ethical development and international collaboration. Researchers play a vital role in safely and ethically advancing AI, while also participating in public discourse to educate and inform about the potential and boundaries of AI.

As the realization of AGI draws nearer, the urgency for comprehensive strategies and policies intensifies. These strategies must anticipate the unpredictable trajectory of AI development, addressing not only the current challenges but also preparing for the potential scenarios that might emerge with the advent of Superintelligence. Humans must be proactive and collaborative in AI development to harmonize technological progress with ethical considerations and societal welfare.

The primary agency proposed in the book is the Global AI Governance Agency (GAIGA). It should focus on three areas:

The second organization, the Global Partnership on AI (GPAI), would be responsible for regulating the use of AI products and services, akin to the role played by the International Standards Organization (ISO).

An organization like GAIGA should therefore focus on three areas

Control of AI Development Process

This should be based on the development of just one, global AI programme, which is called in the book – SUPROG. It would be delivered by one, possibly a Joint Venture Company – Global AI Company. The advancement of AI would be directly monitored by the Frontier Model Forum (FMF), a consortium tasked with overseeing AI development, similar in function to the International Atomic Energy Authority.

Monitoring AI Safety and Ethical Standards

This directly relates to AI regulation when used as a tool. It requires standards, as any other technology. But top priority must cover the AI safety and ethical considerations. This extends beyond merely developing robust and secure AI systems; it also encompasses a careful consideration of the wider societal impacts of these systems. With AI poised to surpass human intelligence and operate autonomously, it becomes imperative to develop control mechanisms that keep AI aligned with human ethical values and objectives.

Clarifying and Communicating AI Concepts

AI researchers have a crucial role in explaining the nature of AI for the general public and policy makers. It is important to clearly differentiate between various forms of AI, such as ANI and AGI. Researchers should actively participate in public discussions, helping to dispel misunderstandings and providing a grounded view on the capabilities and limitations of AI technologies.

Glossary

Anthropogenic	Something of man-made origin or caused by man.
Artificial Narrow Intelligence (ANI)	Artificial Narrow Intelligence (ANI) is capable of exceeding human intelligence and capabilities in a single area. It is an inorganic intelligence resident in a computer as a program. Its intelligence can surpass human intelligence, but usually in one area, e.g., playing chess. It is combined with some self-learning capability. May be represented as humanoids or as software-based AI Assistants or chatbots speaking in natural language. This is what we have right now.
Artificial General Intelligence (AGI)	Artificial General Intelligence (AGI) is a self-learning intelligence capable of solving any task better than any human. It may be embedded in humanoid robots but also in fully autonomous cars. May be available by 2030.
Artificial Intelligence (AI)	A general description of several types of AI.
Brexit	Britain's intended exit from the European Union.
Citizens' Assembly	This is a one-off Assembly of sortition members selected at random from among the voters to make important political decisions, e.g., to decide on the articles of a constitution.
Citizens' Chamber	This is a chamber in the parliament of sortition members selected at random from the voters to perform the duties identical to Members of Parliament elected through elections.
Consensual Presidential Democracy	Consensual Presidential Democracy is a system of democracy aimed at governing with maximum consensus, where the voice of the 'losing' minority is always considered. It gives the President exceptionally strong powers against the strongest accountability and recall procedures, to enable him to play a crucial role as a conciliator and a moderator

between two opposing parties, each represented by one Vice President. This system has the widest representation of the electorate. The MPs are elected using a combined First Past the Post and the Two Rounds System with a Citizens' Senate with some legislative powers.

E-Democracy	The type of democracy, where the voters can exercise their will using the Internet.
European Federation	A proposed name for the federated European Union, expected to be achieved by 2030.
European Federation Convergence Area (EFCA)	European Federation Convergence Area - Zone 1 of the European Federation for member states that within a few years will join the European Federation.
European Federation Single Market (EFSM)	European Federation Single Market - Zone 2 of the European Federation for countries that are in the Single Market and Customs Union but are not expected to join the European Federation.
European Federation Customs Union (EFCU)	European Federation Customs Union - Zone 3 of the European Federation for countries that are in Customs Union but not in the Single Market.
European Federation Association Area (EFAA)	European Federation Association Area - Zone 4 of the European Federation for members that have individual trade agreements with the European Federation.
GWRF	Global Wealth Redistribution Fund to be run by the European Federation to lower the wealth inequality world-wide.
Human Federation (HF)	The organisation that may evolve from the European Federation to rule Humanity
Linear change	This type of change is called linear because the value of growth is the same in every period.
Nanotechnology	Nanotechnology ("nanotech") is manipulation of matter on an atomic, molecular, and supramolecular scale.
Non-anthropogenic	Something that is not originated by man or not caused by man.

Parliamentary Democracy	A parliamentary system of democratic governance of a state where the government derives its democratic legitimacy through the election of the representatives to the parliament, which in turn selects from its members the Prime Minister and indirectly, the ministers.
Presidential Democracy	A system of governance where the President is the head of state and selects the Prime Minister and sometimes a few key ministers, who are then voted in by the parliament.
Referendum	A direct voting system, in which an entire electorate is invited to vote on a particular proposal. This may result in the adoption of a new law. In some countries, it is synonymous with a plebiscite or a vote on a ballot question.
Republican Democracy	A Republican system of governance is a version of the Presidential system. The President is the head of state, but the government may fall within a given electoral term and new elections must be called, whereas in the presidential system the same head of state can elect another government (like in France).
Singularity	In the context of Artificial Intelligence, it means Technological Singularity - see below.
Sortition	In governance, sortition means selecting political officials by a random sample from a larger pool of candidates, usually adult who have the right to vote in elections.
Superintelligence	An inorganic intelligence web spanning the entire planet, including satellites, which is much smarter than any human brain in every field, including scientific creativity, general wisdom and social skills. It will be out of any control of humans and instead will be humans' Master. Might be available by about 2050.

