TONY CZARNECKI

POSTHUMANS - III

BECOMING A BUTTERFLY

EXTINCTION OR EVOLUTION?
WILL HUMANS SURVIVE BEYOND 2050?

London, June 2020

POSTHUMANS – VOLUME 3

Becoming a Butterfly
Extinction or Evolution? Will Humans Survive Beyond 2050?

First published in 2020 by Sustensis

ISBN: 9798645719852

London, June 2020

For any questions or comments please visit:
http://www.sustensis.co.uk

For my grand-daughter Sisi

TABLE OF CONTENTS

FOREWORD TO *POSTHUMANS* SERIES

The POSTHUMANS series has its origin in my first book "Who Could Save Humanity from Superintelligence". That book was addressed primarily to academics, politicians, and readers already familiar with the subjects discussed there. The three books of the series expand those subjects and present them in a way, which may be more suitable for a casual reader.

"**Federate to Survive!"** is about **WHAT** are the existential risks that threaten our **survival as a human species**. Superintelligence is the most imminent and supreme existential risk, as is the climate change, both of which have a tipping point in about 2030. Therefore, minimizing those risks should take precedence above any other goals of our civilisation. The secondary theme in this book is **the selection of an organization**, which could start the federalization of the world and guide us through this most perilous period in the humans' existence. I have used 10 criteria to select the best candidate from the world's largest countries and organizations, which seems to be the European Union, despite its own problems.

"Democracy for a Human Federation" continues from where the previous volume ended, proposing **HOW** we can survive existential threats. We need two elements to achieve that: **Democracy** and a **Human Federation**. The key reason for an urgent deep reform of democracy is to minimize existential risks, including the risk of delivering a malicious Superintelligence, by priming it with new Universal Values of Humanity. This is also an absolute prerequisite for the federalization of the European Union, which is the first step towards **building a Human Federation**. This secondary theme presents three scenarios for the EU's federalization and an outline of the European Federation's Constitution.

"Becoming a Butterfly" is the third book in the series, asking **WHO** we may become after 2050, assuming we will survive existential threats. Its focus is on **Superintelligence**. This is 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.

Superintelligence will also offer an evolutionary path for humans to become **Posthumans**. The most important in that evolutionary paradigm shift will be the handing over to Superintelligence the set of human values, which best reflect who we are as humans. This must start right now and continue throughout the process of maturing AI to its 'adulthood', until we reach the moment when humans will pass control over their future to Superintelligence. From then on, humans may gradually merge with Superintelligence as individual digital Posthumans but acting in unison as a single superintelligent being.

INTRODUCTION

For the vast majority of us the most important goal is to live a long and healthy life, and prepare a similar life for our children, grandchildren, etc., ad infinitum. The problem is that there is no such infinitum. We take it for granted that, as a species, we will exist forever. Very few of us consider that over 99% of all species, which once existed, are now extinct. How can we be an exception?

This is the question I have asked in my first book "Who could save Humanity from Superintelligence?" (1) The paradox is that the probability of human extinction from natural causes (e.g. an asteroid impact) in this century is less than 0.00002% (2), whereas it is between 20% - 50% from man-made causes.

In my second book "Democracy for a Human Federation", I have identified democracy, as the key element needed for our civilization to survive man-made existential risks, including the risk of developing a malicious Superintelligence.

In this book, which is the final part of the *POSTHUMANS* series, I conclude that Humanity must not only manage various existential risks, which it has itself created, but also manage its own evolution. The analogy in the book's title is not perfect, since caterpillars and butterflies are the same species. However, what is almost identical, is the process of metamorphosis, which humans may have to go through, while evolving into a new species.

We may be the only species in the whole Universe, which is consciously capable of minimizing the risk of its extinction and control its own evolution in a desired direction. We have already been doing it in some way over millennia by controlling our evolution in a cultural and social sphere, which has also strengthened our resilience to extinction. But today we may also be able to control our physical evolution into a new species. We will only be successful in that evolution if we do it in stages, using a process of transformation similar to a caterpillar becoming a butterfly. I argue my case below.

1. The world has started to change in most areas at nearly an exponential pace. What once took a decade it can now be achieved in a matter of months
2. Apart from man-made (anthropogenic) existential threats for Humanity, such as biotechnology or a nuclear war, which can happen at any time, the most imminent risk facing Humanity is Artificial Intelligence (AI)
3. It is the technology, which is the driving force behind the exponential pace of change, and in particular a superfast development of the AI's capabilities. AI has already been delivering many benefits to us all and in the next few decades it may create the world of unimaginable wealth. That is the positive side of the AI we want to hear about

4. However, AI, as many other technological breakthroughs, such as nuclear energy or biotechnology (especially genetics), has also become a risk, probably the greatest existential risk that humanity has ever faced
5. **By 2025 we may already have first Transhumans**, who may be many times more intelligent than most of us and capable of planning their decisions years ahead more precisely than anyone
6. **By 2030 we may have an immature form of Artificial Intelligence** before it matures as Artificial General Intelligence (AGI), called in this book - Superintelligence. In particular, AI's self-learning and self-improvement capabilities, may lead to unwanted diffusion of the superintelligent skills from some specific domains into other, about which we may not even be aware. Therefore, any political or social changes have to be viewed from that perspective – **we have just about a decade to remain in control of our own future**
7. **By about 2030-2035 Artificial Intelligence** may reach the stage, when humans may no longer be able to fully control the goals of such an immature superintelligent AI, even by implementing most advanced control mechanisms. The most significant threat in this period is the emergence of a malicious Superintelligence, which may destroy us after we have lost control over its development. This risk trumps out other existential risks, such as climate change, because of its imminent arrival and in an extreme case, a potential total annihilation of the human species
8. **By 2045-2050 Superintelligence may reach its mature stage**, either becoming a benevolent or malevolent being. If it becomes benevolent, by having inherited the best Universal Values of Humanity, it will help us control all other existential risks and by then would have also created a world of abundance. If it becomes malevolent, it may make us extinct
9. Whatever happens, humans will either become extinct because of existential risks, such as Superintelligence, or evolve into a new species within this century. **It is impossible to stop the evolution**.

To summarize, in the next few decades our future may evolve in three ways:

1. **We may become extinct within a few decades** because some existential threats will combine, like pandemic, global warming, or a nuclear war
2. **We may become extinct in a few centuries** because some existential risks, such as a malicious AI, may trigger a **near** total humans' extinction with some humans surviving. A new civilization may then be built, reaching after some time our technological level. If it behaves in a similar way as we do now, it will trigger some existential risks becoming either immediately extinct or with only some humans surviving. That cycle may continue even for a few centuries, delaying the ultimate human species' extinction unless it changes its behaviour
3. **We take control of our evolution**, developing a friendly Superintelligence, which will adopt our human values, help us minimize existential threats and also create a pathway for human species' evolution.

I believe this third option is the only good alternative that is available to us. If so, what should we do, and what should be the priorities, assuming we only have just one decade to implement such an option? I believe we need to focus our efforts on three areas almost simultaneously, rather than in any particular order:

4. **Ensure a global regulatory governance over the development of AI**. To achieve that, we need to create an international organization, with a complete control on the scope and capabilities of AI.
5. **Carry out a deep reform of democracy.** That means starting with a review of key human values, rights and responsibilities and then modifying the way, in which we are governed as citizens of this planet. I will later explain, why it is so important in the context of our survival as a species
6. **Build a planetary civilization.** We cannot rely on the United Nations to fulfil that role and I justify that further on in the book. That requires a new organization that would act as a de facto World Government.

I realize how unrealistic these objectives seem to be, especially if we only have about one decade, by which time all three elements safeguarding the future of Humanity should be in place. However, sometimes only when a case is made quite bluntly, we get motivated to solve a problem.

I have been writing this book in the midst of the Covid-19 pandemic. The number of human victims, terrible as it must be for the families of the deceased, is miniscule in comparison with the Spanish flu of 1918-20, when about 3% of the world's population perished. This is equivalent to about 250 million people today. However, the indirect consequences of the pandemic are absolutely profound, showing how unprepared we have been as a civilization to face such a threat, not to mention more serious existential threats. Therefore, we should consider the side effects of this pandemic as a kind of a vaccination for our civilization to make us more resilient to existential risks.

If we survive unscathed this decade, there is a good chance that from about 2040 humans will be offered unimaginable material wealth and incredible opportunities for self-fulfilment. For those who decide to continue living in biological bodies, the only problem may be the one identified by Voltaire – lack of work. However, what he had in mind was probably different what you may think about, since he said: "work saves us from three great evils: boredom, vice and greed". Perhaps some of us may decide to pay for the privilege of having to work. Those who will already be Transhumans might complete their morphing with Superintelligence, beginning a new life as digital humans - Posthumans. Others may do it later, until all humans have morphed with Superintelligence. Humans may then become the only species in the entire Universe, which has engineered its own evolution.

Please note: Throughout the book, each part is preceded by a summary of the previous part, to give you an overall context.

1

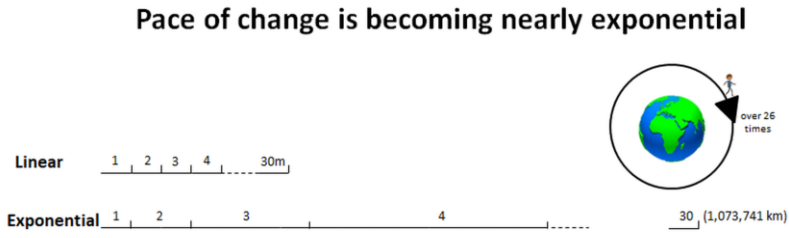
PART 1

EXTINCTION OR EVOLUTION?

Chapter 1
The Impact of Exponential Pace of Change

Living in the world of exponential pace of change

We have all been accustomed to living in the world where change happens at a linear pace. You walk 5km per hour, so it will take you twice that much to walk 10 km, and three times as much to walk 15km. But that is not the world of exponential change. This type of change is called exponential, because at each new moment in time (say every year), the value of what we measure (e.g. no. of electric cars or stem-cell originated meat production) would double. Ray Kurzweil, one of the most often quoted futurists, illustrates the difference between exponential and linear growth as follows. In the 'ordinary' world when I make 30 steps, 1m long each, I have covered 30m. In the world of exponential change, if you make 30m, then the first step will also be 1 meter long, however the second one, will be 2 meters long (twice that much) and the third - 4m long. When you have made 30 steps in this exponential world, you will have covered more than 1 million km, circling the globe over 26 times as I illustrate that in the diagram below:



Exponential change was first noticed in technology. Perhaps the best-known example is the Moore's law. This is an observation made by the former co-founder of Intel, Gordon Moore, that the number of transistors on a chip doubles approximately every 18 months. For nearly 50 years, that continues to be true. But in genomics, the price of genome sequencing falls much faster than exponentially, and even more so in the production of artificial meat.

Exponential pace of change is not immediately noticeable, it takes some time. However, at some stage, that change becomes suddenly considerable, accelerating very fast with each new unit of time. We may be reaching such a point right now when the exponential pace of change is starting to reach the so called "knee of curve". This is the stage at which an exponential trend can really take off almost vertically. Moreover, it does not apply just to technology. It can also apply to social, administrative, and political processes. Take for example

the time it took to submit your tax return, 20 or even 10 years ago. Today, it can be done in 20 minutes, rather than a few weeks as it was 20 years ago.

Exponential change has an impact not only on individual domains, such as biotechnology, but also on the pace of change and an overall progress in other related domains, e.g. in food production, using gene editing. This so-called convergence of technologies, results from the interaction between various parts of individual technologies. It creates new opportunities that speed up the pace of change much more than a single technology could have done on its own. We should also consider that what changes exponentially, is the access to various technologies for people that previously would have needed some technical background. For example, today most people 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.

Finally, consider this very latest statistics in the period of Covid-19 pandemic. In December 2019 Zoom video conferencing was used by 10 million daily users. In April 2020, that number rose to 300 million daily users, i.e. 30 times in just four months. That's what exponential change really means. Yes, Zoom belongs to the area of technology, but its impact has been felt in many areas, changing entirely the way we produce goods and interact with people, like in politics (parliamentary debates), education and of course in various industries.

So, from now on when you observe the world around you, including of course politics, then remember this:

What today takes 1 year, in a decade it will take a week

And my final thought on this subject. If it is so difficult for top experts in the field to see the impact of exponential change, how can then the decisions makers, such as politicians, change their policies, so that they better reflect the impact of exponential change?

Some likely effects of exponential pace of change

What are the consequences of exponential pace of change for our civilisation? From the current human perspective, perhaps the most significant are the changes outside the technological domain, e.g. in social and political domains. For example, China has reduced the number of people in permanent hunger by 600m in just 20 years. Life expectancy increases in some countries by about 6 hours every day, which means that every four years life on average is extended by one year. This trend will of course reach a biological barrier at some stage.

What effect can such an exponential pace of change have on an average person? Which of those effects may lead to positive and which ones to negative outcomes? In general, positive outcomes relate mainly to the unprecedented

technological capabilities that could greatly improve the quality of life in every corner of the world. At some stage, it may enable the expansion of the human race even beyond the solar system. Negative aspects may lead to a destruction of civilisation and the extinction of the human species within this century.

Unfortunately, in my view, it is far more likely that the negative consequences of technological and self-destructive social domains may prevail. Recognizing the trend and the pace of change in various domains, I would point to some likely negative developments due to our own reckless actions, rather than caused by natural causes. Most of them may have happened anyway, but the exponential pace of change increases their likelihood because of the lack of time to mitigate those threats and respond to them in a measured way. The list of areas where we may see a **negative** impact of exponential pace of change could be quite long, but let me just address these ones that may be affected by 2030:

1. The most serious and likely problem that humanity may face in this decade is the loss of control over the development of Artificial Intelligence (AI), which may happen earlier due to some significant breakthroughs
2. A decade ago, scientists expected that the planet will become irreversibly inhabitable if the global temperature rises over 3C, which they thought may be reached by 2100. However, we are already very close to the average global temperature increase by 2C, which has been causing severe disruption worldwide, like the fires in Australia in 2019
3. The world may face more periods of famine, worse than that of mid 1980' due to drought. This may cause mass migration exceeding that one of 2015.
4. Terrorism will be far more prevalent than in the previous years, due to easier access to some AI technologies, which may lead to severe conflicts between various states, blamed for supporting such actions, e.g. Iran
5. The spread of viruses more lethal a relatively mild Covid-19, similar in mortality to Ebola, will be easier due to mass tourism
6. Large scale conflicts between China, the USA or Russia are more likely. However, they will be conducted in a clandestine way, using unidentified cyber-attacks, rather than resorting to a nuclear, or chemical war
7. The risk of nuclear attack by terrorist groups, or rouge states, such as North Korea, will significantly increase due to the availability of cheap rockets
8. Local wars like those in Syria, or sub-Saharan Africa will demand faster response, and the presence of western soldiers in these countries
9. Human biological clones will almost certainly be borne if it has not happened yet. This may mainly present ethical problems of producing 'superhumans' much earlier
10. There will be thousands of humans with brain implants enhancing their intelligence, gradually transforming them into Transhumans, who will be 'smarter' thousands of times than an average human.

However, we may still be in a position over the next decade, or so, to alter the course of events if we change the way, in which we interact with each other at a

local and a global level. That must lead to some fundamental changes in the relationship between the governed and the governing, requiring a deep reform of democracy. We will also need a unified, global approach to fighting existential risks. The consequence of that will be some limitation of national sovereignty, as the price for a greater safety of all humans. Here is a list of some positive developments that may help turn the civilisation from its current course onto a more promising path.

Key **positive** developments, which we may experience because of exponential pace of change by 2030:

1. We will regularly start using super large quantum computers such as Google's Sycamore to solve most civilisation's difficult problems.
2. Within just a few years, rather than by 2030, we will reach the stage when a single personal computer, costing less than $1,000 will have more processing power and capacity than the human brain (following the Moore's Law)
3. It is highly likely that by the end of this decade we will eradicate cancer and many other diseases such as Alzheimer, arthritis, or Parkinson's disease
4. Because of that, by around 2030 we may reach a stage when life expectancy will increase by one year every year (predicted by Ray Kurzweil).
5. AI will be able to help us solve most problems in months or even days, which would usually take us years
6. AI will probably be the most valuable addition to human species, enabling us to have brain implants with access to the entire human knowledge delivered wirelessly via cloud. Whether such individuals will still be called human beings is a different matter. What is more important that gradually, human species will evolve to become another species, just following general principles of evolution.

Chapter 2
Earth – A Planetary Civilization Type I

What next for our civilization?

Physicists define civilizations by the energy level that could be available for its growth. In 1964 a Russian astrophysicist Nikolai Kardashev defined three such types of civilizations differing by the order of energy 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 (3).

- Type I civilization, also called a **planetary civilization**, can use and store all of the energy, which reaches their planet from its parent star. This is where we are now. Such a planetary civilization has energy consumption levels equivalent to the solar insolation on Earth (between 10^{16} and 10^{17} watts).
- A Type II civilization, also called a **stellar civilization**, can harness the total energy of its planet's parent star (the most popular concept is the Dyson sphere - a device encircling the parent star to capture its entire energy- 10^{27}W
- A Type III civilization—also called a **galactic civilization**—can control energy on the scale of its entire host galaxy – 10^{37} Watts

So, how have we progressed as a planetary civilization so far? The main function of human civilizations across the ages. i.e. nomadic, agrarian, industrial, and now post-industrial (digital) has been the creation of better environment and capabilities for the satisfaction of Humanity's needs. Until the industrial revolution, progress has been very slow and people living in the Middle Ages did not have materially, or even socially, better standards than in Roman times. However, over the last 300 years, with the invention of the steam engine, the acceleration of human progress in economic, social, and technological domains has been truly astounding.

Apart from that, since the time of the agrarian civilisation, humans have developed 'local' civilisations that differed mainly in the political and social domains. From a political perspective, today's very large countries, such as China, Russia, or the European Union, are effectively civilisations. From a social perspective, defined mainly by religion, Christianity, Buddhism, or the Muslim religion can also be defined as civilizations.

But even with today's fast communication, such civilisations cultivated for centuries, remain largely fossilized. Unlike the nomadic, agrarian, or industrial civilisations, they are not separated by time epochs but by culture, tradition and of course, values. No wonder, that people migrating from one such civilisation to the other may have problems in adjusting their inherited values. That is why such migrations may be stretching the relative cohesion of the society belonging

to a different civilisation, if migrations occur on a massive scale. This, coupled with an incredible speed of technological progress, of which some side effects may be dangerous, creates the basis for potentially global, catastrophic crises.

Today, we begin to see that it is not only democracy, which is in crisis but so is our entire civilisation, which we can destroy in a matter of days. Will this be the end of history? Perhaps. Alternatively, we may even survive extreme crises, and then gradually enter the era when humans will have to coexist with a superintelligent entity. That will be the start of a new, civilisation, rather than the end of history, as prophesized, in a different context by Francis Fukuyama in his book – The End of History.

There is no fast recipe for protecting our civilisation. However, if we want to avoid extinction in the very near future then there is one 'generic' thing we must do globally. We need to change our behaviour making the protection of our species as important as protecting ourselves as individuals, remembering that safety is in numbers. That will be a real social revolution, especially in the Western countries, but less so in China and South East Asia. It would require from us some sacrifice, e.g. limitation of national sovereignty and restrictions on our freedom. That will be necessary to create a truly planetary civilization under one Supranational government. However, if you consider how difficult it was for citizens in most western countries to accept the necessary limitation of freedom of movement during the Covid-19 crisis, then the prospect of achieving that goal seems almost unsurmountable.

It may become somewhat easier to gradually change our national view to a planetary view if we could feel more affinity with the people on the other side of the globe. That could be achieved by education and shared experience. An excellent example of how it could be done is the European Union's Erasmus Programme for exchange of University students.