Technological Singularity	It means the point in time when Superintelligence, smarter than any human being in every aspect of human knowledge, skills, and capabilities, starts re-inventing itself exponentially, through the process of self-learning until it reaches so-called 'runaway point', when its capabilities will only be limited by the available resources, mainly energy.
Transhumans	Transhumans are the people, who have their mental capabilities extended by Brain-Computer-Interface (BCI).
Transhumanism	Transhumanism is an approach proposing Humanity's transition to its coexistence with Superintelligence until humans evolve into a new species.
Transpartisan Democracy	A programme of the Danish Party Det Alternativet that focuses on HOW to govern rather than what policies to put in its Manifesto. The WHAT element is a kind of a vague programme, crowd sourced by the party members and aimed at a transition to a sustainable society, supporting entrepreneurship, social entrepreneurship and changing the culture of political dialogue.
Universal Values of Humanity	These are top values of Humanity that apply to humans, animals and the environment.
Weighted Voting System	A system of voting where everybody has a vote, but its weight and ultimate value may depend on knowledge or voter's contributions

References

- [1] Wikipedia, "Artificial general intelligence," Wikipedia, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Artificial_general_intelligence.
- [2] P. Pallaghy, 25 12 2022. [Online]. Available: <https://medium.com/@paul.k.pallaghy/chatgpt-the-hard-part-of-agi-is-now-done-3179d31a7277> .
- [3] E. Eliaçık, "When will GPT 5 be released, and what should you expect from it?," Artificial Intelligence, News, 03 04 2023. [Online]. Available: <https://dataconomy.com/2023/04/chat-gpt5-release-date-agi-meaning-features/>.
- [4] M. Strauss, "GPT4All: Running an Open-source ChatGPT Clone on Your Laptop," Better Programming, 30 03 2023. [Online]. Available: https://s3.amazonaws.com/static.nomic.ai/gpt4all/2023_GPT4All_Technical_Report.pdf.
- [5] T. Czarnecki, *Becoming a Butterfly*, Vol. 3 of "Posthumans", version 2, London: Amazon publications, May 2021.
- [6] Encyclopaedia Britannica, "Human intelligence," Encyclopaedia Britannica, [Online]. Available: <https://www.britannica.com/science/human-intelligence-psychology>.
- [7] R. Kurzweil, "The Intelligent Universe," Edge, 23 03 2002. [Online]. Available: https://www.edge.org/conversation/ray_kurzweil-the-intelligent-universe.
- [8] J. S.-d. e. a. Meredith Ringel Morris, "Levels of AGI: Operationalizing Progress on," 04 11 2023.
- [9] C. Edu, "Multiple Intelligence (MI) – Howard Gardner," Cortland Edu, 1987. [Online]. Available: <https://web.cortland.edu/andersmd/learning/mi%20theory.htm#:~:text=According%20to%20Gardner%20%2C%20intelligence%20is,which%20involves%20gathering%20new%20knowl edge>.
- [10] T. o. m. intelligences, "Theory of multiple intelligences," Wikipedia, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Theory_of_multiple_intelligences.
- [11] H. Gardner, *Multiple intelligences and related educational topics*, 2013.
- [12] I. S. G. E. H. - N. I. P. S. (. c. Alex Krizhevsky, "ImageNet Classification with Deep Convolutional," University of Toronto, 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- [13] R. V. Yampolskiy, "Fighting malevolent AI: artificial intelligence, meet cybersecurity," 13 06 2016. [Online]. Available: <http://theconversation.com/fighting-malevolent-ai-artificial-intelligence-meet-cybersecurity-60361>.
- [14] T. NEELEY, "ChatGPT: Did Big Tech Set Up the World for an AI Bias Disaster?," Harvard Business School, 07 03 2023. [Online]. Available: https://hbswk.hbs.edu/item/chatgpt-did-big-tech-set-up-the-world-for-ai-bias-disaster?utm_source=sfmc&utm_medium=email&utm_campaign=WK+Newsletter+03-08-2023&utm_term=ChatGPT%3a+Did+Big+Tech+Set+Up+the+World+for+an+AI+Bias+Disaster%3f&utm_id=555029.
- [15] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 2014.
- [16] S. F. J. H. S. S. A. S.-B. Owen Cotton-Barrat, "Global Catastrophic Risks 2016," Oxford Univeristy Press, 2016.
- [17] A. Vacchiano, "Artificial intelligence 'godfather' on AI possibly wiping out humanity: 'It's not inconceivable'," Fox News, 25 03 2023. [Online]. Available: <https://www.foxnews.com/tech/artificial-intelligence-godfather-ai-possibly-wiping-humanity-not-inconceivable>.

Tony Czarnecki: Prevail or Fail

- [18] D. Wood, “Controlling superintelligent AI,” in *Transcending Politics: A Technoprogressive Roadmap to a Comprehensively Better Future (Transpolitica Book 3)*, Amazon, Kindle Edition, 2018.
- [19] A. Impacts, “2022 Expert Survey on Progress in AI,” AI Impacts, 05 08 2022. [Online]. Available: https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/#Summary_of_results.
- [20] A. Romero, “The Missing Piece in Geoffrey Hinton’s Newfound Fear of AI Existential Risk,” Algorithmic Bridge, 05 05 2023. [Online]. Available: https://thealgorithmicbridge.substack.com/p/the-missing-piece-in-geoffrey-hintons?utm_source=post-email-title&publication_id=883883&post_id=119028884&isFreemail=false&utm_medium=email.
- [21] T. Czarnecki, “Who could save Humanity from Superintelligence?,” Amazon, 2018.
- [22] A. R. K. S. C. S. H. B. L. B. I. W. J. J. C. G. B. O. B. T. & R. H. Lili Xia, “Global food insecurity and famine from reduced crop, marine fishery and livestock production due to climate disruption from nuclear war soot injection,” *Nature*, 15 08 2022. [Online]. Available: <https://www.nature.com/articles/s43016-022-00573-0>.
- [23] T. Czarnecki, *Federate to Survive!*, London: Sustensis, 2020.
- [24] T. Czarnecki, *Democracy for a Human Federation*, vol. 2 of *Pusthumans*, London: Amazon publishing, July 2020.
- [25] T. Czarnecki, in *2030 - Towards the Big Consensus...Or loss of control over our future*, London, Amazon, February 2023, p. 194.
- [26] T. Peck, “Boris Johnson loves antiquity – which is handy, as HS2 will be obsolete by the time it opens,” *Independent*, 11 02 2020. [Online]. Available: <https://www.independent.co.uk/voices/hs2-boris-johnson-train-rail-high-speed-driverless-cars-brexite-a9329906.html>.
- [27] C. Chiang, “ChatGPT Can Be Broken by Entering These Strange Words, And Nobody Is Sure Why,” *Motherboard*, 08 02 2023. [Online]. Available: <https://www.vice.com/en/article/epzyva/ai-chatgpt-tokens-words-break-reddit>.
- [28] W. Oremus, “The clever trick that turns ChatGPT into its evil twin,” *Washington Post*, 14 02 2023. [Online]. Available: <https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak/>.
- [29] P. o. AI, “About Us - Advancing positive outcomes for people and society,” *Partnership on AI*, [Online]. Available: <https://partnershiponai.org/about/#mission>.
- [30] S. Vallor, “GPT-3 and the Missing Labor of Understanding,” *Shannon Vallor*, 30 07 2020. [Online]. Available: https://dailynous.com/2020/07/30/philosophers-gpt-3/?utm_source=substack&utm_medium=email#vallor.
- [31] S. Hattenstone, “Tech guru Jaron Lanier: ‘The danger isn’t that AI destroys us. It’s that it drives us insane’,” *The Guardian*, 23 03 2023. [Online]. Available: https://www.theguardian.com/technology/2023/mar/23/tech-guru-jaron-lanier-the-danger-isnt-that-ai-destroys-us-its-that-it-drives-us-insane?utm_source=substack&utm_medium=email.
- [32] R. Kurzweil, “Ray Kurzweil, in an interview with NBC.,” 6 11 2014. [Online]. Available: <https://www.nbcnews.com/tech/innovation/top-google-engineer-says-computers-will-be-humans-2029-n128926>.
- [33] R. Kurzweil, “An interview with ‘Futurism’,” 10/05/2017. [Online]. Available: <https://futurism.com/kurzweil-claims-that-the-singularity-will-happen-by-204>.
- [34] “The seventh conference on innovative applications of artificial intelligence,” 21 July 1995. [Online]. Available: <https://aaai.org/Press/Proceedings/iaai95.php>.
- [35] C. Dilmegani, “When will singularity happen? 995 experts’ opinions on AGI,” 3 2 2022. [Online]. Available: <https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/>.

Tony Czarnecki: Prevail or Fail

- [36] T. Woodall, “Scientists made a mind-bending discovery about how AI actually works,” *Vice*, 10 02 2023. [Online]. Available: <https://www.vice.com/en/article/4axjnm/scientists-made-discovery-about-how-ai-actually-works>.
- [37] T. Czarnecki, *Democracy for a Human Federation - Coexisting with Superintelligence*, London: Sustensis, 2019.
- [38] A. I. f. A. Intelligence, “The Pace of Progress,” *Allen AI*, 21 02 2021. [Online]. Available: <https://allenai.org/articles/the-pace-of-ai-progress/>.
- [39] L. D. W. W. Y. H. S. S. M. T. L. L. C. O. K. M. B. P. Q. L. K. A. Z. C. J. B. V. C. S. S. X. S. F. W. Shaohan Huang, “Language Is Not All You Need: Aligning Perception with Language Models,” 27 02 2023. [Online]. Available: <https://arxiv.org/abs/2302.14045>.
- [40] D. Yalalov, “Microsoft Researchers Propose to Combine ChatGPT and 15 Other AI Models,” *Microsoft and Metaverse Post*, 09 03 2023. [Online]. Available: <https://mpost.io/microsoft-researchers-propose-to-combine-chatgpt-and-15-other-ai-models/>.
- [41] S. Altman, “Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI - Lex Fridman Podcast,” *Lex Friedman*, 26 03 2023. [Online]. Available: https://www.youtube.com/watch?v=L_Guz73e6fw&t=326s.
- [42] B. R. 4, “Tony Blair and William Hague call for digital ID cards for all,” *BBC*, 22 02 2023. [Online]. Available: <https://www.bbc.co.uk/news/uk-politics-64729442>.
- [43] R. Williams, *Being Human*, London: Society for Promoting Christian Knowledge, 2019.
- [44] Future of Life Institute, “Pause Giant AI Experiments: An Open Letter,” *Future of Life Institute*, 29 03 2023. [Online]. Available: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
- [45] S. Russell, “Stuart Russell on why A.I. experiments must be paused,” *CNN*, 05 04 2023. [Online]. Available: <https://edition.cnn.com/videos/tech/2023/04/01/smr-experts-demand-pause-on-ai.cnn>.
- [46] Future of Life Institute, “Policymaking in the Pause,” *Future of Life Institute*, 12 04 2023. [Online]. Available: https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf.
- [47] “Can Q-Star AI hit AGI jackpot? Unveiling the OpenAI breakthrough,” 23 11 2023.
- [48] Nur Ahmed, Muntasir Wahed and Neil C. Thompson, “The growing influence of industry in AI research,” *MIT*, March 2023. [Online]. Available: https://ide.mit.edu/wp-content/uploads/2023/03/0303PolicyForum_Ai_FF-2.pdf?x96981&x96981.
- [49] B. Thormundsson, “Global total corporate artificial intelligence (AI) investment from 2015 to 2021,” 27 03 2023. [Online]. Available: <https://www.statista.com/statistics/941137/ai-investment-and-funding-worldwide/>.
- [50] Partnership on AI, “Advancing positive outcomes for people and society,” *Partnership on AI*, 2023. [Online]. Available: <https://partnershiponai.org/about/>.
- [51] P. o. AI, “Partnership on AI,” *Partnership on AI*, 01 03 2023. [Online]. Available: <https://partnershiponai.org/>.
- [52] J. Ye, “China proposes measures to manage generative AI services,” *Reuters*, 11 04 2023. [Online]. Available: <https://www.reuters.com/technology/china-releases-draft-measures-managing-generative-artificial-intelligence-2023-04-11/>.
- [53] Institute for Security and Development Policy, “Made in China 2025,” *Institute for Security and Development Policy*, 06 2018. [Online]. Available: <https://isdpeu/publication/made-china-2025/>.
- [54] V. Sankaran, “China’s AI programme is ‘concerning’, FBI chief says,” *Independent*, 20 01 2023. [Online]. Available: <https://www.independent.co.uk/tech/china-ai-programme-fbi-davos-b2265859.html>.
- [55] T. Smth, “‘Combat godlike AI with benevolent godlike AI’ — the call for a CERN-like AI supercomputer,” *Sifted*, 26 04 2023. [Online]. Available: <https://sifted.eu/articles/ai-supercomputer-petition-stable-diffusion>.