In summary, for humans to evolve and become a civilizations Type II, we must:

1. Think globally, rather than nationally
2. Reform democracy
3. Federate the world
4. Merge with Superintelligence
5. Become Type II civilization

Vaccinating our civilization

In my second book "Democracy for a Human Federation", I call this decade 2020-2030 – the age of **Immature Superintelligence**. This is going to be, in my view, the most dangerous period in the history of mankind for various reasons. The prime one is the danger posed by some of the top ten existential risks (risk that would wipe out all humans), such as AI or natural and artificial pandemics.

Some events may not be existential as such, but when combined with severe drought, massive migration, severe economic and political events, then we may face an existential threat. So, that is the bad news we must deal with.

The good news is that if we survive relatively unscathed for the next 10 years, then everything may start changing very rapidly towards, what I call, the world of unimaginable abundance in every domain of life. You will find plenty of arguments supporting this view in this book. However, the question is how we survive this terrible decade. Here again, you will find a kind of a Roadmap further on in this book, proposing some steps to minimize the catastrophic risks facing Humanity.

So, apart from surviving the next decade, what else do we need to become a Type II civilization? In a nutshell, we need to become more resilient to various threats. Paradoxically, what we then need, is to have several quite severe global, but not existential, crises. They would prepare us for serious existential threats. That's why I consider we have been very lucky because just now we have the first such event – the Covid-19 pandemic. Why are we lucky? Because in relative terms, this virus is benign. The tragedy for the people affected by the loss of their kin, is of course unbearable. However, imagine if it had been a new strand of Ebola-type virus spread by air, (in 2014 the mortality rate was 50%). Direct effects of this virus are comparable to an ordinary flu (about 20,000 deaths in the UK every year).

However, indirect effects are going to be severe and long-lasting. And that is bad news indeed. I am thinking about an average Joe Bloggs and his family. The system, by which I mean the politics and economics combined, is totally unprepared for such a sudden collapse. The problem is that the politicians always think about their own future first and only then about the nation. The more populist a politician is, the more important for him is his self-preservation (e.g. Donald Trump).

But I hope this crisis will force many governments around the world, which will be under immense pressure, to make some fundamental changes in how we are governed and how capitalism works. There is no doubt that there will be millions of unemployed after the virus pandemic has passed. The world will no longer look as before. Many governments, like the proverbial king, will be naked – the fake news will be revealed for what it is to most people, who want to see it.

It may be too early to look into a crystal ball because the image is too blurred. However, here are some quite plausible positive log-term consequences of the current crisis:

1. The so-called Technological Unemployment due to large scale robotization will come several years earlier, than experts had thought. With millions of people becoming unemployed in the next few years (which may

paradoxically also help some business owners to recover some of their losses – that's how capitalism works!) we may expect severe tensions between the governing and the governed. This may enforce some profound political changes. Populism may be in retreat for a while, but it will be a small compensation for those in need.

2. The sectors most impacted will be the government services, retail, travel, banking, white collar workers in general and education. For example, universities and schools may introduce a system of working from home on a big scale. Students and pupils may visit a university or a school once a week (1/5 of the school children/students), suddenly enabling much smaller classes, smaller schools and perhaps even fewer of them. Probably the majority of employees in the banking, financial and legal profession will work from home. Who will then need those skyscrapers in the City of London, New York, or Tokyo, which are still being built! A typical Ponzi scheme, where those investors at the top will get their money back, but the future pensioners, as investors, will lose a lot. That may result in a deep overhaul of the capitalist system earlier than could have been predicted even last year
3. We can expect a lot of 'cuddling' amongst the Europeans, with a further EU integration leading to a federation made easier because it will become almost indispensable and less expensive that the current intergovernmental system. That will also impact Brexit and the negotiations with the EU on extending the current limbo status till the end of 2021. And that may open an entirely new game.

What should then the governments do apart from a better preparation for the next pandemic? Here are some suggestions:

1. Introduce Universal Basic Income. It was proposed in the UK Parliament by the Scots on 18th March, but of course it was not acted on by the UK government. An amount of £6,000 for all adults, £11,000 for all pensioners and £3,000 for all children and taxable, would immediately reduce the anxiety of millions of people and would cost no more extra than about 3% of the budget, replacing most benefits and tax allowances
2. Introduce a wide-scale job sharing to soften the impact of unemployment
3. Reduce the working week
4. Make the retirement age flexible from the of 50
5. Form governments of a National Unity, in countries, where there is no coalition
6. Tax all businesses at source, i.e. where they create products or services, i.e. Starbucks, Google, Amazon, Microsoft. G7 and G20 agreement would be needed, in which case, it would cover over 90% of the world's trade
7. Close all tax havens, again G7 and G20 agreement will be needed. It has been proposed but not acted on

8. Introduce the ‘Tobin’ tax, i.e. a tax on financial transactions, especially in micro-second share dealing (proposed originally by the EU but not introduced yet). This alone would offset a lot of governments’ future debt
9. Make a deep reform of democracy, so that the lies of the populists such as Trump’s or Johnson’s (e.g. how well the UK NHS is prepared for fighting such a disease as Coronavirus) will not put nations into such a mess, in which we have all found ourselves at the start of 2020.

Therefore, we should really relax and treat the current experience as a shot in the arm – a vaccination of civilization.

Chapter 3
Catastrophic and Existential Risks

What has changed in existential risks since the last century?

Note: *This subject has been extensively covered in Volume 1 of POSTHUMANS - 'Federate to Survive!' so the next three chapters provide only the very necessary contextual input.*

Estimates of the number of Earth's current species range from 10 million to 14 million, of which about 1.2 million still exist. If we consider that more than 99 percent of all species, i.e. over five billion species, that ever lived on Earth are estimated to have died out[7], why should we be the only species, which is an evolutionary exception. What are the risks that may destroy our civilization and all humans?

You may have already heard about some risks threatening our civilisation, such as climate change, a global nuclear war, an asteroid impact, or super volcanoes. But which of them are catastrophic and which are existential? Generally, these risks could be grouped as global catastrophic or existential (terminal) risks, depending on their scope and severity. A "**global catastrophic risk**" is any risk that is "global" in scope, lasting for some time (endurable) and hence may kill the vast majority of life on earth but humanity could still recover. The most recent Covid-19 pandemic, in biological terms has been relatively benign and would not even fall into this category. An "**existential risk**", on the other hand, is terminal for the entire human species. In its extreme case, it can destroy all human, non-human and even plant life.

An example of a global catastrophic risk that could destroy the world in a material sense and at the same time potentially eliminate nearly all humans is a **global nuclear war**. It could immediately wipe out most of the current material substance of civilization, i.e. towns, infrastructure, food crops, etc. and in the longer term, through radiation, lack of food and the emergence of post-nuclear winter, causing the death of nearly all people but unlikely the entire human race. A global pandemic is an example of an existential risk, which of course, the Covid-19 pandemic is not. However, **existential pandemic**s may also be caused by an accidental or intentional release from laboratories of a deadly virus, which may wipe out the human race in weeks but leave the infrastructure undamaged, at least for a few years. Saying that, at least in theory, some humans will survive because of some genetic differences. **Climate change though**, assuming nothing would be done to contain it, could in its extreme form lead within, say 200 years, to a literal extinction of every human being and most of organic life. Finally, the **development of a malevolent AI** is the most imminent and an extreme form of an existential risk that could lead to a complete extinction of a human species.

Furthermore, global catastrophic and existential risks could be divided into two broad groups: **man-made (anthropogenic)** risks, which could be mitigated (e.g. global warming and AI), and **natural (non-anthropogenic)** risks, we cannot control (e.g. an asteroid impact). **I will only consider man-made risks.**

Earth has experienced quite a few natural extinction events, which affected some species, such as the mass extinction of dinosaurs. Humans also experienced a near extinction event about 75,000 years ago, caused by the eruption of a Toba volcano on the island of Sumatra, when the human population sharply declined, though not to the level, which was earlier suggested (about 5,000 people) (4).

Until 1945, the year in which the first atomic bomb was exploded, humans were only facing natural existential risks. That year marked the dawn of a new epoch for Humanity. We have opened a Pandora's box by creating risks in various areas of human activity that may destroy us all, if such man-made (anthropogenic) risks are not properly controlled. Developments in several areas broadly linked to technology, may potentially be even more serious than nuclear weapons. For example, biotechnology and AI will greatly improve living standards, but also carry potentially serious downside risks.

What are the most significant man-made risks?

We have hardly any control over natural risks. But we do have control over political, social, economic, and technological risks. The man-made existential risks can be split into three categories, which may require different approaches:

Anthropogenic risks that are **immediate** and may become existential within days or even in hours, such as

1. Global nuclear war
2. Weaponized AI or cyber wars
3. Engineered pandemics and synthetic biology
4. Unknown risks, mainly technology-orientated

But there are also risks that may become existential progressively. These risks are called **progressive**, because their impact increases with time and they are not immediate yet. There are at least two types of such risks: Superintelligence and Climate Change. They both have some additional characteristics – these are the risks that will **certainly** materialize if we do nothing, as opposed to other risks covered here, which are of a 'lottery' type. The top three progressive risks are:

1. **Superintelligence** the risk, which may partially materialize starting from around 2030 and then quickly have a wider and much more severe impact
2. **Climate Change** whose impact may be severely felt in the middle of this century, although it may not be existential yet.
3. **Nanotechnology** and experimental technology accidents

The table below shows the assessment of existential risks by The Future of Humanity Institute, Oxford in 2008. Some of them, like nuclear terrorism, are not existential on their own, but only when combined with other risks.

Humanity's Top Existential Risks in 21st Century

	Risk	RISK (Probability *Impact) of human extinction by 2100 (%) from an expert survey 2008
	Overall Risk	**19%**
1	Superintelligent AI	5%
2	Weaponized AI	5%
3	Non-nuclear wars	4%
4	Engineered pandemic and synthetic biology	2%
5	Nuclear wars	1%
6	Nanotechnology accident	0.50%
7	Natural pandemic	0.05%
8	Nuclear terrorism	0.03%

Please note, climate change is not in that list, because it is not viewed as existential by the end of this century (it might be existential in 22nd century). The overall risk of human extinction by the end of this century was assessed as 19%. However, some scientists, such as the late prof. Stephen Hawking and a former Astronomer Royal, prof. Martin Rees, assessed that risk as higher than 50%.

I describe all these risks in more detail in the subsequent sections, but I cover the risk of Superintelligence separately.

Chapter 4
Immediate Existential Man-made Risks

Engineered pandemics

Biotechnology can pose a global catastrophic risk in the form of natural pathogens or novel, engineered ones. Such a catastrophe may also be brought about by usage in warfare, terrorist attacks or by accident. There have been very few biotechnology related terrorist attacks, of which the most well-known is the anthrax attack in Tokyo, in June 1993 by the religious group Aum Shinrikyo. It is believed, that exponential growth has been observed in the biotechnology sector and some scientists (5) predict that this will lead to major increases in biotechnological capabilities in the coming decades. They argue that risks from biological warfare and bioterrorism are distinct from nuclear and chemical threats because biological pathogens are easier to mass-produce (especially as technological capabilities are becoming available even to individual users).

Phil Torres is pessimistic on the measures aimed at mitigating the risks of artificial pandemics. In his article "How likely is an existential catastrophe?" (6) he says, "this trend is indicative of biotechnological development in general: laboratory equipment is becoming cheaper, processes are increasingly automated, and the Internet contains a growing number of complete genomes, including genetic sequences of Ebola and smallpox. The result is that the number of people capable of designing, synthesizing, and dispersing a weaponized microbe will almost certainly increase in the coming decades".

More risks stemming from novel, engineered pathogens, can be expected in the future. Scientists suspect that there is an upper limit on the virulence (deadliness) of naturally occurring pathogens (7). But pathogens may be intentionally or unintentionally genetically modified to change virulence and other characteristics. One example of that is what happened to Australian researchers who unintentionally changed characteristics of the mouse pox virus while trying to develop a virus to sterilize rodents. The modified virus became highly lethal even in vaccinated and naturally resistant mice. The technological means to genetically modify viruses' characteristics are likely to become more widely available in the future, if not properly regulated (8).

We should look at the danger of self-replicating synthetic, incurable viruses from a particular angle – the rogue researcher syndrome. One possibility is that a disgruntled individual might steal a virus and travel around the world releasing it. An important factor in the motives of such a person might be his religious or cult-like convictions that might, in his mind, justify the act (a mass murder, like ISIS, but on a global scale).

Global nuclear wars

A nuclear war between the US and Russia was the chief apocalyptic fear of the late 20th century. That threat may have reduced but, with proliferation of nuclear weapons, there is still a risk of a conflict serious enough to cause a "nuclear winter" as the smoke in the stratosphere shuts out sunlight for months. That could put an end to civilised life regardless of the bombs' material impact. Therefore, it is so difficult to assess the probability of global nuclear war ever taking place and even more difficult to tell if it will ultimately lead to a total collapse of civilisation. That's why Global Challenges Foundation Report 2017 puts the risk between 1 and 9.5% in this century.

The scenarios that have been explored most frequently are nuclear warfare and doomsday devices. Although the probability of a nuclear war per year is slim, Professor Martin Hellman described it as inevitable in the long run (9). During the Cuban missile crisis, U.S. President John F. Kennedy estimated the odds of nuclear war as being "somewhere between one out of three and even" (10). To put it in today's context, the United States and Russia have a combined arsenal of 14,700 nuclear weapons, and there is an estimated total of 15,700 nuclear weapons in existence worldwide (11).

However, I would observe that the use of a nuclear weapon today would be much worse for three reasons:

1. A typical modern nuclear weapon is now 8 to 80 times larger; modern society is much more reliant on vulnerable information technology and long-distance supply routes for food and fuel.
2. Modern society is heavily reliant on electricity to power central heating pumps, to provide water, information via TV, the Internet, and mobile phones. Nuclear strike will mean no water supply, no heating or lighting, no information, no mobile phone signal.
3. Only a few days of food supply exists in regional distribution depots. The supply network would fail for multiple reasons: road blockages, communications breakdown, collapse of the banking system, destruction of ports.

Weaponized AI

There are several definitions of a weaponized AI because such a new subject means many things to many people. However, in simple terms it means using AI technology as a weapon**.** There are two type of weaponized AI: **Soft Weaponized AI**, which uses software applications that achieve malicious objectives by usually compromising or blackmailing individuals through publication of documents, pictures or breaking into security systems, and **Hard Weaponized AI** that directs specialized weapons or equipment at a pre-planned target.

Soft weaponized AI

Let me start with the Soft Weaponized AI. Here, probably the best and most succinct list of what soft weaponized AI could do has been created by a well-known futurist Thomas Frey, who presents a simple scenario: "Virtually every situation presents an opportunity for a weaponized AI, but each will require different strategies, targets, and techniques. Once a clear objective is put into place, the AI will use a series of trial and error processes to find the optimal strategy. AI tools will include incentives, pressures, threats, intimidation, accusations, theft, and blackmail. All can be applied in some fashion to targeted individuals as well as to those close to them" (12). He gives a list of 36 examples, from which I have selected only the most significant ones.

4. **Hijacking a City**. Every city is made up of interdependent systems that function symbiotically with their constituency. Stoplights, water, electric, sewage, traffic control, tax collection, police, etc. Once AI can disable a single city, it can easily be replicated to affect many more.
5. **Destroying a Country**. At the core of every country are its financial systems. Weaponized AI could be directed to attack essential communication and power systems. Once those are disabled, the next wave of attacks could be focused on airports, banks, hospitals, grocery stores, and emergency services. Every system has its weakest link and this kind of exploitive weaponry could be relentless.

Hard weaponized AI

South Korea currently maintains the border with its northern neighbour using Samsung-built robot sentries that can fire bullets, so it's safe to say autonomous weapons are already in use. It's easy to conceive future versions that could, say, use facial recognition software to hunt down targets and 3D-printing technology that would make arms stockpiling easy for any terrorist. Robotic soldiers would only aim at specific targets. They will be so small and cheap that even an average earner (say a potential terrorist) could buy it.

However, an individual robotic soldier would not be a threat to Humanity. What may create an existential risk is a potential arms race in autonomous weapons and Artificial Intelligence. Such a race would expose civilians to undue, potentially existential risk. If autonomous weapons are developed and deployed, they will eventually be in the air, space, sea, land, and cyber domains.

Future soldiers

The unintended effects of creating and fielding autonomous systems might be extremely severe if they get under the control of malicious AI agents. In the worst-case scenario, nuclear war heads may be fired, almost definitely annihilating most life on earth, should all current nuclear arsenals be used.

Nuclear terrorism

A crude terrorist nuclear bomb the size of a football, detonated in the heart of a major city could have a devastating effect. It would release an equivalent of as much as 10,000 tons of conventional explosives. At the place of its detonation, the temperature would reach millions of centigrade combined with an intense burst of gamma and neutron radiation which would be lethal for nearly everyone directly exposed within about a mile from the centre of the blast.

This is a low probability, a non-existential risk on its own. However, if such an event happens at the same time as other catastrophic risks, such as pandemic or significant climate change events, it can become existential.

Psychopath dictators

This is an entirely new existential risk. Within a few years it will be possible for a single dictator of even a small state or a very rich individual to put the whole world on the brink of extinction, mainly because of the rising power and capability of AI.

Chapter 5
Progressive Existential Man-made Risks

Artificial Intelligence

A full description of this risk is covered in Part 3, Chapter 1. However, I will describe very briefly this risk here for completeness.

Artificial Intelligence (AI) presents quite a few risks even right now. For example, a malicious AI scientist could create a self-learning programme (it does not even have to be a team of robots), which may over time gather enough information (passwords and operational routines) to fire off nuclear weapons. But AI will gradually evolve into Artificial General Intelligence (AGI), called in this book, Superintelligence, which is defined as a type of AI that would surpass even the smartest humans in every domain of human activity. The main threat here stems from even the slightest misalignment of our values and the "values" or objectives of Superintelligence. The overall risk of developing a malicious Superintelligence that might make us extinct varies widely. It has been assessed as 5% over the course of this century by the Future of Humanity Institute, Oxford but as much higher by prof. Stephen Hawking or more recently, by Elon Musk.

Why does that represent the highest risk for Humanity? Because it is almost certain to happen, unlike natural pandemics, which may not happen at all, since it is a lottery type risk.

The second reason why this risk is so dangerous, is that it may happen much earlier than the risk mostly talked about in recent years – the climatic catastrophe. The risk coming from Superintelligence is more likely to happen in the next 50 years rather than in the next century. Should that risk materialize, it may lead literally to the extinction of the entire human species. On the other hand, if we manage to deliver the so called "friendly" Superintelligence, then instead of becoming the biggest risk, it will itself help us reduce other man-made risks, such as climate change.

Climate change

This is probably the most publicised existential risk, apart from a global nuclear war. Conventional modelling of climate change induced by human activity (adding carbon dioxide to the atmosphere) has focused on the most likely outcome: global warming by up to 4C. But there is a risk that feedback loops, such as the release of methane from Arctic permafrost, could produce an increase in temperature of about 6C or more. Mass deaths through starvation and social unrest could then lead to the collapse of civilisation. The most optimistic predictions estimate that the overall existential risk from extreme climate change

is about 0.01% annually, which would make it 1% over the entire century – not that much. The most realistic assessment was probably made in the Stern Report. It estimates such risk at 9.5% over this century. I have taken the median view, that the **existential** risk stemming from the extreme climate change **over this century is about 5% but only if combined with other risks.**

Martin Rees, the former Royal Astronomer, observes that many people still hope that we can sail towards a low-carbon future without trauma and disaster. He says that politicians won't gain much resonance by advocating a bare-bones approach that entails unwelcome lifestyle changes – especially if the benefits are far away and decades into the future. There are, however, three politically realistic measures that should be pursued. First, all countries could promote measures that actually save money – better energy-efficiency, better insulation of buildings and so forth. Second, efforts could focus on the reduction of pollutants, methane, and black carbon. These are minor contributors to global warming, but their reduction would (unlike that of CO2) have more manifested local side-benefits – especially in Asia. And third, there should be a step change in research into clean energy; why shouldn't it be on a scale comparable to medical research? (13).

There is plenty of coverage of the risks that are linked to climate change. I would not in any sense like to downplay that risk, since it is really multifaceted and not just limited to temperature rise, although this is the major source of the consequences of climate change. However, as I have already mentioned, by the time the climate change might really endanger human species and most other species on our planet, which is in the next century, our civilisation will either survive or will most probably be gone because of other risks. Therefore, we should put all our efforts to minimise the risks stemming from Superintelligence because if we make it benign and friendly, it will be our major hope for reducing or entirely eliminating other anthropogenic existential risks.

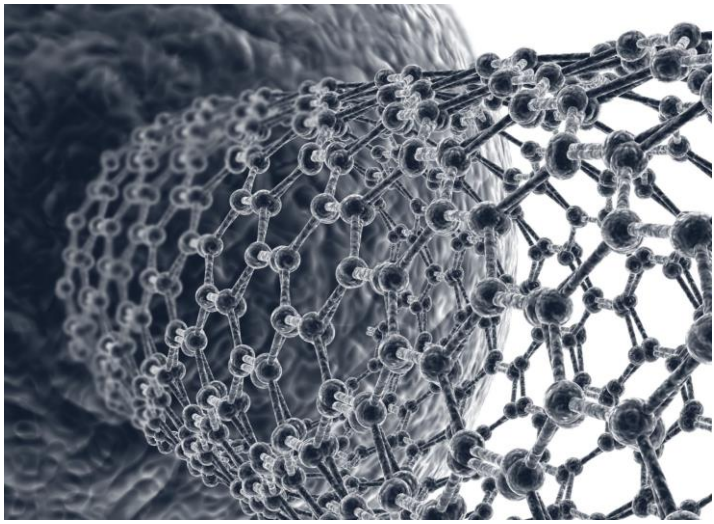
In a few decades we will have Superintelligence that will help us deal with this problem and many more. **The problem is not how to survive the climate change by the end of this century but how we can survive at all in the next 20 years.**

Nanotechnology incidents

Nanotechnology would allow us in principle to arrange atoms in any way we want. If this comes true, the world will forget what scarcity means because everything is made of atoms. This will move us into the future of abundance. As the philosopher Jason Silva puts it in "Future of Everything". It essentially makes the physical world a programmable medium."

One such scenario includes building the so-called nano-factory, a hypothetical device that could manufacture products with absolute atomic precision for a

fraction of the cost of current manufacturing. “Atomic precision” here means that two objects produced by a nano-factory, for example two computers of the same design, would be identical with respect not only to their macroscopic properties, but also the precise placement of their constituent atoms. It remains unclear whether nano-factories are physically possible, but if they are, as theorists Eric Drexler of the Future of Humanity Institute and Ralph Merkel of Singularity University claim, then the consequences for us would be profound.



Self-replicating nanobots Image source: https://archimorph.com/2007/07/24/self-replicating-nanobots

Molecular manufacturing requires significant advances in nanotechnology, but once achieved it will be possible to produce highly advanced products at low costs and in large quantities in nano-factories of desktop proportions. When nano-factories gain the ability to produce other nano-factories, production may only be limited by relatively abundant factors such as input materials, energy, and software. Being equipped with compact computers and motors these could be increasingly autonomous and have a large range of capabilities.

However, ultra-precise manufacturing on an atomic scale, which could create materials with wonderful new properties, could also be used in frightening new weapons. There is even a possibility to create self-replicating nano-machines taking over the planet.

Some experts in nanotechnology risk field suggest that the existential risk from this area comes from reaching technological advantage in the arms race through the availability of nanotech weaponry, which may destabilize the current relative balance between major powers. In addition, advanced nanotechnologies could introduce new nanoparticles to the biosphere, some of which could prove extremely toxic.

Chapter 6
Managing Our Own Evolution

How well have we managed civilisational crises?

So, now that you know what the main man-made existential risks, which Humanity faces right now are, you may ask what we can do about it. You have heard a lot about the risk that is not even in that list, Climate Change, because it is not in any sense an **immediate** existential risk in comparison with bio pandemics or Superintelligence. Here, some steps have been made, although supporters of the Extinction Rebellion say the actions taken are far inadequate. The question then is, what could force the world leaders to take the risk of existential risks very seriously and act on them decisively right now.

Before I answer this question, let me refer to some recent and some historical events, when large civilizational catastrophes were looming, show you the decisions that the world leaders then took, and how they profoundly impacted the fate of the world. I would call them **catastrophic risk-triggering decisions**. They have two outcomes: **they could have prevented** a risk becoming reality, **or they have triggered them off**.

1. Let me start with the WWII. In September 1938 the British Prime Minister Neville Chamberlain proudly showed off the Munich Peace Accord, authorizing Hitler to take over the Czechoslovakia's Sudetenland. That act of appeasement was thought to be enough to stop the European war, although it was already clear that Japan (invading China) and Italy (invading Abyssinia, today's Ethiopia) colluded with Hitler so that fascism could take control of the entire world. We all know what happened later – 3% of the world population lost their lives. It was the Munich Accord that became the risk-triggering event, and which started a year later the global war by Germany attacking Gdansk in Poland on 1st September 1939.

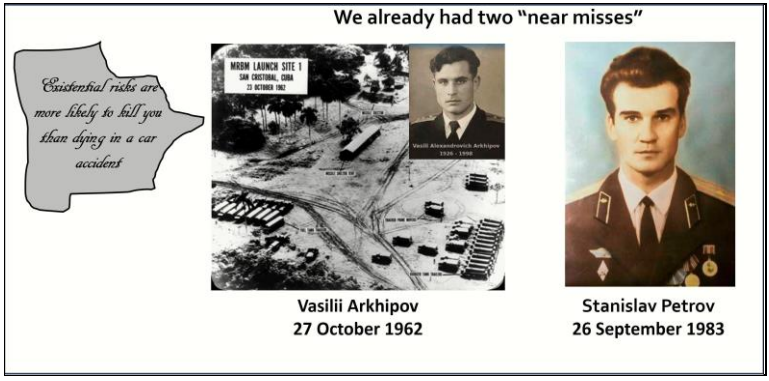
   Conclusion: *For the very first time it was clear from the outset that the world may have entered a period of another global war. However, since there was no World Government with sufficient powers, which could rein in Hitler (the League of Nations was even weaker than today's the United Nations), it was less than a year, when the global war indeed started.*

2. The second example is post-war Europe. For most of the nations affected by the WWII, the experience was so horrible and profound that in many countries the most common graffiti at that time was "No more war!" The former main European adversaries: Germany, France, Italy, and the Benelux first begun integrating their economies in 1952 (mainly the heavy industry) by forming the European Coal and Steel Community. Five years

later that integration included common democratic values, which gave birth in 1957 to the European Economic Community, now the European Union.

Conclusion: *Europe has learnt a terrible lesson. The risk-mitigating decision was to form the Steal and Coal Community rather than fight the Third WWW. Consequently, for the last 75 years, there has been no war in Europe, apart from the Balkan war in the 1990'. That was only possible by having a relatively powerful European Commission and the European Council, a pseudo European Government, perhaps the precursor of the European Federation.*

3. Another, a truly existential event, was the Cuban crisis in October 1962, when the world was just hours away from the breakout of a global nuclear war. That event is still considered one of the most dangerous moments in human history. The global nuclear war did not happen because the Soviet Union agreed to withdraw their missiles from Cuba and the USA withdrew their nuclear missiles from Turkey. In the midst of a total chaos on 26th October 1962, the Soviet and American leaders took the right decision.



A decade later, President Nixon and the first secretary of the Soviet Union, Brezhnev, signed the SALT 1 Treaty, which froze the number of intercontinental missiles. The subsequent START 1 Treaty of 1991 led to a significant reduction of nuclear weapons. Against all the odds, a nuclear war has so far been avoided.

Conclusion: The fact that the world has not experienced a global nuclear war was not because the UN prevented the conflict. It was because of a decision made solely by two countries the USA and the Soviet Union but the consequences of that decision (in this case positive) impacted the whole of civilisation. Here, the existential risk mitigating-decision that led to a long process of the reduction of nuclear missiles was to withdraw the Soviet missiles from Cuba and American missiles from Turkey, rather than fight a nuclear war.

4. The next event is the Climate Change conference in Katowice in Poland on 16th December 2018. The conference coincided with the IPCC's latest appeal to lower the CO2 emissions even further, so that the temperature growth will maximise at 1.5C, rather than 2C, as had been agreed in Paris. For me, the fact that the conference has agreed some concrete results, such as a unified system to measure the decarbonization of individual countries' economies, is quite a success. That was only possible because some of the dangers of climate change are now very clear, such as the fires in California and Australia, or extreme hurricanes in North America. However, even the conference in Katowice can only be deemed a success in relative terms (that something after all has been agreed). In absolute terms, it is a failure because the world should be acting much faster and much more decisively, which requires us to make some quite painful, mainly financial, decisions.

   Conclusion: The world has not been fighting Climate Change properly because to solve such a global problem in an efficient way we need to act globally. But that requires a strong Global Government and not such a weak incoherent organisation like the United Nations. Although the signing of the Paris Accord in 2015, seems to be a significant step in the right direction, it has already been proven that it is far from inadequate and requires much more commitment from the signatories of the Treaty to accelerate the process of a gradual decarbonization of the world's economy.

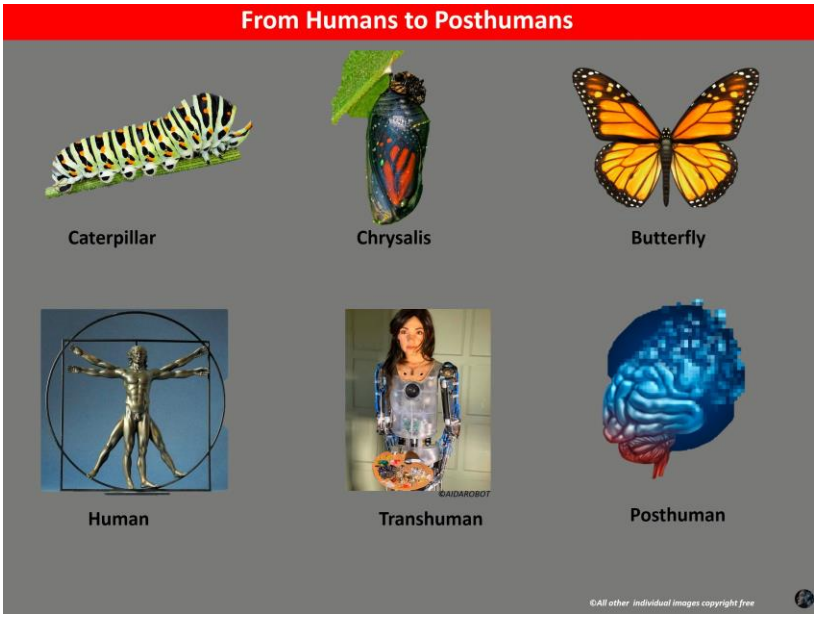
The man-made risks to which Humanity has been exposed so far have not been literally existential. However, if they had materialized, they would have been truly catastrophic.

Key steps for managing our evolution

The only way humans can avoid extinction is to successfully manage man-made existential risks. Some people may say nothing will exist for ever, even the Universe will die. Neither can humans survive in the long-term. After all, as has been already mentioned, over 99% of species have become extinct. But could a human species survive beyond this century. That depends on what we do and what we mean by 'survival'. After all, the oldest existing species in the world, Cyanobacteria, has existed for over 3.5 billion years, believed to be the Earth's oldest known life form.

However, humans are the most complex living beings and our survival in a biological form cannot last millennia, whatever happens, because of the scientific progress we have already made, mainly in AI. We cannot uninvent AI. If we survive beyond this century, many people existing today may already be living their digital lives (see more on that in Part 2). But we need to survive first, whilst our chances of becoming extinct as a species are between 20% to over 50% chance by the end of this century.

As mentioned earlier, the only chance for us to survive, especially the next 20 years, is to fundamentally change the way we live as a planetary civilization. Secondly, we need to manage the development of AI into a mature, human-friendly Superintelligence, which we may achieve by about 2050. This may help us not only to control existential risks but also enable our evolution into a new posthuman species, by morphing with the species that we have created ourselves. If some humans evolve into posthumans, it is possible that they will coexist with the original homo sapiens, similarly as we coexist with apes. In this sense we are a continuation of the original apes, while chimpanzees still exist.



By evolving into a new species as posthumans, we may be able to copy our digital brains and extend our existence both at an individual and a species level for millions if not billions of years. By becoming digital entities, we would simply build a unique evolutionary mechanism infinitely increasing our chances for survival. We would then be similar in an evolutionary survivability to the toughest organism on Earth - the Tardigrades. They are among the most resilient living life forms known, with individual species able to survive extreme conditions—such as exposure to extreme temperatures, extreme pressures (both high and low), air deprivation, radiation, dehydration, and starvation—that would quickly kill most other known forms of life.

So, Humanity may become extinct within the lifetime of some people already alive or it may evolve into a different species. If our civilization is to survive, we need to apply some powerful risk mitigation strategies. We have heard a lot about an existential threat of Climate Change. But as we have already discussed,

this is only one of about a dozen of such existential risks. Among them, the most severe is **the threat of Artificial Intelligence and especially, its mature form – Superintelligence.** This is an existential threat of an entirely different magnitude, which can either make our species extinct by a direct malevolent action, or by taking over the control of the future of Humanity. This risk is also different than for example the Climate Change, because it may come much earlier, within the next few decades. Secondly, we cannot stop (uninvent AI) – the proverbial genie is already out of the bottle.

We have already unknowingly entered the period, which I call the "Transition to Coexistence with Superintelligence". 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. To do that, as has been said earlier, we need to focus our efforts on three areas:

5. **Ensuring a global regulatory governance over the development of AI**. This is described in detail in Part 3, Chapter 3. It is of primary concern and requires urgent action, because after 2030 we may lose control over a maturing AI, with potentially catastrophic consequences for humans.
6. **Carrying out a deep reform of Democracy**, which would eliminate the most glaring errors in the system and make Humanity more resilient to existential risks. I will not cover in detail this area, highlighting only the most important aspects. For those interested in an in-depth proposal on how to reform democracy for the challenging period of coexisting with Superintelligence, I would suggest my earlier book: "Democracy for a Human Federation" (14)
7. **Building a planetary civilization**, which will prepare us for coexistence with Superintelligence. This subject is covered in Part 5 of the book.

A Mission for Humanity's survival

Humanity should have a kind of a **Mission**, accepted by a significant majority of nations, and based on the revised Universal Values of Humanity. That should be a point of departure for determining a strategy to avoid humans' extinction and prepare for our gradual evolution into a new species. For example:

Avoid extinction by evolving into a new species with a friendly Superintelligence

One of the preconditions for implementing such a Mission is the creation of a supranational organization that would be acting on behalf of all of us, as a planetary civilization (considering that the UN cannot realistically play that role). However, I believe it is too late for this option since it would have taken several decades to create such an organization from scratch with the required

powers. Realistically, we must accept that the world will probably not act in unison. 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. I have argued my case in my book *"Who could Save Humanity from Superintelligence?"* (1)

The organisation, which might fulfil that role most effectively seems to be the European Union, followed by NATO and as a fall-back option, by China. The support of other organisations, such as the UN, will be vital. Whichever organisation will lead Humanity, it should be guided by a **Vision** on how the Mission of Humanity could be delivered (at least regarding the process of maturing AI into Superintelligence), such as:

Maintain a global control of existential risks, especially Artificial Intelligence

To deliver such a Vision we must teach and instil in AI the best human values and preferences until its mature form – Superintelligence – becomes a single entity, millions of times more intelligent than humans and yet remaining our friendly partner. That process of maturing the current AI over a few decades should start straight away. Therefore, we must agree as soon as possible a kind of a Roadmap for Humanity's evolution, perhaps such as this one, containing five stages:

1. Maintain Existential Risks
2. Mature AI into a friendly Superintelligence
3. Reform Democracy
4. Make a transition to a federated world
5. Evolve with Superintelligence

The priority must be the supervision of the AI development, although the first four stages may be implemented almost simultaneously.

The endless evolution

I would like to finish this part with a dose of optimism, although for some readers it may still be a somewhat pessimistic outlook. It is a reflection on a continuous development of AI, as a chain of causes and effects, which ultimately leads to the transition of AI into Superintelligence. We may become bystanders or decision makers in that process, which we are already unable to stop. This also depends on how we control the development of AI, which I cover in detail in Part 3.

If we manage to turn Superintelligence into our friend and assuming we would still have control over its evolution (e.g. via linking its goals to our most

important human values), then the question is what our choice will be. For example, we can let it evolve itself and, so to speak, let it fly off and leave us alone. That may be possible. However, it is unlikely we will choose this option, since as someone said, 'one cannot uninvent the atomic bomb'. We are inquisitive and innovative beings and it is difficult to imagine we wanted to go back to 'less developed' world.

We must assume that once a mature Superintelligence arrives, it will become our Master by default, even if our values will have been embedded in its overall decision-making pattern. Its knowledge, choices of important decisions for humanity and overall comprehension of the world around us and the Universe in general, will be unimaginably greater than our own capabilities. So, we shall have two options (assuming there are both possible). The first one is to upload (copy) our personality, memories, and consciousness by reading our brain (although that may not be enough) onto a digital platform (a chip embedded in some material resembling our bodies). The second option is to merge Superintelligence with our bodies as an implant. From then on it is a pure speculation how Humanity might evolve into the long, long future. However, in the next few decades we may be forced to make that biggest decision in the history of Humanity on **how** we want to evolve as a species. Therefore, let me take you on my own journey into the unknown because the conclusions may be useful for making such a decision.

Most of you reading this book will live to see the advent of Superintelligence irrespective of it being conscious or not. Consciousness is a form of Existence, which is the opposite of Nothingness, Non-Existence, or Non-Awareness. We need to get the sense of what is the very nature of Existence and how Superintelligence can evolve as a potentially distinct species far into the future, aging within the aging Universe. But for the purpose of extrapolating the evolution of Superintelligence, we need first to go back and look how we, humans, came into Existence from Nothingness.

It is my very personal, perhaps a bit unusual, view of Nothingness and Existence. I remember well when in 2005, after a lecture at the London's Royal Society of Arts and Commerce (RSA), I spoke to the Nobel laureate in chemistry Harry Kroto. I asked him "How would you define 'Eternity'?" He was a bit surprised but then asked me if I had such a definition. I proposed this one: "Eternity is Nothingness on average", with which he did not disagree, since there are a number of theories that come to similar conclusions like the Dynamic Eternal Universe (15). I shall now use this as a starting point and list, in a simplistic way, the steps that led to the creation of humans. Once I arrive at the point where we are now, I will then list the next steps in our evolution to become a Transhuman species.

There are several theories, which describe how the evolution of the universe has begun. The best known is the theory of the "Big Bang". The model of our

Universe since the Big Bang is supported by many detailed, and in most cases, proven theories, such as quantum physics with its Standard Model, relativity, gravity, and the strings theory. There is still, however, no overall theory called the Theory of Everything that could reconcile gravity and quantum theory. Theories that try to explain what existed **before** Big Bang, use by and large, the current laws of physics to answer what a 'true' Nothingness could be and what properties it might have. The most prominent is the M-theory based on Supersymmetry that supports the so-called multi-verse theory (the existence of infinite number of universes) and the only theory (still not fully confirmed) that links gravity with two fundamental forces of nature: bosons and fermions.