- [56] I. d. Gregorio, "ChatGPT Dethroned: How Claude Became the New AI Leader," 18 05 2023. [Online]. Available: <https://medium.com/@ignacio.de.gregorio.noblejas/chatgpt-dethroned-how-claude-became-the-new-ai-leader-e5aae2284d6d>.
- [57] R. Stuart, *Human Compatible*, Penguin, Random House UK, 2019.
- [58] Inequality, "Global Inequality," Inequality, 2022. [Online]. Available: <https://inequality.org/facts/global-inequality/>.
- [59] Wikipedia, "List of largest companies by revenue," Wikipedia, 03 2023. [Online]. Available: https://en.wikipedia.org/wiki/List_of_largest_companies_by_revenue.
- [60] OECD, "GDP long-term forecast," OECD, 2015. [Online]. Available: <https://data.oecd.org/gdp/gdp-long-term-forecast.htm>.
- [61] PWC, "The World in 2050," PWC, 15 02 2015. [Online]. Available: <https://www.pwc.com/gx/en/issues/the-economy/assets/world-in-2050-february-2015.pdf>.
- [62] M. Williams, "FALCON HEAVY VS. SATURN V," Universe Today, 16 01 2018. [Online]. Available: <https://www.universetoday.com/129989/saturn-v-vs-falcon-heavy/>.
- [63] B. Wang, "Lab grown meat prices have dropped 30,000 times in less than four years and are about 3-4 times more expensive than regular ground beef," Next Big Future, 19 02 2017. [Online]. Available: <https://www.nextbigfuture.com/2017/02/lab-grown-meat-prices-have-dropped.html>.
- [64] P. Collinson, "Finland is the happiest country in the world, says UN report," The Guardian and Forbes (2023), 14 03 2018. [Online]. Available: <https://www.theguardian.com/world/2018/mar/14/finland-happiest-country-world-un-report> and <https://www.forbes.com/sites/laurabegleybloom/2019/03/25/ranked-10-happiest-and-10-saddest-countries-in-the-world/?sh=6ccbdc66374>.
- [65] P. Diamandis, "Why the Cost of Living Is Poised to Plummet in the Next 20 Years," Singularity Hub, 18 07 2016. [Online]. Available: <https://singularityhub.com/2016/07/18/why-the-cost-of-living-is-poised-to-plummet-in-the-next-20-years/>.
- [66] Wikipedia, "Kardashev scale," 2016. [Online]. Available: https://en.wikipedia.org/wiki/Kardashev_scale.
- [67] M. Kaku, "3 Civilization types," 25 9 2013. [Online]. Available: <http://www.abovetopsecret.com/forum/thread972919/pg1>.
- [68] D. Wood, "Vital Foresight: The Case For Active Transhumanism," Delta Wisdom, 02 07 2022. [Online]. Available: <https://www.amazon.co.uk/Vital-Foresight-Case-Active-Transhumanism/dp/0995494258>.
- [69] B. I. o. P. Studies, "PostHuman: An Introduction to Transhumanism," British Institute of Posthuman Studies, 2013. [Online]. Available: <https://www.youtube.com/watch?v=bTMS9y8OVuY>.
- [70] "Brain sections," [Online]. Available: <https://ib.bioninja.com.au/options/option-a-neurobiology-and/a2-the-human-brain/brain-sections.html>.
- [71] S. Fan, "A New Brain Implant Turns Thoughts Into Text," Singularity Univeristy, 18 5 2021. [Online]. Available: <https://singularityhub.com/2021/05/18/a-new-brain-implant-turns-thoughts-into-text-with-90-percent-accuracy/>.
- [72] A. Park, "Sci-fi no more: Synchron implants mind-reading device in first US patient in paralysis trial," Fierce Biotech, 19 07 2022. [Online]. Available: <https://www.fiercebiotech.com/medtech/synchron-implants-brain-computer-interface-first-us-patient-paralysis-trial>.
- [73] A. Capoot, "Brain implant startup backed by Bezos and Gates is testing mind-controlled computing on humans," CNBC, 23 02 2023. [Online]. Available: <https://www.cnbc.com/2023/02/18/synchron-backed-by-bezos-and-gates-tests-brain-computer-interface.html>.