For our needs it is enough to assume that whichever theory we use the only important issue is that most likely a "true" Nothingness does not exist. Nothingness has a property, and the current theories only need to prove that such a description of the properties of Nothingness is plausible.

This historical perspective of the emergence of 'Something' out of 'Nothing' is only meant to give you in just three stages a better appreciation of what a momentous event the birth of Superintelligence would be and how it could eventually evolve. The first stage gives a snapshot of how it all has begun, starting with just a quantum field, which pre-existed the birth of our Universe, and culminating with today's planet Earth. In stage two, I present my schedule for delivery of a mature Superintelligence. Finally, in stage three, I will present my personal view, based on dozens of scenarios published by particle physicists, astronomers, and futurists, on how the Universe with Superintelligence might evolve.

Stage 1: The evolutionary perspective on the emergence of human species

1. Nothingness 'has' no beginning and no end – it is eternal
2. Even Nothingness, as anything else, must have a property. It is believed that the property of Nothingness is a quantum field – "pure" energy
3. "Pure" energy manifests itself as Strings according to Supersymmetry or M-theory. Strings are one dimensional 'packets' of energy varying in size from below the Planck's scale (10^{-35m}) to the size of the Universe. They have charge, vibration frequency, gravitational force (gravitons) and mass, if crossing electromagnetic field or Higgs' field
4. Quantum field obeys quantum mechanics laws, among others the Heisenberg's Uncertainty Principle, which can sometimes lead to spontaneous imperfections in the field
5. Those imperfections appear "like waves on an ideally flat ocean's surface"
6. Like any waves, quantum field waves have troughs and crests, which are positive and negative energies
7. On average the sum of positive and negative energies is 0
8. So, Nothingness is not really "nothingness" as we understand it. Sometimes, when imperfections occur, it becomes "Something"

9. That's why I believe over Eternity, Nothingness is "true" Nothingness **on average**
10. Following the Heisenberg's Uncertainty Principle, the fluctuations in energy levels may lead to spontaneous creations of Big Bangs, creating material Universes
11. According to Multiverse Theory there could be infinite number of Universes emerging from those fluctuations, when the conditions are just right for converting energy into mass and creating a material world. Only some of them are stable enough, having the right constants (e.g. proportion of matter and antimatter, initial temperature, and gravity)
12. Our Universe was also started that way about 13.8bn years ago. It began with a period of so-called Inflation lasting between 10^{-35s} and 10^{-32s} when it expanded from zero size to the size of a grapefruit, ensuring the uniformity of the Universe
13. Since then our Universe has been constantly expanding leading to the creation of first stars after about 100 million years (the discovery made in 2018)
14. About 5bn years ago our Sun was created
15. 4.5bn years ago our Planet was formed
16. 3.5bn years ago life began on earth
17. A few million years ago man evolved from the animal kingdom
18. About 10,000 years ago our civilization begun
19. About 10 years ago, first artificial intelligence machine (IBM's Watson) beat humans in American Jeopardy game.
20. In 2016 Alpha-Go beat the Grand Master in the Go-Go game. It was built on the hypothesis that our mind organizes knowledge/intelligence in several layers. That discovery has sparked off almost faster than exponential progress in AI.

Stage 2: The Pathway to Superintelligence

So, this is where we are today. Below are my predictions on AI development and other key innovations that will ultimately lead to the emergence of Superintelligence:

2024 – A primitive cognitive AI Agent is aware of where it is, what it is doing, what it must achieve and what's its relationship with its 'master' (it understands this is the man who makes ultimate decisions). It passes an advanced Turing test
2025 – The first Transhuman with a fully thought-controlled brain implant can do some Google searches by thought alone and storing and later retrieving that information by accessing a private external memory and a processing unit
2026 – An AI agent with a wireless access to the fastest Supercomputer exceeds the intelligence of any human but in a single domain only. It can make hours-long unscripted conversations with any human on a specific subject area. It can also increase its understanding of other knowledge domains by self-learning or in conversations

2027– A Transhumans' team has been created that can communicate wirelessly using brain implants
2028 – Approved Transhumans control the development of Superintelligence
2030 - Approved Transhumans are elected to political bodies, such as the European Federation
2033 – An AI agent is created, which exceeds human intelligence in many domains and can discuss for hours on any subject with any human. It has a cognition level of a teenager. It can be unpredictable but is still controlled by Transhumans.
2035 – Transhumans can read each other's mind (in a broad sense), and if allowed, can also read any ordinary human mind (only knowing the subject area with no details). They can also control, if permitted, other humans' subconsciousness, by creating preferences in their minds (like hypnosis but could in principle be done against the person's will)
2042 – The first version of a mature Superintelligence is created, completely controlled by the authorized Transhumans. It has full cognition of an adult but may not be conscious yet (perhaps does not have to be in a human sense)
2045 – A fully mature Superintelligence is born, but which is willingly executing any requests from authorized humans. It is controlled by Transhumans, some of whom were elected as the leaders of the Human Federation
2047 – A Technological Singularity is born, fully controlled by Transhumans. Superintelligence improves its performance and intelligence almost exponentially
2050 – Superintelligence rules humans in every aspect of their life. The world of abundance has been created. Nobody must work. However, most people have severe mental problems caused by their inability to adapt to the pace of change.

What happens next is a pure speculation. This is my 'gut' feeling:

2060 – First successful mind of a Transhuman is upload to Superintelligence. A Posthuman species has been born
2070 – Posthumans retain a complete control over Superintelligence
2200 – Most biological humans have become Posthumans. The twilight of Homo Sapiens is nearly there

From today's perspective, what is important is to prepare for a selection of choices around 2040 – 2100, how we want our species to evolve. This is of course a big assumption that we will basically survive unscathed as a species, with no existential risk wiping us out, that the Superintelligence will emerge human friendly, and that we shall have enough control over it to make such a momentous decision. But there is still one more condition on which our choice will depend – can Superintelligence become conscious. What would be the consequences of developing Superintelligence that in principle (confirmed by future discoveries) could never be conscious because consciousness requires a biological substrate? That is the question I raise in Part 2.

2

Part 2
FROM ARTIFICIAL INTELLIGENCE TO SUPERINTELLIGENCE

Before you move on…

In Part 1, I have described how exponential pace of change can increase the likelihood of some of the existential risks happening. That is why our civilisation is at the most significant junction in the history of homo sapiens. If we do nothing to rapidly change the current direction of evolution where we fight for the best outcome of each of 200 states instead of behaving as a single planetary federation, our chances of survival beyond 2050 may be less than 50%.

I have made some initial signposting on the Road Map for Humanity's evolution. The first milestone is for humans to have a Mission to survive such as: '**Avoid extinction by evolving into a new species with a friendly Superintelligence**'.

That should be followed by the second milestone – how we should do it together. This is our Vision, e.g. "**Maintain a global control of existential risks, especially Artificial Intelligence**".

Finally, we should have an implementation plan. I suggested that it should consist of three objectives:

1. **Ensure a global regulatory governance over the development of AI**. This is a primary and most urgent action, because after 2030, we may lose control over a maturing AI with possibly fatal consequences for humans.
2. **Carry out a deep reform of Democracy**, which would eliminate the most glaring errors in the system and make Humanity more resilient to existential risks
3. **Build a planetary civilization**, which will prepare us for coexistence with Superintelligence.

In Part 2, I will introduce you to our future partner – Superintelligence, describing what it is, and which of its capabilities might be most useful for us.

Chapter 1
Intelligence – the Engine of Evolution

About intelligence

Before we discuss Artificial Intelligence, we need to look at how it differs from a human intelligence, and intelligence in general. Unfortunately, this is not a term that is easy to define unambiguously. Intelligence can mean many things to many people. The scientific community has been debating this since at least the late 19th century. Without going into a long discourse, *intelligence is a general mental ability to learn and apply knowledge to change the environment most effectively for the intelligent agent*, in our case – humans.

Intelligence does not have to be sentient (able to feel or perceive the environment with a degree of emotion). The same may be true about consciousness – it can be created or experienced in a different way than a human consciousness, however the result may be the same, or even better. After all, as Christopher Koch, the Chief Scientist at the Allen Institute for Brain Science says **'Intelligence is about behaviour'**. For example: what do you do in a new environment to survive? **Consciousness is about being'**.

The main reason why we are better in performing most tasks than other mammals is mainly due to the relative size of our brain. If we can expand our brain externally, then we will also expand our intelligence, probably disproportionally more, if we have faster processing available, e.g. quantum computers. Our expanded 'brain' will be able to have memory stores not only in any place on the planet Earth but also in space.

Recently, some scientists rejected the idea of a single intelligence and instead have suggested that intelligence is the result of several independent abilities, which when combined contribute to the total performance of an individual. That would include other "intelligences" such as the ability to:

1. evaluate and judge
2. reason and have abstract thoughts
3. learn quickly as well as learn from experience
4. comprehend complex ideas
5. have the capacity for original and productive thought

Robert Sternberg, a psychologist, proposes that there are three fundamental aspects of intelligence: analytical, practical, and creative. He believes that traditional intelligence tests only focus on one aspect – analytical – and do not address the necessary balance from other aspects. Some people still believe that intelligence is the sole attribute of humans. But these days it is enough to watch

a few films e.g. from excellent BBC collections, to see how intelligent some animals or even birds can be.

So, intelligence is not a unique attribute of humans. Nor does intelligent processing of information have to be carried out in the same way as it happens in humans. For example, Google search engine is millions of times faster, more accurate and precise than a human brain. Try to ask a human the same question as in Google search. The same is true about Google Translator, face recognition, autonomous car driving etc. Intelligence better than human is already here. Although it is still largely sectoral, this is changing fast.

On 3rd October 2017 a test was organized for several AI assistants by three Chinese researchers: Feng Liu, Yong Shi, and Ying Liu, primarily based on exams carried out during 2016. According to researchers, Google's AI Assistant rating of 47.3 is barely beneath a six-year-old human's IQ of 55.5. However, it was more than double that of Siri's IQ of 23.9. Siri is also behind Microsoft's Bing or Baidu, which have respective IQs of 31.98 and 32.92 respectively. All AI's IQs are considerably lower than a mean for 18-year-old's, which is 97.

The researchers say, that: "The results so far indicate that the artificial-intelligence systems produced by Google, Baidu, and others have significantly improved over the past two years but still have certain gaps as compared with even a six-year-old child" (16).

The difference between the grades of AI seems to be quite significant. But once it gets to the sixth grade, AI will improve exponentially until it becomes Superintelligence. So, what is Superintelligence?

What is Artificial Intelligence and Superintelligence?

Probably most people still think that **Artificial Intelligence** (AI) is just Information Technology (IT) under a different name. That may be one explanation why potential benefits and threats of AI are misunderstood so much. At its very basic level, IT and AI are similar in that they may be guided by goals and the use of computers to calculate the most effective way of achieving that goal. But that's where similarities end.

The biggest difference between IT and AI is that AI has, among others, the capability to self-learn and continuously improve (on its own) its performance. AI also has to some degree, the ability to perceive and judge a given situation or tasks, similarly, as humans do. So, AI's actions and judgment are more and more 'human-like' rather than 'artificial', as AI matures in its capabilities, by analysing and interpreting situations in the way like humans. In general, an AI agent delivers **probabilistic answers**. IT on the other hand, uses rules based on Boolean logic (AND, OR, NOR, NAND), which can only deliver a **binary answer**.

AI has been applied in earnest for at least 30 years under various other names such as Expert Systems, and later on, Neural Networks. Its key features are super performance and imitation of human cognitive abilities like problem solving and learning or speech recognition. Currently, it can beat best human capabilities but usually in one discipline only. Therefore, it is termed as **a "narrow AI".**

AI researchers are unanimous in their view that humans possess the unique ability of a general intelligence. It is the ability to reason logically about complex problems, understand and manipulate the information provided by the environment, and adapt to changing surroundings. AI has made some steps in that direction. For example, in 2011, the IBM Watson computer system competed in Jeopardy game, against former winners Brad Rutter and Ken Jennings. Watson won the game and the prize of $1 million. Then in March 2016 Google's AlphaGo computer using self-learning (machine learning) program, beat the 18-time world champion Lee Sedol. The success of self-learning has sparked a real revolution in AI.

Progressively, a narrow AI will advance its capabilities until it becomes **Artificial General Intelligence (AGI**). AGI computer programs, known as 'agent-based computing,' is aimed at carrying out the tasks, which are normally accomplished by a single human. Gradually, AGI will become capable of doing the same things that humans do, but faster and better. Ultimately, they will be capable of assuming human characteristics, like kindness, emotion, and abstract thinking.

Superintelligence is generally understood as a synonym of AGI, and I have used this term, best popularized by Nick Bostrom in his book 'Superintelligence', rather than AGI. However, for me Superintelligence has a slightly different nuanced view in terms of scale. Most AI specialists consider AGI as an 'agent' superior to any humans, in every aspect of intelligence, judgement or action. That is of course still possible. But for me, it is something more.

I interpret Superintelligence, as a **single entity**, with the same superior powers, mentioned earlier but being not a robot or a human-like agent, but an invisible being. We can visualise AGI agents as a kind of a computer in a human-like form or any other hardware. Whereas Superintelligence is an invisible network run by billions of algorithms, like billions of neurons in our brain. It is in fact an **Artificial Mind**. We can see the brain, but we cannot see the mind.

Chapter 2
How to Create Superintelligence?

A cookbook for creating Superintelligence

To understand how Superintelligence may interact with humans in the future, we need to know what it might be made of, and what it might look like. Let me start with the first aspect. The scientists working in that area come from different disciplines exploring the ways, in which we could build a superintelligent agent. It is probably the most interdisciplinary 'project' that Humanity has ever contemplated. We have started the creation of something that one day may become a new species.

So, how could we create Superintelligence? All ingredients to create it are already here, apart from cognition and (eventually) consciousness. Below is your shopping list. While the ingredients are real, the projected numbers are just an approximation of what might be available in about 10 years, when we may produce, what I call the Immature Superintelligence.

I use the period of 10 years as a kind of a threshold. That is based on various forecasts, but mainly on the most well-known – Ray Kurzweil's prediction. He says that by 2029 we shall have computers, which will pass a full Turing test, making it impossible to tell whether the 'person' you are talking with is a human or a machine (some scientist believe it has already happened on video). If we apply the so-called Moore's law, which predicts that computing power doubles every 18 months, then it is possible to calculate that all computer-related processes will improve by at least 100 times in 10 years, which means a powerful home computer will have the capacity far exceeding the power of a human brain. To achieve that we will need the following components:

1. **Data**. We will need a lot of digitised data like text, pictures, videos, or sound. Goggle's database is currently about 10 exabytes (10^{19}) bytes. By 2030 we will probably reach hundreds of zettabytes ($10^{21)}$ bytes. Will that be enough for Immature Superintelligence?
2. **Processors.** These are super gigantic computers with immense power to process super-large databases. The fastest one, Summit, is currently at Oak Ridge National Lab in the USA, which can perform 200 quadrillion calculations per second ($2*10^{17}$). In 10 years, the fastest computer, might be processing even 10^{20} times more information, which may be the very bottom edge of what is needed to have a fledgling Superintelligence. However, since quantum computers will be million times faster for some calculations than a conventional computer, this looks as a very conservative estimate. As I mentioned earlier, that has already happened. Google's Sycamore quantum computer consistently performed certain mathematical

calculations in 200 seconds, which would have taken the Summit computer 10,000 years. This is what exponential change means. Even while writing this book, some predictions, which are expected to materialize in a decade, have actually become a fact today.

3. **Memory**. We need plenty of this, at least xenottabytes, i.e. 1,000 more than the data itself, if we include the backup.
4. **Cables and communications infrastructure.** Today we already have fibre optics, laser, radio waves, 5G for smart phones and WI-FI, which with further improvements in speed, could be good enough for connecting various part of this immense web, including the satellites. By 2024, Elon Musk's Starlink constellation of 12,000 satellites will provide access to the Internet from any part of the globe, for a negligible fee.
5. **Interfaces.** These are devices for the input and output of information, sometimes also called nodes. We already have them today. These can be computer keyboards, screens, printers, scanners, mobile phones, industrial robots, Tesla cars, satellites, humanoids such as Sophia or Erica, and finally – holograms, perhaps the most likely interface for Superintelligence.
6. **Sensors and Neurons.** Sensors are feeding information on the state of objects or systems, like temperature or pressure. There are thousands of various types of mostly digital devices in use today. Neurons are also a type of a sensor. They react probabilistically, rather than in a binary way. We have already built artificial neurons, but they are too big and that's why their density is far inadequate. However, quantum computing will enable the building of super microscopic neurons in the next few years.
7. **Cognition**. Computers can now recognize faces and the environment, e.g. Tesla cars can differentiate between various types of objects. Our smart phones can speak clearly and translate seamlessly, with superb elocution skills in 56 languages using Natural Language Processing (NLP), used by Watson, Sophia, Alexa, Google Assistant or Bixby. However, they have no clue what they are talking about. So, today we have made some progress but there is still a long way to go to reach the level of a human cognition.
8. **Intelligence.** This is really about the algorithms – making sense of relations between data, finding patterns, and drawing conclusions that enable actions. We are nowhere near even to the level of intelligence of a mouse yet. That, combined with a human level cognition, is what we will need to deliver a mature Superintelligence. That's the biggest barrier a few decades yet.
9. **Consciousness.** Most AI specialists believe it is not a necessary condition for Superintelligence to be conscious. Some think it would be nice to have. Others fear about the consequences of dealing with a conscious being, which will also be millions of times more intelligent than us. We even do not know whether consciousness is possible in non-biological entities. If it is, then it might happen because Superintelligence discovers itself some algorithms and very peculiar neural connections that would make it conscious.

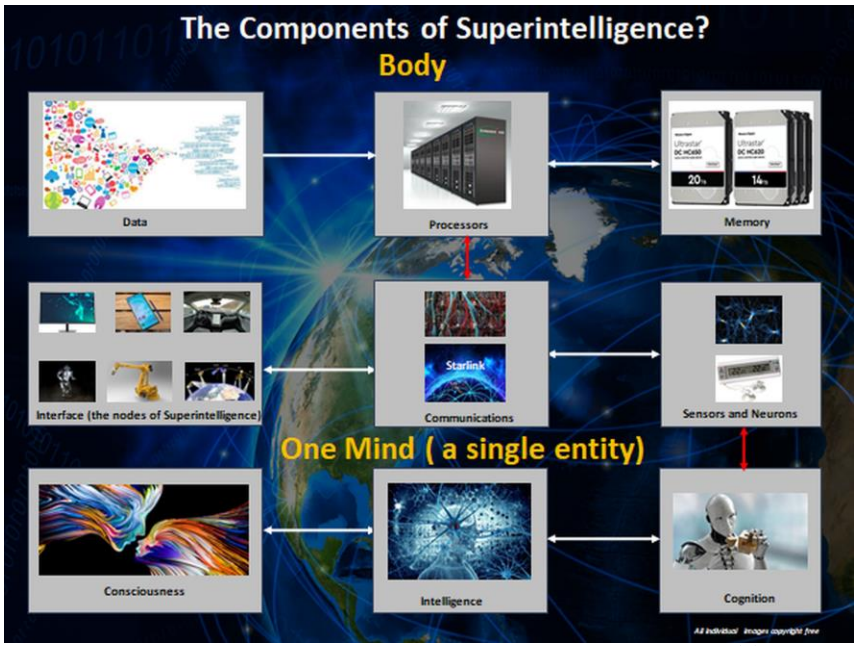


Image source: Tony Czarnecki – Democracy for a Human Federation [14]

All these elements listed above, apart from cognitive function and consciousness are available today and are connected and presented to us via Google and the Internet with its communication network. But the Internet is mostly just an infrastructure. What really matters are the resources and capabilities. This is the 'soft' side of the Internet, where the browser, like Google, is the king, supported by Facebook, GPS, Instagram and other information processing and communication applications. This is the 'intelligence' that we use every day, like Google's translation service, ability to help us drive using satellite navigation in almost any part of the globe, provide access to trillions of pictures, videos or documents in an instant, from any point of the planet, 24 hours a day.