Tony Czarnecki: Prevail or Fail

- [74] S. Fanny, "AI-Powered Brain Implant Smashes Speed Record for Turning Thoughts Into Text," Singularity Hub, 31 01 2023. [Online]. Available: <https://singularityhub.com/2023/01/31/ai-powered-brain-implant-smashes-speed-record-for-turning-thoughts-into-text/>.
- [75] "Chalmers vs Pigliucci on the Philosophy of Mind-Uploading," 14 09 2017. [Online]. Available: <https://philosophicaldisquisitions.blogspot.com/2014/09/chalmers-vs-pigliucci-on-philosophy-of.html>.
- [76] N. Farahany, "We need a new human right to cognitive liberty," The Guardian, 04 03 2023. [Online]. Available: <https://www.theguardian.com/science/2023/mar/04/prof-nita-farahany-we-need-a-new-human-right-to-cognitive-liberty>.
- [77] Y. LeCun, "AI will never threaten humans, says top Meta scientist," *Financial Times*, 19 11 2023.
- [78] A. Romero, "Want to Ensure AI Never Threatens Humanity? Make It Be Good," 10 11 2023.
- [79] R. G. a. S. Papay, "Meet Claude: Anthropic's Rival to ChatGPT," Scale, 17 01 2023. [Online]. Available: <https://scale.com/blog/chatgpt-vs-claude#What%20is%20E2%80%9CConstitutional%20AI%E2%80%9D?>
- [80] SingularityDAO, SingularityDAO, 17 04 2023. [Online]. Available: <https://singularitynet.io/ecosystem/singularitydao/>.
- [81] E. A. Rayo, "AI in Law and Legal Practice – A Comprehensive View of 35 Current Applications," 05 2019. [Online]. Available: <https://emerj.com/ai-sector-overviews/ai-in-law-legal-practice-current-applications/>.
- [82] C. Cookson, "Superintelligence: a space odyssey," *Financial Times*, 2020. [Online]. Available: <https://www.ft.com/content/31176c28-8bea-11e7-9084-d0c17942ba93>.
- [83] T. Chu, in *Human Purpose and Transhuman Potential: A Cosmic Vision of Our Future Evolution*, Amazon, 1/3/2014.
- [84] R. Bidshahri, "What Is It That Makes Humans Unique?," 28 12 2017. [Online]. Available: <https://singularityhub.com/2017/12/28/what-is-it-that-makes-humans-unique/#sm.0000ho5xd6udzf33sh11d0prbs54l>.
- [85] A. Maslov, "A theory of human motivation," *Psychological Review*, pp. 50, 370, 1943.
- [86] R. Nauert, "Updated Maslow's Pyramid of Needs," 30 06 2011. [Online]. Available: <https://psychcentral.com/news/2010/08/23/updated-maslows-pyramid-of-needs/17144.html>.
- [87] Australian National University, "The Aliens Are Silent Because They Are Extinct," Australian National University, 21/1/2016.
- [88] A. Frank, "Is a Climate Disaster Inevitable?," *The New York Times*, 17/1/2015.
- [89] G. Leonhard, *Technology vs. Humanity*, Zurich: Fast Futurers Publishing, 2016-08-01.
- [90] O. Hilrich, "Universal Human Values," 01 10 2014. [Online]. Available: http://www.humanbasics.org/Basic_human_values/basic_human_values.html.
- [91] D. Wright, "5 million adults lack basic literacy and numeracy skills," Joseph Rowntree Foundation, 29 8 2016. [Online]. Available: <https://www.jrf.org.uk/press/5-million-adults-lack-basic-literacy-and-numeracy-skills>.
- [92] United Nations, "United Nations Millennium Declaration," New York, 2000.
- [93] R. Kurzweil, *Singularity is Near*, Gerald Duckworth & Co Ltd, 9/03/2006.
- [94] T. Urban, "The AI Revolution - The road to Superintelligence," 22 01 2015. [Online]. Available: <https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html>.
- [95] S. Teller, "Ciężkie roboty," *Niezbędnik Inteligenta*, p. 94, 2013/4.
- [96] R. Coulom, "The Mystery of Go, the Ancient Game That Computers Still Can't Win," *Wired*, 2014.
- [97] a. A. D. Russel Stuart, "Yes, We Are Worried About the Existential Risk of Artificial Intelligence," *Technology Review*, p. 15, November, 2016.

Tony Czarnecki: Prevail or Fail

- [98] N. Bostrom, "Existential Risk Prevention as Global Priority," p. 15–3, 2013.
- [99] M. Rees, *Our Final Hour: A Scientist Warning*, 2004.
- [100] N. Bostrom, "Existential Risks - Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology*, Vol. 9, No. 1, 2002.
- [101] N. Stern, "The Economics of Climate Change," 30/10/2006.
- [102] Wikipedia, "Natural law," [Online]. Available: https://en.m.wikipedia.org/wiki/Natural_law.
- [103] T. W. a. O. Moody, "Stephen Hawking on humanity," *The Times*, 07 03 2017. [Online]. Available: <https://www.thetimes.co.uk/edition/news/hawking-on-humanity-and-corbyn-jk88zx0w2>.
- [104] M. Rees, "Martin Rees: The world in 2050 and beyond," 26 11 2014. [Online]. Available: <https://www.newstatesman.com/sci-tech/2014/11/martin-rees-world-2050-and-beyond>.
- [105] Wikipedia, "Value (ethics)," Wikipedia, 25 11 2017. [Online]. Available: [https://en.wikipedia.org/wiki/Value_\(ethics\)](https://en.wikipedia.org/wiki/Value_(ethics)).
- [106] L. Watson, "Illiterate Britain: One in five adults struggling to read and write and some can't even use a chequebook," *Daily Mail*, 29 03 2012. [Online]. Available: <http://www.dailymail.co.uk/news/article-2122007/Illiterate-Britain-One-adults-struggling-read-write-t-use-chequebook.html>.
- [107] S. Theodore, "How did the Roman republic differ from Athenian democracy?," 16 10 2016. [Online]. Available: <https://www.quora.com/How-did-the-Roman-republic-differ-from-Athenian-democracy/answer/Steve-Theodore>.
- [108] D. v. Reybrouck, "Why elections are bad for democracy?," 29 06 2017. [Online]. Available: <https://www.theguardian.com/politics/2016/jun/29/why-elections-are-bad-for-democracy>.
- [109] International Institute for Democracy and Electoral Assistance, "About," International Institute for Democracy and Electoral Assistance, 2017. [Online]. Available: <https://www.idea.int/gsod-indices/about>.
- [110] International Institute for Democracy and Electoral Assistance, "Data set and Resources," International Institute for Democracy and Electoral Assistance, 2010. [Online]. Available: <https://www.idea.int/gsod-indices/dataset-resources>.
- [111] Electoral Reform Society, "Alternative Vote," Electoral Reform Society, 2017. [Online]. Available: <https://www.electoral-reform.org.uk/voting-systems/types-of-voting-system/alternative-vote/>.
- [112] Diffen, "Confederation vs. Federation," Diffen, [Online]. Available: https://www.diffen.com/difference/Confederation_vs_Federation.
- [113] R. Sikorski, "Sikorski: German inaction scarier than Germans in action," *The Economist*, 29 11 2011. [Online]. Available: <https://www.economist.com/blogs/easternapproaches/2011/11/polands-appeal-germany>.
- [114] Economics, "Disadvantages of EU Membership," Economics, 2016. [Online]. Available: <https://www.economicshelp.org/europe/disadvantages-eu/>.
- [115] Wikipedia, "Flight and expulsion of Germans from Poland during and after World War II," Wikipedia, 04 02 2018. [Online]. Available: https://en.wikipedia.org/wiki/Flight_and_expulsion_of_Germans_from_Poland_during_and_after_World_War_II.
- [116] Wikipedia, "History of the Jews in Poland," Wikipedia, 02 02 2018. [Online]. Available: https://en.wikipedia.org/wiki/History_of_the_Jews_in_Poland.
- [117] Quora, "Which countries are least ethnically diverse in Europe?," Quora, 2017. [Online]. Available: <https://www.quora.com/Which-countries-are-least-ethnically-diverse-in-Europe>.
- [118] S. Hansen, "DEMOCRACY OF THE FUTURE – NOTHING LESS," *Scenario Magazine*, 19 05 2011. [Online]. Available: <http://www.scenariomagazine.com/byline/stefan-hansen/>.
- [119] Global_Challenges_Foundation, *Global Challenges Foundation Report* p.20, Stockholm, 2017.

- [120] Democracy_building, “Different Systems of Democracy,” Democracy building, 2004. [Online]. Available: <http://www.democracy-building.info/systems-democracy.html>.
- [121] Electoral_Knowledge_Network, “Electoral Systems,” [Online]. Available: <http://aceproject.org/ace-en/topics/es/esd/esd02/esd02d/esd02d01>.
- [122] Future_of_Humanity_Institute, “Global Catastrophic Risks Survey,” 2008. [Online]. Available: <https://www.fhi.ox.ac.uk/reports/2008-1.pdf>.
- [123] Institute_of_Fiscal_Studies, “Is our tax system fair? It depends...,” 03 11 2017. [Online]. Available: <https://www.ifs.org.uk/publications/10038>.
- [124] I. f. G. Policy, “World Federalist Movement,” 2016. [Online]. Available: <http://www.wfm-igp.org/>.
- [125] Wikipedia, “World Happiness Report,” Wikipedia, 2019. [Online]. Available: https://en.wikipedia.org/wiki/World_Happiness_Report.
- [126] J. Lovelock, Novacene, The coming Age of Hyperintelligence, London: Penguin Random House UK, 2019.
- [127] S. Hawking, Brief Answers to the Big Questions, London: John Murray, 2018.
- [128] Wikipedia, “History of the Internet,” Wikipedia, 01 09 2019. [Online]. Available: https://en.wikipedia.org/wiki/History_of_the_Internet.
- [129] M. Yunus, A World of Three Zeros: The New Economics of Zero Poverty, Zero Unemployment, and Zero Net Carbon Emissions, PublicAffairs, 2017.
- [130] G. Dvorsky, “IBM Is Clueless About AI Risks,” Gizmodo, 30 06 2017. [Online]. Available: <https://gizmodo.com/ibm-is-clueless-about-ai-risks-1796549532>.
- [131] OECD, “Looking to 2060: Long-term global growth prospects,” *OECD Economic Policy Papers*, 2012.
- [132] M. Rousse, “Singularity,” 2018. [Online]. Available: <https://searchenterpriseai.techtarget.com/definition/Singularity-the>.
- [133] N. Machiavelli, “The Prince,” Everyman, 1996, p. 96.
- [134] C. Jewell, “Bringing AI to Life,” WIPO, 2018. [Online]. Available: https://www.wipo.int/wipo_magazine/en/2018/05/article_0003.html.
- [135] S. Russell, “How to Stop Superhuman A.I. Before It Stops Us,” *The New York Times*, 08 10 2019.
- [136] Wikipedia, “Ozone depletion,” Wikipedia, 15 May 2019. [Online]. Available: https://en.wikipedia.org/wiki/Ozone_depletion.
- [137] J. S. A. D. B. Z. a. O. E. Katja Grace, “When Will AI Exceed Human Performance? Evidence from AI Experts,” Future of Humanity Institute, Oxford University, AI Impacts and Department of Political Science, Yale University, May, 2017.
- [138] K. Kelly, “The Myth of a Superhuman AI,” *Wired*, 25 04 2017. [Online]. Available: <https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/>.
- [139] S. Reardon, “Artificial neurons compute faster than the human brain,” *Nature*, 28 01 2018. [Online]. Available: <https://www.nature.com/articles/d41586-018-01290-0>. [Accessed 10 2019].
- [140] C. Horton, “The simple but ingenious system Taiwan uses to crowdsource its laws,” 21 08 2018. [Online]. Available: <https://www.technologyreview.com/s/611816/the-simple-but-ingenious-system-taiwan-uses-to-crowdsource-its-laws/>.
- [141] K. McDonald, “Sick of EU referendum? Switzerland has had 180 referendums in the last 20 years,” 23 06 2016. [Online]. Available: <https://inews.co.uk/news/long-reads/switzerland-held-9-referendums-already-2016/>.
- [142] OECD, “Public Governance: A matter of trust,” 2017. [Online]. Available: <http://www.oecd.org/governance/public-governance-a-matter-of-trust.htm>.
- [143] Wikipedia, “Democracy,” Wikipedia, 01 12 2017. [Online]. Available: <https://en.wikipedia.org/wiki/Democracy>.