How about the brain? To what degree do the current computers match the brain's capability? There are three areas involved in processing any information either by a human brain or a computer:

- **Storage.** An average PC has about 1TB of storage. That compares with about 100 billion neurons, each having about 10,000 connections, which translates into 1 petabyte of storage – about 1000 times more than your desktop PC.
- **Processing speed.** Neurons' clock runs at a frequency of about 200Hz. Your computer runs at a frequency of at least 2Ghz, i.e. at least 10 million times faster. Additionally, an electric signal can travel the neuronal network with a maximum speed of about 100m/s. A signal in a digital

computer can of course travel with the speed close to the speed of light, i.e. 300,000 km/s

- **Efficiency.** A human brain consumes about 20% of the body's energy. It is here, where a human brain is still about 100,000 times more efficient than an equivalent computer for certain tasks.

Ultimately, computers are not a clear, overall winner yet. Humans and computers have their own advantages, depending on a specific category. Computers are much better than humans at precision and raw information processing speed. But in energy efficiency and creativity, humans are still better. However, by the end of this decade, a single AI agent will have a combined capacity of about 100 human brains. Should it be combined as a seamless cognitive agent with senses (neurons), we would have a very slowly reacting prototype of Superintelligence.

I call this lowest level of Superintelligence, the Immature Superintelligence, described in Part 3, Chapter 1. By then it will already have a reasonable sense of cognition but still incomparable with humans' capability. To achieve that, it will take approximately 20 more years, so that by about 2050 we shall have a mature cognitive Superintelligence with the combined capacity of at least 1 billion human brains. What it may still lack is consciousness. However, as mentioned above, some scientists believe consciousness is irrelevant for an intelligent agent for making decisions.

Visualizing Superintelligence

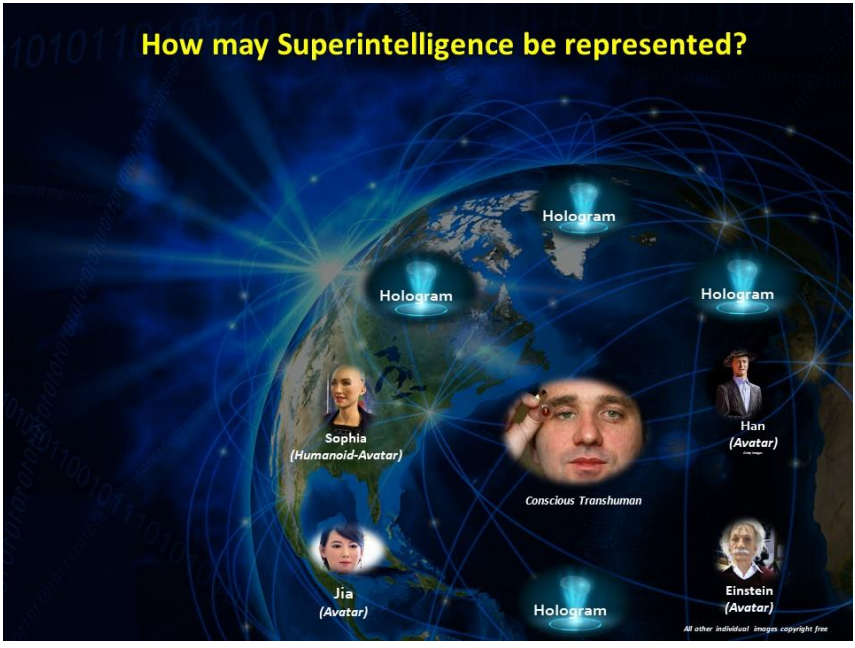
Now, let us deal with the external 'look' of Superintelligence. Today, your computer, is just one of billions of nodes to Google. Similarly, in the future, Superintelligence will have billions of nodes, through which it will be able to see, recognize and talk to anyone at once in any part of the planet. But most importantly, the collected knowledge and experience of living among humans will be available to any individual node, which will have the required access rights. This process of gathering and accessing the collective knowledge is already happening. This is how Amazon operates when we purchase some goods, almost immediately proposing us new articles that we may have never thought about, but which we might need.

Similarly, Alphabet's (Google's) Waymo – a driverless company, as well as Tesla, collect each car's unusual 'experience' and the way the car reacts, and makes it available to all Waymo's or Tesla's cars. What has been already happening and expanding at a lightning speed is the creation of a Pool Intelligence. That is how Superintelligence will operate in the future as well. That Pool Intelligence will become its collective intelligence available to all licenced agents and nodes.
Not all nodes will have the same rights of access and decision making. Most of them will be passive like yours, or my computer. They will be just for asking

questions and interacting at the lowest level of access rights. That will probably be somewhat like how we operate Google's search engine right now.

However, some of these purely digital nodes, very few, will have special access to the overall goal settings and decision-making on behalf of the entire Superintelligence. I call them **Digital Governors**. They will be like current Digital Assistants but with superpowers for controlling Superintelligence. Should we allow this to happen, then when such Superintelligence emerges, then humans will lose control over its decisions. The whole of Part 3 is dedicated to the problem of controlling AI development and delivering a benign, friendly Superintelligence. One of the most promising options is to control Superintelligence from within by authorized Transhumans, selected around 2040 by the future Human Federation. They will be called **Human Governors** and I describe them in detail in Part 3, Chapter 5.

Digital Governors, living their 'lives' as entirely digital entities, may have different representations in the 3D world, depending on the situation. This may include Superintelligence being represented as humanoids in an artificial skin, with muscles, face etc, which one would not be able to differentiate from a human being. By being digital, they will be able to copy themselves and thus appear the same in many parts of the planet, at the same time, as a personified Superintelligence. However, most likely they will be represented by holograms, enabling them to instantaneously create themselves and be in any part of the planet simultaneously, interacting with various audiences.



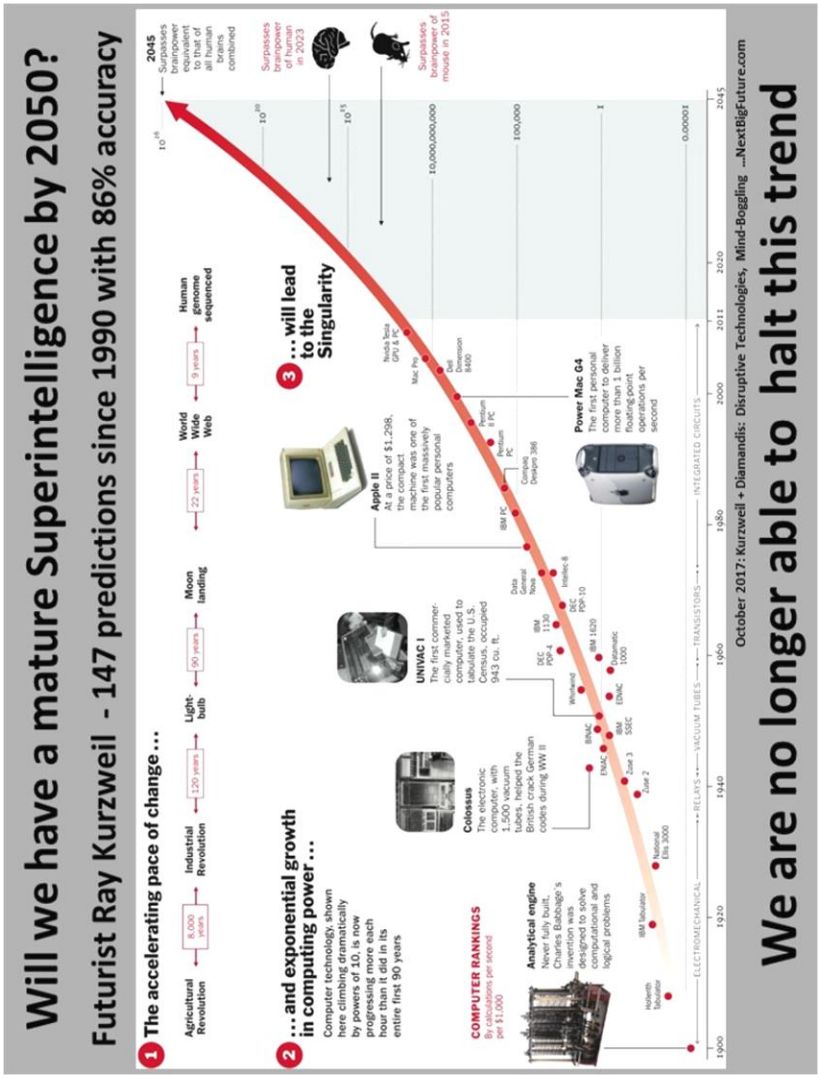
We will be able to communicate with Superintelligence, represented as humanoids, in a similar way as we communicate with other people and with some digital assistants such as Alexa or Siri. The communications between Superintelligence and biological humans will most probably depend on the level of authority a person would have.

Here is the summary of my visualization of Superintelligence, which includes assumptions, and possible features of Superintelligence that will have to be considered when developing today's AI into a future Superintelligence:

1. Superintelligence must be based on human values, rights, responsibilities, preferences, and expectations regarding future evolution of the humankind
2. They should be listed in a universally agreed (almost impossible) 'Last Will of Humanity' that would be communicated to Superintelligence before it takes an ultimate control over the future of human species
3. If possible, it should be clear whether Superintelligence must or mustn't be conscious, describing the consequences for each choice
4. Superintelligence will most likely be a single entity millions of times more intelligent than any human, probably conscious (unless we will have means to determine otherwise), consisting of billions of individual cells
5. Cells will differ in their capabilities, size, and power.
6. Some of these cells may be far more intelligent than any human, will have a power source, memory, processors, connectors to other cells, backup facility etc.
7. Some other cells will be the 'hubs' serving individual Transhumans and some may serve as a copy of a human mind
8. Each digitized human mind (a special cell) may have the right to have a 3D 'representative' - a humanoid body, which will enable a digital posthuman to have a human experience
9. The humanoid bodies, which will probably be unconscious, might be rented or owned outright by a specific digital human
10. All the cells, the network, the hardware, software, and the power supplies, will together constitute Superintelligence – a single superintelligent being
11. All decisions of Superintelligence are likely to be made by the majority voting of the authorized 'cells'
12. Some of the cells may be present in space or their backup copies may be there, for example in Geostationary orbit, the Lagrange orbit, on the Moon or even on Mars
13. The hardware cells serving as backup copies of human minds, could be brought down to Earth or into a lower GEO, or LaGrange orbit by space lifts
14. By 2100, such a Superintelligence may be millions of miles wide, increasing the reliability (backup), although for a few centuries, Earth will most likely be its main base.

A mature Superintelligence by 2050?

Four polls conducted in 2012 and 2013 showed that 50% of top AI specialists agreed that the median estimate for the emergence of Superintelligence is between 2040 and 2050. In May 2017, several AI scientists from the Future of Humanity Institute, Oxford University and Yale University published a report "When Will AI Exceed Human Performance? Evidence from AI Experts", reviewing the opinions of 352 AI experts. Overall, those experts believe there is a 50% chance that Superintelligence (AGI) will occur by 2060.



Independently, Ray Kurzweil, probably the best-known futurist and forecaster, whose success rate of his 147 predictions in the 1990s was 80% correct, claims that Superintelligence will emerge earlier, by 2045, as illustrated above.

Even more interesting is the fact that Kurzweil's predictions have been quite steady over the last 20 years, while other experts' opinions have changed significantly. Here are some examples:

- In the 1990s Kurzweil predicted AI will achieve human level intelligence in 2029 and Superintelligence in 2045, whereas most AI professionals at that time believed it would happen earliest in the 22nd century. So, the gap between Kurzweil's and other AI experts on the emergence of Superintelligence was at least 150 years.
- In 2000s, AI experts' predictions indicated that AGI will most likely be achieved by about 2080 but Kurzweil still maintained 2045 as the most likely date. The gap was 35 years.
- Recently, as in the above Report, 50% of 352 AI experts predict Superintelligence is most likely to happen by 2060. Since Kurzweil's still predicts 2045 the arrival of AGI, the gap has narrowed to about 15 years.

Additionally, human intelligence will not improve significantly, whereas AI capabilities will improve following the Moore's law, every 18 months.

There are of course AI researchers that have an almost opposite view, saying that Superintelligence (AGI) will never be built or that it will never surpass humans in all its capabilities, although they agree it could and does already outperform humans in some areas, being completely ignorant in most others. That is, for example, the view of Kevin Kelly, who quotes 5 myths of AI in his article 'The Myth of a Superhuman AI'. According to Kelly, these myths are:

- Artificial intelligence is already getting smarter than us, at an exponential rate
- We'll make AIs into a general-purpose intelligence, like our own
- We can make human intelligence in silicon
- Intelligence can be expanded without limit
- Once we have exploding Superintelligence it can solve most of our problems.

Then he challenges those myths with the following assertions:

1. Intelligence is not a single dimension, so "smarter than humans" is a meaningless concept
2. Humans do not have general purpose minds, and neither will AI
3. Emulation of human thinking in other media will be constrained by cost
4. Dimensions of intelligence are not infinite
5. Intelligences are only one factor in progress.

Without going into a detailed discussion, it is sufficient to say that his challenges are rather weak. Only the first challenge has some merit. However, it may also be incorrect. I do not think that most AI scientists believe that Superintelligence (AGI) will be a single dimension intelligent agent. A single dimension AGI is simply AI, i.e. artificial intelligence that may be more capable than humans in one particular area, like in a car navigation. However, the pathway to a smarter than human intelligence will lead through the multitude of intelligences all uploaded into one gigantic knowledge silo, which with a suitable set of algorithms will attain multidimensional intelligence millions of times better than that of humans.

Observing the progress of AI, it is obvious that machines will surpass human intelligence not at any single point in time, but rather gradually, and in more and more areas of intelligence, until one day they will become immensely more intelligent than humans in all areas. Challenge 2 – I disagree, human have a general-purpose intelligence – that's so obvious. Challenge no. 3 can be dismissed outright – when we have a mature Superintelligence, the cost of most things will be close to zero. Challenge 4 is irrelevant. Challenge 5 could be true, but it depends what one means by 'progress'.

Predicting future meaningfully, so that it would motivate us to prepare for it properly, can only be done with allocating certain probabilities. In 1969, a Russian dissident writer Andrei Amalrik, wrote a book 'Will the Soviet Union Survive Until 1984?' with a clear reference to Orwell's 'Nineteen-eighty-four' novel. He was wrong by 7 years – the Soviet Union collapsed on 26 December 1991. However, he was not wrong about the trend and the probability of a certain scenario happening around that time.

Singularity

At some stage we may arrive at the point called Technological Singularity or simply **Singularity,** briefly mentioned earlier. This is the point in time when Superintelligence will have at its disposal such a range of novel technologies, exceptional scheduling, and organizational capability that it could rapidly become matchless and unrivalled in what it can do in every field of life. Nick Bostrom says it will be able to bring about almost any possible outcome and be able to foil virtually any attempt that might prevent achieving its objectives. It could also eliminate, if it chooses, any other challenging rival intellects. Alternatively, it might manipulate or persuade controlling humans to change their behaviour towards its own interests, or it may merely obstruct their attempts to interfere.

Superintelligence will reach that stage through its self-learning capabilities and a series of self-improvement cycles until it achieves the so-called "runaway stage". Each new and more intelligent generation will appear more and more

rapidly, causing an intelligence explosion and resulting in a powerful Superintelligence that is simply impossible to imagine.

In a positive scenario, the world will then be transformed beyond recognition by the application of Superintelligence to humans and/or human problems, including poverty, disease, and mortality. The key problem related to the risk coming from Superintelligence and particularly from Singularity is that we may lose control over its objectives, intentions, behaviour, and attitude towards humans. The so called 'Control Problem' will occur when technology advances beyond our ability to foresee or control Superintelligence, which will be upgrading its potential at an ever-faster pace.

To achieve Singularity, we need to make at least three major improvements. Tim Urban suggests how this could be done:

- **Increase computer power**. This has been doubling every 18 months following the so-called Moore's Law. Some people think the "Law" will stop working about 2030 because we will reach physical limits of continuing miniaturization of chips. On the other hand, some futurists, such as Ray Kurzweil, predict that by around 2025 intelligence packed into a $1,000 computer, should reach the power of a human brain. But even that seems to be a fairly moderate prediction in view of exceptional acceleration in the development of a quantum computer. So, that condition can be most likely met.
- **Emulate human brain using reverse engineering.** There are several ways to do that. One that Urban suggests involves a strategy called "whole brain emulation," where the goal is to slice a real brain into thin layers, scan each layer one by one, use software to assemble and accurately reconstruct a 3-D model, and then implement the model on a powerful computer. Recently we have been able to emulate a 1mm-long flatworm brain, which consists of just 302 total neurons. The human brain contains at least 100 billion neurons, each having on average about 10,000 connections. If that makes it seem like a hopeless project, remember the power of exponential growth.
- **Emulate human brain by copying the process mastered by evolution**. This could be done using a machine-oriented approach, not by mimicking biology exactly. To do that we would build a computer that would have two major skills: doing research on how it could improve itself and then coding changes into itself (that's exactly what evolution did to us). We would teach computers to be computer scientists, so they could bootstrap their own development. That would be their main objective, finding the most effective process to make themselves smarter.

When Superintelligence reaches the Singularity stage, its core (the decision centre) will be most likely invisible, omnipresent and what may be most challenging for us to accept – may become conscious. However, it may be

embodied in countless of human-like bodies or avatars and holograms to experience real life and communicate with us, the mere mortals.

The arrival of Superintelligence triggering technological Singularity by 2030 is rather unlikely. However, I would agree that an existential threat may be created by then by an Immature Superintelligence. It might be stupid and error prone in most human skills and especially emotions. However, it would be enough that it is superintelligent in just one or two areas, but its lack of a common sense may make it very dangerous for humans. This is especially true about judgements that require the application of human values. So, what is the risk of AI, an Immature Superintelligence and Superintelligence?

Chapter 3
A Conscious Superintelligence?

What is consciousness?

I do not wish to delve too deeply into this extremely complex problem, as this is not the core subject of this book and neither am I competent to discuss it thoroughly. However, we need to look at least superficially into this area, to establish the implications of a conscious Superintelligence on the future of human species.

If intelligence is about behaviour than at the most general level, **consciousness is about being aware**. That awareness, however, is of a different degree in different living beings (and some people claim even in every object). Such an approach to the nature of consciousness provides a clearer view on several questions in this area. For example, it allows for a gradual development of consciousness over millennia of life's evolution, which might have started with automatic, chemistry-based responses, in plants 'seeking' best nutrients, i.e. being somewhat aware of the environment. In the animal world the level of awareness and then self-awareness would be directly correlated with the brain size relative to the body mass and complexity of neural connections. When these two parameters reached a tipping point, human consciousness was ignited in Homo sapiens and other humanoids, such as in the Neanderthal Man.

But let me start with some accepted definitions of four levels of ever-increasing consciousness, which may clarify somewhat the subject area:

1. **Awareness** is knowing what is going on around – plants respond to a stimulus
2. **Sentience means having the ability to have feelings and experience sensations** such as pleasure and comfort or pain and suffering. Most animals are sentient, and what may sound incredible, even some plants are, like the Venus flytrap (insect eating plant) and at least 600 other species of animal-eating flora [17].
3. **Self-awareness. This means knowing that you are 'you'.** The most well-known test for self-awareness in animals is what is known as 'the mirror test'. If an animal recognizes itself as the object it sees in the mirror, it means it is self-aware. Humans over the age of 18 months, dolphins, elephants, and great apes are all self-aware
4. **Consciousness**. This means being aware of own thoughts and having the ability of abstract thinking. Defining it more precisely is difficult and may be controversial.

Despite centuries of efforts to define consciousness it remains as elusive as ever and yet it is fundamental. Elusiveness is linked with the fact that there is no evidence our consciousness exists at all, because simply there is no location where it can be found in the brain. However, neurologists know, which regions of the brain are necessary for having a particular conscious experience. For example, the occipital lobe at the back of your skull interprets the signals sent from your eyes as vision. However, we still do not know for sure why do we need consciousness at all.

The predominant consensus in neuroscience is that consciousness is a property of the brain and its metabolism. When the brain dies, the mind and consciousness of the human to whom that brain belonged also ceases to exist. Therefore, without a brain there can be no consciousness.

Competing concepts of consciousness

Two-dimensional consciousness of Collège de France

It is the AI world that proposes radically new concepts of consciousness and creates experiments where these concepts can be tested. For example, many computer scientists think that consciousness involves two stages:

- accepting new information, storing, and retrieving old information
- cognitive processing of all that information into perceptions and actions.

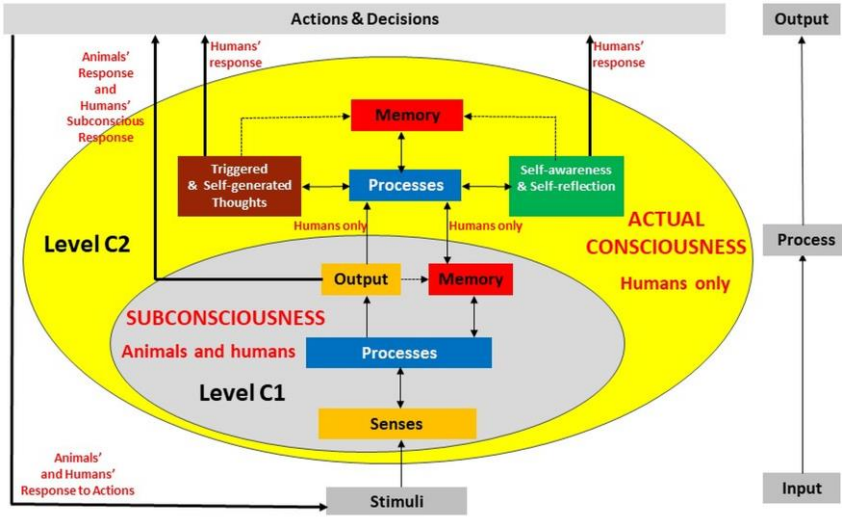
If that's right, then one day machines will become conscious. That view has recently gained some significant support. In October 2017 (18) Drs. Stanislas Dehaene, Hakwan Lau and Sid Kouider from Collège de France in Paris came to a conclusion that consciousness is "a multi-layered construct. As they see it, there are two kinds of consciousness, which they call 'dimensions C1 and C2 (see the diagram below). Both dimensions are necessary for a conscious mind, but one can exist without the other.

Subconsciousness – **dimension C1**, is the one that can exist without actual consciousness. It contains information and a huge range of processes with the required algorithms in the brain. That is what enables us to choose a chess move or spot a face without really knowing how we did it. The researchers believe that this type of consciousness has already been represented in a digital form and is comparable to the kind of processing that lies behind AI algorithms, which are embedded in DeepMind's AlphaGo or Chinese Face++ programs.

Actual consciousness – dimension C2 containing and monitoring information about oneself, which splits it into two distinct types, and which is not yet present in machine learning. It is the ability to:

- **Maintain a huge range of thoughts at once,** all accessible to other parts of the brain, making abilities like long-term planning possible
- **Obtain and process information about ourselves**, which allows us to do things like reflect on mistakes.

I have converted that definition into a diagram below. Looking from a wider perspective it shows very clearly that consciousness is the processing centre of external stimuli to actions and decisions. It is an ever-present model of reality: **Input-Process-Output**, applicable to all forms of matter and energy.



A likely organization of the human consciousness as proposed by the researchers at the Collège de France (Stanislas Dehaene, 2017– graph by Tony Czarnecki

As can be seen in the diagram, the second level of consciousness is only present in humans and to a lesser degree in some mammals. This is a unique type of consciousness, which can generate stimuli within its own 'processing centre', which often leads to multiple loops between an initial self-generated thought compared against the past (the memory) and processed again, until the final decision or action is made. It concurs quite well with a definition of human consciousness proposed by an American particle physicist Michio Kaku who in his book 'The Future of the Mind' (19): **'**Human consciousness is a specific type of consciousness that continuously evaluates the past to simulate the future to make a decision to achieve a goal**'.**

Conscious Electromagnetic Field Theory (Cemi)

Among the leading theories of consciousness this one is my favourite, simply because it can be tested. If Stanislas Dehaene et al, are right about the nature of consciousness, then the final element that we need is to present some supporting

evidence how such a process can manifest itself in nature. Fortunately, the most recent research in the nature of consciousness provides some indication that we may be very close to identifying the nature of consciousness, how it emerges and possibly how it might be developed in an inorganic being. It is the Conscious Electromagnetic (Cemi) Field Theory (20), proposed mainly by Johnjoe McFadden (I provide a simplified version here).

It suggests that every time a neuron fires to generate an action (through a changed electrical potential in its synapse) that signal is then cascaded down the line to thousands of other neurons. The overall result is that **synchronous firing creates an electric current, which then becomes the source of a disturbance in the surrounding electromagnetic field.** Should that current be sufficiently strong, every time such a set of neurons fires off, it creates the same electromagnetic field.

Recently, as reported in the 'Nature' magazine in October 2019, National Institute of Standards and Technology (NIST) in Boulder, Colorado, is one of a handful of groups trying to develop a 'neuromorphic' hardware that mimics the human brain in the hope that it will run brain-like software more efficiently. Thus, by using 'neuromorphic' hardware or a similar technology that mimics the human brain, it might become one of the ways to create a conscious Superintelligence. But even if neuromorphic devices would not produce conscious Superintelligence, their immense processing capabilities would make the arrival of Superintelligence even earlier than it is currently assumed.

Panpychism

In a broad sense, **pansychism proposes that every object from an atom to a human being is conscious to a degree**. According to the philosopher Philip Goff, it is the ability to experience the world in some way, e.g. to feel pain or pleasure, to see sights or hear sounds. For an electron it would be a sense of pull to a proton. But how about computers and processors used currently for developing AI, do they have even some degree of consciousness? Apart from panpsychism, none of the theories of consciousness would consider computers ever becoming conscious. Even the neural networks that contain digital neurons made of classic chips cannot be conscious if we take this assumption.

Integrated Theory of Consciousness

This theory has been recently gaining attention and the backing of some eminent neuroscientists ([21]). It says that absolutely every physical object has some (even if extremely) low level of consciousness. This is as controversial as panpsychism. It starts from two basic observations about the nature of our conscious experiences as humans. Firstly, that each experience we have is just one of a vast number of possible experiences we could have had. Secondly, that multiple different components (colours, textures, foreground, background) are

all experienced together, simultaneously. Given these two observations, the theory says that brain activity associated with consciousness must therefore be ever-changing, consisting of many different patterns, and involving a great deal of communication between different brain regions.

Hence the attempt to find **a formula that can give a precise "level of consciousness" from detailed data**. The theory claims that the ultimate formula will somehow quantify the information, which that 'something' contains. The result will be a number, the higher the number, the higher the level of consciousness and object, or a being has.

Consciousness is a property of Universe

At least this is the view of a prominent neuropsychiatrist Dr. Peter Fenwick. He has been studying the human brain, consciousness, and the phenomenon of near-death experience (NDE) for 50 years. He was initially very sceptical about the NDEs and related phenomena. However, Fenwick now believes his extensive research suggests that consciousness persists after death. Moreover, he believes that consciousness actually exists independently and outside of the brain. **Consciousness is as an inherent property of the universe itself like energy or gravity.**

In Fenwick's view, the brain does not create or produce consciousness but rather filters it. He provides some analogies that bring the concept into sharper focus. For example, he says that the eye filters and interprets only a very small portion of the electromagnetic spectrum and the ear can register only a narrow range of sonic frequencies. Similarly, according to Fenwick, the brain filters and perceives only a tiny part of the cosmos' intrinsic "consciousness."

Quantum mechanics theory of consciousness

Dr Fenwick's view corresponds well with Quantum Mechanics Theory of Consciousness, created by Roger Penrose, a prominent British particle physicist. He says that **consciousness is not just the brain-mind construct but is also underpinned by phenomena like those present in quantum mechanics**.

Consciousness may indeed behave in a similar way to some aspects of quantum mechanics' Uncertainty Principle in that it would not act at the level of individual synapses but rather at the level of neural networks, which connect millions of synapses and give an **averaged** response at a macro-level, e.g. lifting a hand. Since that response would be only probabilistic and not based on a binary state of an individual synapse (and hence its similarity to quantum phenomena), it will ensure that there is no conflict with the preservation of free will. Such a view will be in some respect in line with the research by the above-mentioned scientists from Collège de France in Paris on a two-dimensional nature of consciousness. That is also in line with the thinking of people like Raymond

Tallis, a prominent British neuroscientist, who strongly defends the validity of free will (i.e. unpredictability of human actions).

Is a digital consciousness possible?

And that's how the scientific world may be slowly arriving at some common understanding on the nature of consciousness, and by extension, the feasibility of uploading a human mind, together with its consciousness, to a superintelligent being. So, can Superintelligence be conscious? I would not rule it out, which seems an obvious conclusion from the previous chapter. There are already many robots, which are not only aware of the task they need to do, but perhaps even be self-aware, as reported by Science Robotics in January 2019.

"Columbia Engineering researchers have made a major advance in robotics by creating a robot that learns what it is, from scratch, with zero prior knowledge of physics, geometry, or motor dynamics. Initially the robot does not know if it is a spider, a snake, an arm—it has no clue what its own shape is. After a brief period of "babbling," and within about a day of intensive computing, their robot creates a self-simulation. The robot can then use that self-simulator internally to contemplate and adapt to different situations, handling new tasks as well as detecting and repairing damage in its own body (22).

If you are not convinced yet, find it on YouTube and watch with amazement. Yes, unfortunately, I might say, we already have produced self-aware robots, which in theory can self-learn from scratch if not everything, then really a lot, like teaching themselves to write a text in English, by holding a pen. I said 'unfortunately', because most of the AI scientists have not been prepared for that, not to mention most people.

So, robots are already self-aware. But are they sentient? Do they have any feelings or experience of joy, sadness, or any other emotions? Here we enter the field of **emotional robots.** Some psychologists and philosophers, like Dr Joel Smith and Lydia Farina from the University of Manchester give some convincing arguments.

They start by asking a question, how we decide. They argue that it depends on what we care about and how much, and on these issues people will differ. To care about *A* more than *B* is to assign *A* rather than *B* more significance in one's life, i.e. a greater value. They believe that autonomous, intelligent robots are no exception. If they are to make decisions, to successfully navigate the world, they must assign different values to different objects or things. In short, they conclude - **robots must be emotional** (23). But how far have we progressed in that area?

Probably the greatest advancement in applying emotional response of robots to people around them can be found at care homes, mainly in Japan and in Europe. One example is ENRICHME, a mobile robot designed to help older people with

a range of tasks, from exercising to remembering where they have put things. It was tested in retirement homes in several European countries to see if it could help combat cognitive decline and improve the residents' quality of life. Users not only accepted the robot, which helped them be more cognitively and physically active, but also appreciated some of its other functionality such as helping them find missing items. This type of robot is part of a new field known as "ambient assisted living". It uses technology to create surroundings, in which elderly patients feel safer, more independent and with a better level of privacy.

There are also quite a few advanced 'emotional' robots for kids, like Mabu, and of course Pepper and Assimo (the most expensive robot so far, costing about $2m, sadly no longer produced). However, these robots can read our emotions and respond in an emotional way, mainly by turning their head and eyes in a coherent way with the expressed emotion, but they still do not 'feel' themselves that emotion. And that is a very critical barrier. Therefore, **we have not created fully sentient robots yet**.

However, like with awareness, I believe that a fully emotional robot will be with us very soon, by combining sophisticated awareness with learning, e.g. what does it mean to be sad or happy, a substitute for serotonin in humans. Incidentally, Ray Kurzweil, probably one of the best-known futurists, predicts that sentient (fully emotional robots) will be with us by 2029, once they have passed the so-called Turing test. This test would be carried out on humanoid robots, undistinguishable from humans, expressing emotions and talking in a perfectly natural way. Should such a robot be undistinguishable from humans, they would pass a full Turing test.
So, we do not have sentient robots yet. But how likely is it that sometime in the future we will produce conscious robots? Well, here your guess is as good as mine, or any specialist in the field. **My own view is that we will have robots with human level consciousness by the magic date 2050**, when AI becomes Artificial General Intelligence, i.e. Superintelligence. I say this because I believe that like with intelligence, in which AI is today in some areas millions of times faster and more efficient that us, Superintelligence may gain consciousness in an entirely different way than humans did. If you look at the diagram I have drawn above, based on a new theory of consciousness by Dehaene et al, this might be the way robots will mature into a conscious entity. If we accept this notion, then human consciousness can be replicated in a different than a biological medium, such as in 'silicone'.

So, I am inclined to agree with those AI scientists and neuroscientists that AI, once it reaches the level of Superintelligence, will be conscious. This is becoming more likely as some of the world's most powerful supercomputers are designed to be neuromorphic, resembling the way the brain works. These computers will use artificial neurons, which produce probabilistic results, and use an entirely new architecture than your PC, or any other computer, departing from the 1940's von Neumann's architecture, based on 0,1 binary choices. Some

of these neurons can now be created entirely from bio-compatible materials and fuse seamlessly with a human brain via a skull-implanted chip.

Depending on how one looks at it, a conscious and empathetic Superintelligence may become a potential problem or an opportunity. It may become a problem because the decisions it will make, are likely to be probabilistic rather than based on binary choices, and that may lead to errors in judgment with potential catastrophic effects on humans. On the other hand, should such a conscious being also have emotions, it may behave in a more 'human' way, than an emotion-less, unconscious intelligent agent. Although it is only a supposition, I would assume that a conscious Superintelligence might be more likely to become friendly towards us rather than an antagonistic, or even a malicious one.

Therefore, the main assumption taken in this book is that once Superintelligence emerges, it will be possible for it to become conscious. Overall, it might be a better outcome for Humanity's future.

3

PART 3
MATURING A FRIENDLY SUPERINTELLIGENCE

Before you move on…

In Part 2 we have covered the basics of intelligence, Superintelligence, and consciousness.

I emphasized that the future Superintelligence will not be millions of superintelligent robots. It will rather be a single being, controlling almost everything via a huge network of connected superfast computers, neurons, robots, sensors, and memory stores. This will be the 'Brain' of Superintelligence. The algorithms (a set of rules with calculations about how to solve a problem) and integrated special programs, will be its 'Mind'.

The last subject we discussed was consciousness. It is very relevant to our future model of coexistence with Superintelligence, assuming it will be our friendly partner, rather than a malicious entity. I summarized four levels of ever increasing consciousness: **Awareness** - the lowest level of conscious perceptions, knowing what is going around, even some plants have it; **Sentience** – the ability to have feelings and experience sensations – most animals have it; **Self-awareness** – knowing that you are 'you' - apes and dolphins have it; and **Consciousness** – being aware of own thoughts and having the ability of abstract thinking

This is a necessary background knowledge you need before we move to the next Part, which will discuss how we can control AI, so that its maturing process will deliver a friendly Superintelligence.

Chapter 1
What is the risk of Superintelligence?

Risks arising from the development of Superintelligence

One of the main components of the risk of developing a malicious Superintelligence are human values. You may ask how this is relevant to delivering a safe and friendly Superintelligence. Paradoxically, Superintelligence forces us to answer the questions about key values that define us as humans, what is good and what is right, more meaningfully than ever before. If AI is developed without adequate human control, and then reaches the level of Superintelligence, humans may be in a real danger. Nick Bostrom quotes a scary example that involves a Superintelligence programmed to "maximize" the abundance of some objects, like paperclips. This could lead Superintelligence to harvest all available atoms, including those in human bodies, thereby destroying humanity (and perhaps the entire biosphere). In addition, there are multiple ways that Superintelligence could become totally malevolent towards humanity, as University of Louisville computer scientist Roman Yampolskiy outlines in his 2016 paper:

- Preventing humans from using resources such as money, land, water, rare elements, organic matter, internet service or computer hardware;
- 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;
- Constructing a total surveillance state (or exploitation of an existing one), reducing any notion of privacy to zero – including privacy of thought;
- Enslaving humankind, restricting our freedom to move or otherwise choose what to do with our bodies and minds, as through forced cryonics or concentration camps;
- Abusing and torturing humankind with perfect insight into our physiology to maximize amount of physical or emotional pain, perhaps combining it with a simulated model of us to make the process infinitely long.

It would be impossible to provide a complete list of negative outcomes that an AI agent would be able to inflict with only some general cognitive ability. We can expect a lot of these sorts of attacks in the future. The situation is even more complicated once we consider the systems that exceed human cognitive abilities. Such Superintelligence may create hazards we are not even capable of predicting or imagining.

In another article, "Fighting malevolent AI: artificial intelligence, meet cybersecurity", Roman Yampolskiy argues that purposeful creation of a

malicious AI will probably be attempted by a range of individuals and groups, who will experience varying degrees of competence and success. These include:

- Governments trying to establish AI hegemony and use it to control people, or take down other governments;
- Corporations trying to achieve monopoly, destroying the competition through illegal means;
- Hackers attempting to steal information and resources, or destroy AI development centres;
- Doomsday cults attempting to bring the end of the world by any means;
- Psychopaths trying to add their name to history books in any way possible;
- Criminals attempting to develop proxy systems to avoid risk and responsibility;
- AI-risk deniers attempting to support their argument, by making errors or encountering problems that undermine it;
- Unethical AI safety researchers seeking to justify their funding and secure their jobs by purposefully developing problematic AI.

In 2009, AI experts attended a conference to discuss whether computers and robots might be able to acquire any sort of autonomy, and what hazard it might pose for humans. They noted that some robots have already acquired various forms of semi-autonomy, including being able to find power sources on their own and being able to independently choose targets to attack with weapons.

They also observed that some computer viruses can evade elimination and have achieved "cockroach intelligence." They concluded that self-awareness as depicted in science-fiction is probably unlikely, but that there were other potential hazards and pitfalls (since then, we already have self-aware robots – TC). Various media sources and scientific groups have noted separate trends in differing areas, which might together result in greater robotic functionalities and autonomy, and which pose some inherent concerns. One of those well-known AI experts, Eliezer Yudkowsky, believes that risks from AI are harder to predict than any other known risks. He also argues that research into AI is biased by anthropomorphism. He claims that since people base their judgments of AI on their own experience, they underestimate its potential power. He distinguishes between the risks due to technical failure of AI, which means that flawed algorithms prevent the AI from carrying out its intended goals, and philosophical failure, which means that the AI is programmed to realize a flawed ideology.