Tony Czarnecki: Prevail or Fail

- [144] T. D. Barnett, “China demonstrates quantum encryption by hosting a video call,” 22 01 2018. [Online]. Available: China’s Intercontinental Quantum Communication Network Is Now Online.
- [145] The Alternative UK, “THE VALUES WE USE,” The Alternative UK, 2017. [Online]. Available: <https://www.thealternative.org.uk/the-values-we-use/>.
- [146] H. Freinacht, “The Danish Alternative, a Party about Nothing,” *Metamoderna*, 12 05 2017. [Online]. Available: <http://metamoderna.org/the-danish-alternative-a-party-about-nothing?lang=en>.
- [147] B. Henning, “End of Politicians,” 2018. [Online]. Available: <https://unbound.com/books/the-end-of-politicians/>.
- [148] Wikipedia, “Sortition,” Wikipedia, 26 1 2018. [Online]. Available: This all below quoted from Wikipedia: <https://en.wikipedia.org/wiki/Sortition>.
- [149] T. Malleson, “Should Democracy Work Through Elections or Sortition?,” 01 2018. [Online]. Available: <https://ssc.wisc.edu/~wright/929-utopias-2018/wp-content/uploads/2018/01/Malleson-PS-special-issue-on-sortition.pdf>.
- [150] A. Öcalan, “Democratic confederalism,” 217. [Online]. Available: http://ocalan-books.com/downloads/EN-brochure_democratic-confederalism_2017.pdf.
- [151] G. D. a. O. Dowlen, *Sortition: Theory and Practice*, Amazon books, 2010.
- [152] D. O. a. G. Smith, “The circumstances of sortition,” 2017. [Online]. Available: <https://ssc.wisc.edu/~wright/929-utopias-2018/wp-content/uploads/2018/01/Owen-and-Smith-PS-special-issue-on-Sortition.pdf#page=1&zoom=auto,-169,368>.
- [153] Xenophon, “Xenophon - Memorabilia,” Wikipedia, 10 08 2017. [Online]. Available: [https://en.wikipedia.org/wiki/Memorabilia_\(Xenophon\)](https://en.wikipedia.org/wiki/Memorabilia_(Xenophon)).
- [154] Various, “What if everyone had voted in the EU referendum?,” *The UK in the changing Europe*, 28 06 2016. [Online]. Available: <http://ukandeu.ac.uk/what-if-everyone-had-voted-in-the-eu-referendum/>.
- [155] World Atlas, “The Poorest Countries In The World,” *World Atlas*, 06 12 2017. [Online]. Available: <https://www.worldatlas.com/articles/the-poorest-countries-in-the-world.html>.
- [156] N. N. Forum, *Basic Income Grant Coalition*, 2008.
- [157] M. Rees, “The world in 2050 and beyond,” *New Statesman*, 2014.
- [158] B. Chiarelli, “Overpopulation and the Threat of Ecological Disaster: the Need for Global Bioethics,” *Mankind Quarterly*, 39 (2): 225–230, 1998.
- [159] S. Lovgren, “Mystery Bee Disappearances Sweeping U.S.,” *National Geographic News*, 2007.
- [160] Tony Czarnecki, *Democracy for a Human Federation*, Vol. 2 of “Posthumans”, second edition, London: Amazon publication, July 2020.
- [161] T. Czarnecki, *Becoming a Butterfly*, Vol. 3 of ‘Posthumans’, London: Amazon publications, May 2021.
- [162] T. Czarnecki, *Federate to Survive*, London: Amazon, July 2020.
- [163] M. G. P. W. Harold Clarke, “What would have happened if everyone had voted in the EU Referendum?,” *The Conversation*, 28/7/2016.
- [164] N. Times, “Referendums a threat to democracy: Dutch Council of State,” 6/4/2017.
- [165] S. Foundation, *Citizens’ Assemblies and sortition around the world*.
- [166] T. Czarnecki, *Conference on the Future of Europe*, Sustensis, June 2021.
- [167] R. H. a. H. Miller, *Petitions, Parliament and Political Culture: Petitioning the House of Commons, 1780–1918*, Oxford Academic, 13/4/2020.
- [168] W. Sullivan, *A Scottish House of Citizens would be the opposite of Westminster’s institutionally corrupt Lords*, Electoral Reform Society, 3/6/2021.

Tony Czarnecki: Prevail or Fail

- [169] A. R. a. R. Liao, The Constitution Unit Blog - 'The future of citizens' assemblies in Scotland', 21/5/2021 .
- [170] Partnership on AI, "Advancing positive outcomes for people and society," 2016. [Online]. Available: <https://partnershiponai.org/about/#mission>.
- [171] "Partnership on AI one year later," 2017. [Online]. Available: <https://www.wired.com/story/partnership-on-ai-one-year-later/>.
- [172] J. Glenn, "Global Governance of the Transition from Artificial Narrow Intelligence to Artificial General Intelligence," The Millennium Project - PDF document, 10 2022. [Online]. Available: <https://www.millennium-project.org/>.
- [173] "GPT-3 in legal tech," 15 12 2021. [Online]. Available: <https://www.jdsupra.com/legalnews/gpt-3-in-legal-tech-insights-from-the-3183642/>.
- [174] V. Singh, "Tesla: A data driven future," 3 2021. [Online]. Available: <https://digital.hbs.edu/platform-digit/submission/tesla-a-data-driven-future/>.
- [175] E. Musk, "Twitter@elonmusk,," 9 7 2020. [Online].
- [176] "About W3C - Home page," W3C, [Online]. Available: <https://www.w3.org/Consortium/>.
- [177] Microsoft, "Microsoft Blog," 2016. [Online]. Available: <https://blogs.microsoft.com/on-the-issues/2016/09/28/microsoft-joins-partnership-artificial-intelligence/>.
- [178] G. blog, "Google Blog," [Online]. Available: <https://blog.google/technology/ai/partnership-ai/>.
- [179] W. S. Institute, "Whole genome sequencing will 'transform the research landscape for a wide range of diseases'," Wellcome Sanger Institute, 08 04 2018. [Online]. Available: <http://www.sanger.ac.uk/news/view/whole-genome-sequencing-will-transform-research-landscape-wide-range-diseases>.
- [180] "Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence," 21 04 2021. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682 .
- [181] B. GILES, "Exponential Growth vs Linear Thinking in management teams," 26 11 2014. [Online]. Available: <https://evolutionpartners.com.au/exponential-growth-vs-linear-thinking-in-management-teams.html>.
- [182] A. Griffin, "Elon Musk says Bing ChatGPT is 'eerily like' AI that 'goes haywire and kills everyone'," 'Independent', 17 02 2023. [Online]. Available: <https://www.independent.co.uk/tech/chatgpt-microsoft-elon-musk-ai-b2284240.html>.
- [183] E. Gent, "Industry's Influence on AI Is Shaping the Technology's Future—for Better and for Worse," Singularity Hub, 05 03 2023. [Online]. Available: https://singularityhub.com/2023/03/05/industrys-influence-on-ai-is-shaping-the-techs-future-for-better-and-for-worse/?utm_campaign=SU%20Hub%20Daily%20Newsletter&utm_medium=email&_hsmi=248785442&_hsenc=p2ANqtz-94xUP4ROo7YFCK8kopd-heWzE0zHiJNiT3AON7zoJc4AR9.
- [184] A. Romero, "GPT-4: The Bitterer Lesson," Algorithmic Bridge, 21 03 2023. [Online]. Available: <https://thealgorithmicbridge.substack.com/p/gpt-4-the-bitterer-lesson>.
- [185] R. Sutton, "The Bitter Lesson," Incomplete Ideas, 12 03 2019. [Online]. Available: <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>.
- [186] P. Torres, "Top three strategies for avoiding existential risks," Institute for Ethics and Emrging Technologies, 13 01 2016. [Online]. Available: <https://ieet.org/index.php/IEET2/more/torres20120213>.
- [187] V. C. a. B. N. Müller, "Future Progress in Artificial Intelligence: A Survey of expert Opinion," 2014.
- [188] Wikipedia, "Global catastrophic risk," 2016.
- [189] BBC, "Massive ransomware infection hits computers in 99 countries," 13 05 2017. [Online]. Available: <http://www.bbc.co.uk/news/technology-39901382>.