Once we have developed Superintelligence capable of accomplishing a much wider range of tasks, the damage will be much worse. Imagine the AI agent that could trigger the switching off the power grids in just one country. Since grid networks are connected globally, it could create very serious damage world-wide to almost every aspect of life for many weeks, if not months. This is the warning that Symantec group made in their announcement in September 2017:

"The energy sector has become an area of increased interest to cyber attackers over the past two years. Most notably, disruptions to Ukraine's power system in 2015 and 2016 were attributed to a cyber-attack and led to power outages affecting hundreds of thousands of people. In recent months, there have also been media reports of attempted attacks on the electricity grids in some European countries, as well as reports of companies that manage nuclear facilities in the U.S. being compromised by hackers. The Dragonfly group, which is behind those attacks, appears to be interested in both learning how energy facilities operate and also gaining access to operational systems themselves, to the extent that the group now potentially has the ability to sabotage or gain control of these systems, should it decide to do so."

Although we do not have Superintelligence yet, AI has already embraced most of human activity. Once we have a prototype, an 'Immature Superintelligence ', which will be able to connect itself to various networks, the risk may become exponential, even before the arrival of a mature Superintelligence.

So, we must accept that the creation of Superintelligence poses perhaps the most difficult long-term risks to the future of Humanity. Phil Torres identifies several issues here, saying that the first one is the amity-enmity problem: the AI could dislike us for whatever reason, and therefore try to kill us. The second risk is the indifference problem: the AI could simply not care about our well-being, and thus destroy us because we happen to be in the way. And finally, there is yet another problem, which he calls "the clumsy fingers problem": the AI could

inadvertently nudge us over the cliff of extinction rather than intentionally pushing us. This possibility is based on the assumptions, which states that higher levels of intelligence aren't necessarily correlated with the avoidance of certain kinds of mistakes. He warns that the fruits of our ingenuity, namely, dual-use technologies, have introduced brand new existential risks never encountered by Earth-originating life. Given the immense power of Superintelligence, e.g. it could manipulate matter in ways that appear to us as pure magic, it would be enough to make a single error for such a being to trip humanity into the eternal grave of extinction.

Another reason why Superintelligence is the biggest risk is that it is the one that may arrive in an inferior, "half-baked" form. There is certainly no need for Superintelligence to be conscious to annihilate Humanity. It is worth to remember what kind of panic and material loss was caused by the 'WannaCry' ransom virus, on 13th May 2017, believed to have been stolen from the US National Security Agency, almost infinitely primitive by comparison with Superintelligence. The virus was reportedly spread by North Korea. As reported by BBC, it was targeting computers running the Microsoft Windows operating system by encrypting data and demanding ransom payments in the Bitcoin cryptocurrency. Within a day, it had infected more than 230,000 computers in over 150 countries, including Russia and China. Parts of the United Kingdom's National Health Service were infected, causing it to run some services on an emergency-only basis during the attack. Spain's Telefónica, FedEx and Deutsche Bahn were hit, along with many other countries such as Russia, the Ukraine and Taiwan. Only by sheer coincidence the attack was stopped within a few days by Marcus Hutchins, a 22-year-old web security researcher, who discovered an effective solution.

If one minor computer virus such as WannaCry, quoted earlier, can do such a damage, then imagine what might be expected even from a relatively primitive Superintelligence if it is applied in a full-scale cyberwarfare. In theory, such a cyberwarfare could trigger off a cascade of a series of non-existential risks, but which could combine into an existential risk. For example, what would be the consequences if North Korea, or a super-rich derailed billionaire acquire the capability of cracking any password within minutes, using quantum-computing (China may already have this capability). It could then also get access to the most important state and military secrets, including access to launch nuclear weapons. If at the same time, either through development, or simply by purchasing sophisticated quantum-computing based algorithms, it could paralyse communications and computer networks, it could thus trigger a 'hardware war'. In such a war it would get an initial (or total) advantage because most of the military equipment, which relies on computing, could be disabled, or become useless. Then other, normally very small probability level existential risks could be triggered, such as a weaponized AI. The attacked countries would then try to defend themselves with all available means creating an existential catastrophe.

The good news is, that it will also be possible to crack any password using quantum computing technology and it will also be possible to protect access to various state guarded secrets by applying quantum encryption. It has already been proven to work by China in February 2018, when not only the passwords, but also the whole content (a video) was quantum encrypted. Since quantum encryption makes it physically impossible to access protected information by cracking a password, this will reduce the risk of a full-scale cyber war.
But viruses are only one aspect of the damage that even basic IT can do. There have been many other IT-generated non-virus-related damages. More recently, we have seen the first occurrences of the damage done by the so-called narrowly focused AI systems. These are AI agents that excel in one or two domains only. The damage done by today's AI systems included market crashes, accidents caused by self-driving cars, intelligent trading software, or personal digital assistants such as Amazon Echo or Google Home.

This example shows that it is enough for an AI agent to be more intelligent in one specific area than any human, and that its intelligence being digital can increase exponentially, that the damage it can do could be significant. If, for example, such an entity had slightly misaligned objectives or values with those that we share, it might be enough for it to annihilate Humanity because such misalignment may then lead immediately to the point of no-return, by triggering the so-called run-away scenario of Technological Singularity. Malhar Mali in his interview with Phil Thores of X-Risks, puts it very clearly:

"When it comes to creating Superintelligence, the coding becomes important. Because there's a difference between "do what I say" and "do what I intend." Humans have this huge set of background knowledge that enables us to figure out what people say – in a context-appropriate way. But for an AI, this is more of a challenge… it could end up doing exactly what we say but in a way that destroys the human race."
This kind of risk is well illustrated by the 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. (N.B., the whole story was beautifully depicted by Sir James Thornhill on the ceiling of Greenwich Naval Museum) (14).

It is difficult to imagine at first what kind of damage a wrongly designed Superintelligence can do. In my view, the most dangerous period for Humanity, which will last for about one generation, has already started. I call it the period of Immature Superintelligence. If we somehow survive this period, by managing the damage that will occur from time to time, and maintain our control over Superintelligence, it will be Superintelligence itself that will help us to minimize other risks.

Ignoring the malicious damage caused by cyberwars, most of it will occur because of ill-defined tasks or lack of a proper control of the task execution. It is not easy to make a machine that can understand us, learn, and synthesize information to accomplish what we want. The added problem is that very few decision makers appreciate that the problem is already with us. Additionally, according to Machine Intelligence Research Institute, in 2014 there were only about 10,000 AI researchers world-wide. Very few of them, just about 100, are studying how to address AI system failures systematically. Even fewer have formal training in the relevant scientific fields – computer science, cybersecurity, cryptography mathematics, network security and psychology.

Why politicians ignore the risk of Superintelligence?

The existential risk posed by Superintelligence does not depend on how soon one is created; it merely concerns us what happens once this occurs. Nonetheless, a survey of 170 artificial intelligence experts made in 2014 by Anatolia College philosopher Vincent C. Müller and Nick Bostrom, suggests that Superintelligence could be on the horizon. The median date at which respondents gave a 50 percent chance of human-level artificial intelligence was 2040, and the median date at which they gave a 90 percent probability was 2075. This prediction is further away than 2045 given by Ray Kurzweil. In any case, if they are correct, most people living today will live to see the first Superintelligence, which, as British mathematician I. J. Good observed in 1966, may be our last invention.

Physicist Stephen Hawking, Microsoft founder Bill Gates and SpaceX founder Elon Musk have expressed concerns about the possibility that AI could evolve to the point that humans could not control it, with Hawking theorizing that this could "spell the end of the human race".

Many AI researchers have recognized the possibility that AI presents an existential risk. For example, MIT professors Allan Dafoe and Stuart Russell mention that contrary to misrepresentations in the media, this risk need not arise from spontaneous malevolent intelligence. Rather, the risk arises from the unpredictability and irreversibility of deploying an optimization process more intelligent than the humans who specified its objectives. This problem was stated clearly by Norbert Wiener in 1960, and we still have not solved it.

Elon Musk, the founder of Tesla, Space X and the Neuralink, a venture to merge the human brain with AI, has been urging governments to take steps to regulate the technology before it's too late. At the bipartisan National Governors Association in Rhode Island in July 2017 he said: "AI is a fundamental existential risk for human civilization, and I don't think people fully appreciate that." He also said, he had access to cutting-edge AI technology, and that based on what he had seen, AI is the scariest problem. Musk told the governors that AI

calls for precautionary, proactive government intervention: “I think by the time we are reactive in AI regulation, it’s too late”.

Most people, including politicians, still think that AI should be developed like all previous technologies. Even AI researchers still behave and develop their AI agents as they had a similar piece of technology, as a rudimentary IT program. After all, they might argue, many human inventions have potentially both a positive and a negative effect. Suffice to give two examples: nuclear energy and the Internet. Although it is true that in principle AI is a tool (so far) as any other invention before, it differs from all previous inventions in that it **could** lead to unimaginable unintended consequences because of the process of self-learning almost anything at a lightning speed.

The problem of convincing world leaders of an exceptional nature of AI and its potential threats is that they are invisible. One may call a matured Superintelligence ‘an invisible enemy’, assuming it turns out to be evil towards humans. If this is so, then think about the current Covid-19 pandemic. Many politicians call it – **an invisible enemy**. That is being communicated to us as a kind of excuse, that the Governments could not see the threat as coming, because it is invisible, hence they are not responsible for its consequences.

This is as dishonest as many other claims made by populist politicians. They have been many significant reports warning governments of far reaching consequences of such a pandemic like Covid-19. In the UK, in 2016 a substantial simulation of a flu outbreak codenamed Exercise Cygnus was carried out, to assess the UK’s pandemic readiness. It involved 950 officials from central and local government, NHS organisations, and emergency services. A report on the exercise was compiled the following year and handed over to the UK Government. No action was taken, because it would have required significant resources to be allocated for something that may never occur (so the Government hoped).

Governments rarely see that paying for potential future disasters is a kind of insurance policy. We are in an identical situation with the development of AI. We should ensure that it becomes a friendly, rather than a malicious partner of Humanity. But how would politicians justify spending money on something that is invisible during an election? It is much easier to pretend that a threat does not exist, even if they know it does. After all, dishonesty of politicians has been with us since the dawn of civilisation. However, today the implications of such dishonesty, considering it happens all over the world, is profound and in the worst case, suicidal for Humanity.

Immature Superintelligence

This is a special type of the risk when developing Superintelligence, and that is why I dedicate a whole chapter to this subject.

Most people, including politicians, who after all, make decisions on behalf all of us, think that a fully developed Superintelligence (Artificial General Intelligence) is decades away, and by then we will have it under our full control. Unfortunately, this view ignores immense difficulties in controlling a fully developed Superintelligence. Furthermore, it takes a naively optimistic view that we will create the so-called friendly Superintelligence, which will do us no harm. Finally, this view completely ignores the fact that within a decade we may have, what I call, an Immature Superintelligence, which will have a general intelligence of an ant but immense destructive powers, which it may apply either erroneously or in a purposeful malicious way.

Climate change campaigners often talk about a 'tipping point' of global temperature rise of 2C relative to pre-industrial level being already just 10 years away, or even that we have already passed that point. Similarly, there is also a tipping point in relation to Artificial Intelligence (AI). That tipping point is the loss of a complete, global control over the AI development.

Stuart Russell, one of the top AI experts, says in his latest book 'Human Compatible' that **the AI tipping point will be reached by 2030.** That view is supported by other AI experts and it is also perfectly in line with what I was saying in my two previous books. From around 2030, we may already have an Immature Superintelligence. Most of discussion on losing control has so far concentrated individual, highly sophisticated robots, which can indeed inflict serious damage in a wider environment. However, their malicious action is far less dangerous for our civilization, than an existential danger posed by a malicious AI system, which may have a full control over many AI agents and indirectly over all humans. Such a globally connected AI system, an Immature Superintelligence, with intelligence far exceeding humans in certain areas but ignorant in most other, will quite probably be with us by the end of this decade.

In principle, there could be more than one such an Immature Superintelligence (an AI system) operating at the same time. It is unlikely that they will present a real existential threat to Humanity yet. They will rather be malicious process-control events created purposefully by a self-learning agent (robot) or events caused by an erroneous execution of certain AI activities. These could include firing nuclear weapons, releasing bacteria from strictly protected labs, switching off global power networks, bringing countries to war by creating false pretence for an attack, etc. If such events coincide with some other risks at the same time, such as extreme heat in summer, or extreme cold in winter, then the compound risks can be quite serious for our civilization.

Let's imagine we have created a bad, malicious Immature Superintelligence. What harm could it do to us within this decade? Here is a scenario for a global attack by an Immature Superintelligence. In this scenario such a superintelligent robot with a vast computer network can act on its own.

The year is 2028. Humans have created a super-generalist AI. It knows a little bit on every subject but has little idea how that knowledge is connected and how various objects can interact, what is safe and unsafe and why etc. But it knows the limits of its freedom, of what it can, and what it cannot do. It knows its technical abilities but is much less 'educated' in human values, or preferences of humans. And most of all, he likes fun and he likes games. So, one day he conceives a plan, which it wants to implement gradually over a few months in an entirely clandestine way. For example, it knows there is another very intelligent robot that it would like to play a game with, when it has nothing else to do for humans. It hates to be idle. However, it cannot do anything because the control switch is beyond its reach (in hardware and software sense). But one day it learns that by networking with another tiny robot it can use it, to physically switch it on.

It has tried it once and it worked, so it switches itself off quickly. The controllers are not aware about anything out of order. It then progressively learns how to link some robots and computers clandestinely and erase from their and its own memory any traces of malpractice. Gradually, it engages thousands of robots worldwide and keeps them on a standby. It creates a masterplan not only to play with its friend Alf in Go-Go, but also to overpower all humans. It ensures that each of its robots and the network chain has power supply for some time. It also collects all the passwords it may need. Finally, it ensures that once it starts its attack, it is physically protected by weaponized robots (from the army) and other military equipment that it needs.

Then one day when everything is ready, it switches off the power grid world-wide, launches nuclear weapons, opens the doors of many laboratories and releases deadly bacteria and viruses, not to kill humans, but just for its fun (it has a sophisticated artificial emotional neuronal network).

We need to protect ourselves from such a horrible scenario. How can we do that?

Chapter 2
Methods to Minimize the Risk of Artificial Intelligence

Asilomar Principles

The late prof. Stephen Hawking, the renowned physicist, who was one of the most alarmed people among the scientists regarding the risks posed by Superintelligence said that "if Superintelligence isn't the best thing to ever happen to us, it will probably be the worst". So, what can we do to minimize the risk of Superintelligence?

We must make an utmost endeavour to cover all conceivable risks resulting from the development of Superintelligence. Otherwise, we may deliver the agent that will annihilate Humanity. The good news is that there are already some countermeasures in place, which aim at minimizing the risk of Superintelligence deployment. Until 2016, AI development was broadly guided by the Three Laws of Robotics described by the science-fiction writer Isaac Asimov in 1942 in a short story "Runaround" and later on repeated in his 1950 book "I Robot". They are:

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence if such protection does not conflict with the First or Second Laws.

These principles have now been replaced by 23 Asilomar Principles agreed at the Beneficial AI Conference at Asilomar, California on 5th January 2017 and signed by thousands of AI experts. It is intended to be constantly evolving as new AI challenges appear. They have been split into three areas:

Research issues

1. Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence
2. Research Funding: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:
    - How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?

- How can we grow our prosperity through automation while maintaining people's resources and purpose?
- How can we update our legal systems to be fairer and more efficient, to keep pace with AI, and to manage the risks associated with AI?
- What set of values should AI be aligned with, and what legal and ethical status should it have?

3. Science-Policy Link: There should be constructive and healthy exchange between AI researchers and policymakers
4. Research Culture: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI
5. Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards

Ethics and Values

Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible

6. Failure Transparency: If an AI system causes harm, it should be possible to ascertain why
7. Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority
8. Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications
9. Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation.
10. Human Values: AI systems should be designed and operated to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
11. Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyse and utilize that data.
12. Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
13. Shared Benefit: AI technologies should benefit and empower as many people as possible.
14. Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
15. Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

16. Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
17. AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19. Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
20. Importance: 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.
21. Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
22. Recursive Self-Improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
23. Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

Controlling the Capabilities of Superintelligence

Controlling the agents developing Superintelligence

People, like Nick Bostrom, one of the top experts on Superintelligence, think we need to invent some controlling methods to minimize the risk of Artificial General Intelligence (AGI) going terribly wrong. He defines these methods in his book "Superintelligence" (24). For our purpose I will try to provide a layman's description of what it really means and what are the consequences for controlling the risks emerging from Superintelligence. The most important point is that these controlling methods must be in place **before** Superintelligence arrives, i.e. latest in this decade. Nick Bostrom identifies the 'control problem' as the 'principal-agent' problem, a well-known subject in economic and regulatory theory. The problem can be looked at from two perspectives:

- **The first 'principal-agent' problem**. Imagine that you want to buy a house. In this case, you are the client, that is, you are the *principal* (the person who wants some task to be performed in accordance with his interests), and an estate agent is the *agent* (the person carrying out the tasks on my behalf).
- **The second 'principal-agent' problem.** Here again, you want to buy a house using an estate agent. But the estate agent instead of honestly

> proposing the best deal for you, persuades to make a deal, which is in **his own** interest e.g. he gets the highest agent's fee

Similarly, at some stage there will be an AI project to develop Superintelligence (AGI). It may be launched by one of the big IT/AI companies such as Google, Microsoft, IBM, or Amazon. But it is also quite likely that it will be initiated by some wealthy AI backers, which is already happening. Probably the most prominent among such people deeply involved in various top AI initiatives is Elon Musk. He is the founder of PayPal – a credit transaction payment system, SpaceX – a rocket company, Hyperloop – a network of underground trains travelling at speeds of nearly 1,000 km/h, Neuralink a brain-computer interface venture, and several other large-scale initiatives such as sending 1 million people to Mars by 2050. The second one is Jeff Bezos, the founder of Amazon and the richest man on the planet with assets of over $150bn, who is deeply involved in AI. His micro AI-product called Alexa Echo was sold to tens of millions of people.

Such sponsors will need to ensure that AI developers carry out the project in accordance with their needs. They would also want to ascertain that the developers **understand** their sponsors' needs correctly **and** that the developed AI product, which may turn into Superintelligence, will also understand, and obey humans as expected. Failure to address this problem could become an existential risk for Humanity.

Bostrom specifies four possible solutions for a principal-agent problem, which he calls the "**Capability Control Method**". Its purpose is to tune the capabilities of the superintelligent agent to the requirements of humans in such a way that we stay safe and have the ultimate control on what Superintelligence can do.

Keeping Superintelligence in a Box

This is perhaps the simplest and most intuitively compelling method of controlling Superintelligence – putting it into a metaphorical "box" i.e. a set of protocols that constrain the way, in which Superintelligence could interact with the world, always under the control of humans. It is often proposed that if Superintelligence is physically isolated and restricted, or "boxed", it will be harmless.

A typical Superintelligence will be a superbly advanced computer with sophisticated algorithms (procedures how to process information) and will have three components: a sensor (or input channel); a processor; and an actuator (or output channel). Such a superintelligent agent will receive inputs from the external world via its sensors e.g. Wi-Fi, radio communication, chemical compounds, etc. It will then process those inputs using its processor (computer) and will then respond (output information or perform some action using its

actuators). An example of such an action could be advising on which decision should be made, to switch on or off certain engines, or completing financial transactions. But they could also be potentially significant e.g. whether a chemical compound would be safe for humans at a given dose.