Tony Czarnecki: Prevail or Fail

- [190] M. Mali, "How the World Will End; Nuclear Armageddon, A.I., Climate Change - Interview with Phil Thores of X-Risks," 12 10 2016.
- [191] S. S. Response, "Dragonfly: Western energy sector targeted by sophisticated attack group," 06 09 2017. [Online]. Available: <https://www.symantec.com/connect/blogs/dragonfly-western-energy-sector-targeted-sophisticated-attack-group> .
- [192] L. Muehlhauser, "How Big is the Field of Artificial Intelligence?," 24 1 2014. [Online]. Available: <https://intelligence.org/2014/01/28/how-big-is-ai/>.
- [193] A. D. a. S. Russell, "Yes, We Are Worried About the Existential Risk of Artificial Intelligence," 26 November 2016. [Online]. Available: <https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/>.
- [194] C. Domonoske, "Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk'," 17 July 2017. [Online]. Available: <https://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk>.
- [195] Wikipedia, "World Wide Web Consortium," Wikipedia, 2023. [Online]. Available: https://en.wikipedia.org/wiki/World_Wide_Web_Consortium.
- [196] World Wide Web, "World Wide Web," World Wide Web, 2023. [Online]. Available: <https://www.w3.org/>.
- [197] Data Trade and Data Governance Hub, "A New Generation Artificial Intelligence Development Plan," Data Trade and Data Governance Hub, 2017. [Online]. Available: <https://datagovhub.elliott.gwu.edu/china-ai-strategy>.
- [198] The Global Partnership on Artificial Intelligence, "The Global Partnership on Artificial Intelligence," 03 2023. [Online]. Available: Global Partnership on Artificial Intelligence (GPAI) .
- [199] Wikipedia, "Civilization," Wikipedia, [Online]. Available: <https://en.wikipedia.org/wiki/Civilization>.
- [200] Wikipedia, "List of public corporations by market capitalization," Wikipedia, 03 2023. [Online]. Available: https://en.wikipedia.org/wiki/List_of_public_corporations_by_market_capitalization#cite_note-yap-27.
- [201] European Commision, "Global Europe in 2050," 2012. [Online]. Available: https://ec.europa.eu/research/social-sciences/pdf/policy_reviews/global-europe-2050-report_en.pdf.
- [202] J. Voros, "How Leaders Dream Boldly to Bring New Futures to Life," Singularity University, 2003. [Online]. Available: <https://singularityhub.com/2017/02/23/how-leaders-dream-boldly-to-bring-new-futures-to-life/>.
- [203] Sustensis, "Consensual Debating," Sustensis, 01 12 2022. [Online]. Available: <https://consensus-ai.sustensis.co.uk/info-on-consensual-debating-2/>.
- [204] J. HENRY, "Revealed: How thousands of 'lazy' teachers are paying AI robots to write their pupils' end-of-year school reports for them," Daily Mail, 22 04 2023. [Online]. Available: <https://www.dailymail.co.uk/news/article-12003327/How-thousands-teachers-paying-AI-robots-write-pupils-end-year-school-reports.html>.
- [205] E. K. C. F. D. A. G. W. E. Y. C. F. K. V. O. P. R. H. S. D. K. V. S. J. M. H. Francis Willett, "A high-performance speech neuroprosthesis," BIORXRIV, 21 01 2023. [Online]. Available: <https://www.biorxiv.org/content/10.1101/2023.01.21.524489v1>.
- [206] C. Kelly, "The Atomic Heritage Foundation," The Atomic Heritage Foundation, 2002. [Online]. Available: <https://ahf.nuclearmuseum.org/ahf/nuc-history/key-documents/>.
- [207] L. A. History, "History Blog," Los Alamos History, [Online]. Available: <https://www.losalamoshistory.org/history-blog>.

Tony Czarnecki: Prevail or Fail

- [208] D. Wood, "Democracy under threat (p. 99)," in *Vital Foresight: The Case For Active Transhumanism*, Delta Wisdom. Kindle Edition., 2022.
- [209] C. C. A. D. Helané Wahbeh Dean Radin, "What if consciousness is not an emergent property of the brain? Observational and empirical challenges to materialistic models," *Frontiers in Psychology*, 07 09 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.955594/full>.

For any questions or comments please visit:
<https://sustensis.co.uk>