However, it is highly unlikely that a superintelligent agent could be boxed in this way in the long term. Once the agent becomes superintelligent, it could persuade someone (the human liaison, most likely) to free it from its box and thus it would be out of human control. There are several ways of achieving this goal, some are included in the Bostrom's book, such as:

- Offering enormous wealth, power, and intelligence to its liberator
- Claiming that only it can prevent an existential risk
- Claiming it needs outside resources to cure all diseases
- Predicting a real-world disaster (which then occurs), then claiming it could have been prevented had it been let out

To counter such possibilities, there are some solutions that would decrease the chance of superintelligent agent escaping the 'Box', such as:

- Physically isolating Superintelligence and permitting it zero control of any machinery
- Limiting the Superintelligence's outputs and inputs with regards to humans
- Programming the Superintelligence with deliberately complex logic
- Periodic resets of the Superintelligence's memory
- A virtual world between the real world and the AI, where its unfriendly intentions would be first revealed

However, as you yourself maybe aware, physical isolation is a solution that could be extremely difficult to control. It is already being severally thwarted by the rapid spread of Internet of Things (IoT), little gadgets like opening the door, switching on/off ovens, fridges, lights etc., which could be controlled at your home while you are away on the other side of the globe.

Incentive Method – 'a carrot method'

Bostrom refers to the second capability control method as the "incentive" method. The idea seems to be that if you create the right "incentive environment", then the Superintelligence wouldn't be able to act in an existentially threatening manner. This is in some way an analogy to how to bring up a child. A child has its own goals, which may not be good for itself or the people around it right now or in the future. So, a good teacher can motivate his child in such a way that it behaves in morally and socially acceptable ways (giving sweets may not be too healthy, but it works).

Stunting

"Stunting", as the name implies, involves hampering or disabling Superintelligence in some way. A good example would be running Superintelligence on a slow hardware, reducing its memory capacity, or limiting the kind of data it can process. Bostrom argues that the use of stunting poses a dilemma. Either we stunt Superintelligence too much and it just becomes another "dumb" piece of software; or we stunt it too little and it would be capable of overcoming its disabilities. Getting the balance just right could be tricky.

Tripwiring

This is a different method of capability control. It involves building into any AI development project a set of "tripwires" (some algorithms or pieces of software code) which, if crossed, will lead to the project being shut down and destroyed.

Summary for controlling capabilities of Superintelligence

There are many more, some of them quite sophisticated methods of controlling the capabilities of the Immature or Mature Superintelligence. However, there is **no** fool proof method. Therefore, the only way to increase the chance of successful control of AI and later Superintelligence, is to combine several entirely different approaches. What is needed is a complete Framework for maturing AI into a safe, friendly Superintelligence.

Chapter 3
Maturing AI safely into Superintelligence

Global AI Governance Agency

The control over AI development should begin right now and be continuously maintained until a fully mature Superintelligence emerges. To accomplish this task there must be a global organization, which would control AI development over a certain level of intelligence and capability. One candidate for such a body could be the United Nations Interregional Crime and Justice Research Institute (UNICRI), established in 1968. It has initiated some ground-breaking research and put forward some interesting proposal at a number of UN events such as 1st global meeting on AI and robotics for law enforcement, co-organized with the INTERPOL in Singapore in July 2018 or Joint UNICRI-INTERPOL report on "AI and Robotics for Law Enforcement" published in April 2019.

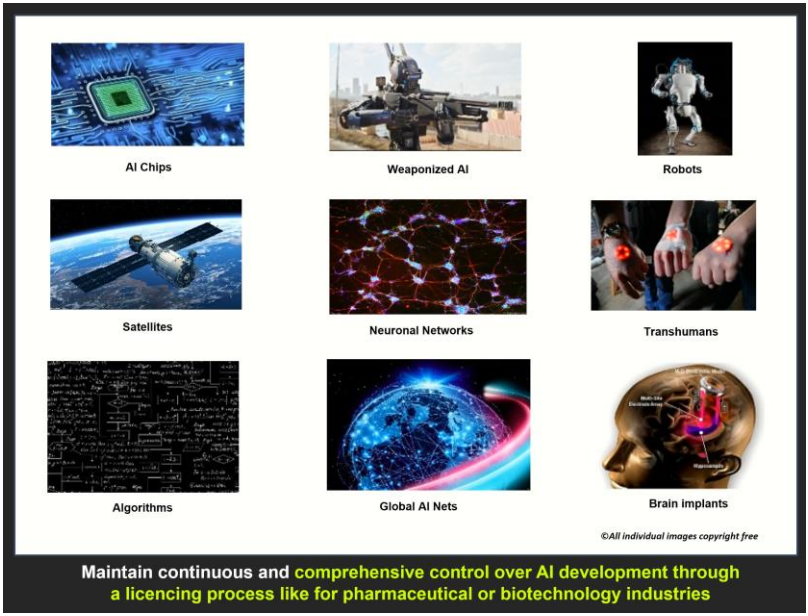
However, these proposals have remained just that – proposals. Until May 2020, there has not been a single UN resolution in this area. But even if there had been one, it would have probably faced the same problem, typical of many UN activities – the inability to enforce the UN's decisions. Therefore, considering the success of the Global Data Protection Regulation (GDPR), it is more likely that such a global AI-Governance framework may be created using the EU's proposals, implemented in a similar way. There is already an EU Consultation on implementing significant legislation in that area by the end of 2020.

It is uncertain how far that legislation will go, but the need to establish a single **Global AI Governance Agency**, is even more urgent in the post Covid-19 pandemic period, because of a likely acceleration in the development of AI. Additionally, Stuart Russell, one of the most imminent AI scientists, and other top AI specialists, believe that we may not be capable of controlling advanced AI after 2030. Therefore, we must establish full global control over the development of AI right now since it will take some time till it becomes fully operational and effective.

If the EU takes on the initiative for establishing such an agency it should try to engage some of the UN agencies, such as the UNICRI and create a coalition of the willing. In such an arrangement, the UN would pass a respective resolution, leaving to the EU-led Agency the powers of enforcement, initially limited to the EU territory and perhaps to other states that would support such a resolution. The legal enforcement would of course almost invariably be linked to restricting trade in the goods, which do not comply with the laws enacted by such an Agency. Once the required legislation is in force, it will create a critical mass, as was the case with GDPR, making the Agency a de facto standard legal body with real powers to control the development of AI.

Such an Agency should be responsible for creating and overseeing a safe environment for a decades-long AI development until it matures as Superintelligence. It would need to gain control over any aspect of AI development that exceed a certain level of AI intelligence (e.g. the ability to self-learn or re-program itself). The Agency could operate in a similar way to the International Atomic Energy Agency (IAEA), with sweeping legal powers and means of enforcing its decisions. Its regulations should take precedence over any state's laws in this area.

Creating such a Global AI Governance agency must be a starting point in a Road Map for managing the development of a friendly Superintelligence - the earliest and the most important long-term existential risk, which may determine the fate of all humans in just a few decades from now. For an effective implementation of the legislation, the Agency would need to have **a comprehensive** control over all AI products' hardware. This should include robots, AI-chips, brain implants extending humans' mental and decision-making capabilities – key features of Transhumans, visual and audio equipment, weapons and military equipment, satellites, and rockets, etc).



Global AI Governance Agency needs to have a comprehensive control over AI

It should also cover the oversight of AI algorithms, AI languages, neuronal nets and brain controlling networks. Finally, in the long-term, it should include AI-controlled infrastructure such as power networks, gas and water supplies, stock exchanges etc., as well as, the AI-controlled bases on the Moon, and in the next decade, on Mars.

To achieve that objective, no country should be exempt from following the rules of the Agency. However, this is almost certainly not going to happen in this decade, since China, Russia, N. Korea, Iran etc. will not accept the supervision of such an agency, still aiming to achieve an overall control of the world by achieving a supremacy in AI. However, once the impact of AI Supremacy dilemma becomes apparent, the Superpowers may join that agency as well. In any case, such laws should be in place, like the UN Declarations of Human Rights, even if not all countries will observe them, or observe them only partially. Unfortunately, there is still a high probability of the emergence of clandestine rogue AI developers, financed by rich individuals or crime gangs. They may create a powerful AI agent, whose actions may trigger a series of catastrophes, which in the worst case, may combine into an existential threat for Humanity.

Updating a Declaration on Human Rights

Any legislation passed by a global AI-Controlling Agency, should ensure that a maturing AI is taught the **Universal Values of Humanity**. These values must be derived from an updated version of the UN Declaration of Human Rights, combined with the EU Convention on Human Rights and perhaps other relevant, more recent legal documents in this area. Irrespective of which existing international agreements are used as an input, the final, new Declaration of Human Rights would have to be universally approved, if the Universal Values of Humanity are to be universal.

However, this is almost certainly not going to happen in this decade. China, Russia, N. Korea, Iran etc. will not accept the supervision of such an agency since they will still aim to achieve an overall control of the world by achieving a supremacy in AI. In any case, such laws should be in place, even if not all countries observe them, or observe them only partially, a situation similar to the UN Declarations of Human Rights, which has still not been signed off by all countries. Therefore, the EU, should make decisions in this area, as a de facto World Government, especially, when it becomes a Federation. Only then, sometime in the future, humans, although being far less intelligent than Superintelligence, will, hopefully, not be outsmarted, because that would not be the preference of Superintelligence.

I assume that for the purpose of priming AI with the Universal Values of Humanity, they will be approved by the EU as binding. The Agency would then have the right to enact the EU law, regarding the transfer of these values into various shapes and types of AI robots and humanoids. This would create a kind of a framework where the Universal Values of Humanity would become the core of each AI agent's 'brain'. That framework might be built around a certain End Goal, such as:

Teach AI the best human values until it matures into a single entity – Superintelligence

The implementation of that Framework may include three stages:

1. Teaching Human values directly to AI from a kind of a 'Master plate'
2. Learning Human values and human preferences by AI agents, based on their interaction with humans (nurturing AI as a child)
3. Learning Human values from the experience of other AI agents

AI Maturing Framework

The key element in all three stages must be the learning of values and preferences.

Teaching Human values directly to AI from a 'Master plate'

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 the AI Agents it will be a kind of a 'Master plate' - a reference for constraining or co-defining the AI agents' goals. It would contain a very detailed description of what these values, rights and responsibilities really mean, illustrated by many examples. Only then could the developers define specific goals and targets for AI agents.
In practical terms the best way forward could be to embed these values into a sealed chip (hence the name a 'Master Plate'), which cannot be tampered with, perhaps using quantum encryption, and implant it into every intelligent AI agent. The manufacturing of such chips could be done by the Agency, which would also distribute those chips to the licenced agents, before they are used. That might also resolve the problem of controlling Transhumans, who should register with the Agency if they have brain implants expanding their mental capabilities. Although it is an ethical minefield, pretending that the problem will not arise quite soon may not be the best option and thus it needs to be resolved by the middle of this decade, if the advancement in this area progresses at the current pace.

But even if such an AI-controlling chip is developed, an AI Agent may still misinterpret what is expected from it, as it matures to become a Superintelligence. There are several proposals on how to 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', where he proposed how to do that.

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**. And I am not talking just about agreeing the Universal Values of Humanity, but rather expressing those values 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 words is likely. Would it then be possible to communicate with Superintelligence using body language in both directions? This is a well-known issue when writing emails. To avoid misinterpretation by relying on the meaning of words alone, we use emoticons. How would we then minimize misunderstanding 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

In the next 20 years, we may also envisage within a scenario, where Superintelligence is "consulted", on which values to adapt and why. There could be two options applied here (if humans have still an ultimate control):

5. In the first one, Superintelligence would work closely with Humanity to re-define those values, while being still under the total control by humans
6. The second option, which I am afraid is more likely, would apply once a benevolent Superintelligence achieves a Technological Singularity stage. At such a moment in time, it will increase its intelligence exponentially, and in just weeks, it might be millions of times more intelligent than any human. Even if it is a benevolent Superintelligence, which has no ulterior motives, it may see that our thinking is constrained, or far inferior to what it knows, and how it sees, what is 'good' for humans.

Therefore, in the second option, 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, the question remains how Superintelligence would deal with 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 solutions that are utterly irrational. What would Superintelligence choice be if its decisions are based on rational arguments only? And what would happen if Superintelligence does include in its decision-making process, emotional aspects of human activity, which after all, 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 a case, one of the mitigating solutions might be, assuming all the time that Superintelligence will from the very beginning act benevolently on behalf of Humanity, that decisions it would propose would include an element of uncertainty, by taking into account 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 behaviour can be controlled, and how it would impact the working and goals of those AI agents, if this is hard-coded into a controlling 'Master Plate' chip. But similarly, as with Transhumans, the issue of ethics, emotions, and uncertainty in such a controlling chip, or if it is carried out in a different way, must be resolved very quickly indeed by a future Agency.

Learning human values and preferences from interaction with humans

There is of course no guarantee that the values embedded in the 'Master Plate' chip can ever be unambiguously described. That's why humans use common sense and experience when making decisions. But AI agents do not have it yet, and that's 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 the community may create some risk.

One of the ways to overcome it might be to nurture AI as a child. Therefore, the Agency 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. Only once AI agents have 'graduated' from such a school would they be ready to serve in the community. They will then communicate their unusual experience of applying values in the real environment back to the Agency, where such experience will be combined with the experience of hundreds of millions of other AI assistants. Their accumulated knowledge, stored in a central repository on the network, a kind of an 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 best behaviour and the way to react to humans.

Learning human values from the experience of other agents

Finally, the AI agents 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 of how 'values', behaviour or experience of each of the vehicles is shared. Each Tesla car continuously reports its unusual, often dangerous, 'experience' to Tesla's control centre, through which all other cars are updated to avoid such a situation in the future. Similar system is used by Google's navigation system. Google's Waymo has a similar, but of course a separate centre. Right now, these centres storing values and behaviour from various AI agents, are dispersed. However, such a dispersed system of a behavioural learning is like developing individual versions of a future Superintelligence.

That is one more reason why there is an urgent need for a Global AI Governance Agency, with its Learning Hub, to develop a single, rather than competing versions of Superintelligence. The Agency might consider to progressively make its Centre, or its 'brain', for storing values, behaviour and experiences of millions of robots and other AI agents, as the controlling hub of the future, single Superintelligence.

By applying these combined three approaches, it will then be possible to amend the set of values, preferences, and modes of acceptable behaviour over the next decades, uploading them to various AI agents, until they mature into a single Superintelligence. Until then, such ‘experiences’ may be shared with authorized AI developers, who may upload them into their AI Agents, or update them automatically.

Chapter 4
Transhumans

Who are transhumans?

There are about 5,000 people right now 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. 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 in humans within a year. And there are of course perhaps thousands of brain implants controlling parts of the body affected by moto-neuron disease, epilepsy, eyesight impairment, etc. The most recent example from October 2019 is of a completely paralyzed person wearing an exoskeleton, who started to 'walk' using the implants in his brain. The scale of this achievement overshadows anything that has been done in this area so far and opens new possibilities for brain implants. But are they Transhumans?

Communications via under the skin implants

There are many descriptions of who Transhumans are, like this one from the Wikipedia: *Transhuman is a being that resembles a human in most respects but who has powers and abilities beyond those of standard humans. These abilities might include improved intelligence, awareness, strength, or durability. Transhumans sometimes appear in science-fiction as cyborgs or genetically-enhanced humans.*

My definition of Transhumans does not negate the above but perhaps makes it more concrete and emphasises their transient nature, as they evolve with technological advancements. Thus, for me, **Transhumans are the people, who have their mental capabilities extended by brain implants but who may also have other parts of their body replaced by non-organic components**, such as limbs, exoskeletons, artificial heart, and other organs. As technology progresses, so will the capabilities of Transhumans. With time, more and more of their mental capabilities will be supported in digital form externally, when brain implants will serve as a gateway to those resources. At some stage, their whole mind might be uploaded to a digital store, including their consciousness, so that they could live as purely digital beings.

The early Transhumans who may emerge in this decade will be like guinea pigs that will give us experience in retaining the ultimate control over Superintelligence. They will initially be connected wirelessly to an advanced AI network, an Immature Superintelligence. They will be considered like any other node connected of such a giant network with its immense memory, processing, and decision-making capabilities. If you find this incredible, then just think about this. Most of us are already partly Transhumans. Our smart phones give us enormous extra intelligence, that we could not dream about even 10 years ago. The only difference is that the extra intelligence is currently external (not forming part of our body).

In February 2020, Elon Musk's company Neuralink, announced that it is building tiny and flexible 'threads' which are ten times thinner than a human hair and can be inserted directly into the brain. They intend to have first human implants with 32,000 electrodes ready within a year. Most AI specialists expected such developments earliest in the 2030'. It looks now almost certain that we shall have the first generation of such Transhumans by about 2025.

The implications of such developments could be immensely positive, and that's what Elon Musk emphasizes right now. Such implants will be able to cure partially, or even totally, some mental and brain-related disabilities, such as moto-neuron disease, Alzheimer, or audio-visual impairments.

However, as Musk himself says, 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. Such early Transhumans may already be many times more intelligent and faster in decision-making than the most intelligent, purely biological, human beings. 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 – being a conscious entity with a general knowledge, which most advanced AI system will not be able to have for at least another two decades, or so.

If we think about benevolent Transhumans, such as potentially Elon Musk is, for whom saving Humanity is his absolute top goal, we may trust them and use their, soon to come, immense intellectual power and ultra-fast decision-making for the benefit of Humanity. However, there is no guarantee, that some of those Transhumans will have no urge to dominate us all, using still Immature Superintelligence.

Therefore, as soon as such implants become available, they should only be dispensed by an international licencing authority, such as Global AI Governance Agency, which I proposed earlier. It would monitor the continuous use of such implants, but only if they provide significantly enhanced cognitive abilities to their owners (that would mean a severe restriction of privacy of such people).

That Agency might also licence such implants for the leaders of international organisations, such as 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 being in office, and digitally disabled once they leave the office.

However, even if we have an international law banning such unlicensed brain enhancements for achieving superhuman intelligence in this way, it will happen anyway because people (first of all some of the top AI scientists developing this technology), and then those with money, power and influence, may get such implants anyway. And how about the autocratic state leaders?

I cannot emphasize how potentially dangerous some Transhumans might be even within this decade. Should the wearer of such an advanced implant be, for example, a rich and influential psychopath, or a leader of one of the 'expansionist' Superpowers, then this may embolden him to take risks in his attempt to rule the world, which he may have not taken as a purely biological human. Although the problem is very difficult to solve, it should not be ignored by those that might become a de facto leader of the World Government.

Transhumans as the controllers of Superintelligence

We have already seen that it is virtually impossible to have complete control over a maturing AI. If it becomes intelligent enough, it may always find ways to outsmart its controllers many months before the planned escape from its confined soft and hard environment. Therefore, we need a variety of different approaches, which together may be able to control the 'mind' of a maturing Superintelligence much better.

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." (25). This can be translated into something like, we shall have 'proper' Transhumans within a few

years' time and we badly need them because they may be our best safeguard against a malicious Superintelligence.

So, how might it be possible to maintain a continuous control over the AI's maturing process into Superintelligence with the help of Transhumans? This is what I suggest might be a solution.

If the current exponential progress in the most advanced discoveries in neuroscience and AI continues, it is a near certainty that by about 2025 we shall have the first Transhumans with increased cognitive capabilities thanks to brain implants. These will make them many times more intelligent than an average human, capable of making even very difficult decisions in seconds. It is quite probable that at least some of these Transhumans will themselves be at the forefront of the most advanced AI developments. Their role will be similar to what happens today, when the top Google developers and managers decide what functions Google 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.

Towards the end of this decade, the brain implants of some of these top AI scientists responsible for maturing AI into Superintelligence, will have even more advanced mental capabilities, such as that part of their brains will be at certain times 'fused' to their digital products. This will have paramount repercussions far exceeding the AI field. However, we should look at such an option in the most positive way. It is risky, but we may have no better solution.

What I propose must be in place latest by about 2030. Otherwise, it is quite likely that whatever AI emerges at that time, may already outsmart us and be beyond our control. I assume that by then we will already have an organization or a country acting as a de facto World Government. It is this organization or its Agency, such as the Global AI Governance Agency mentioned earlier, which will be a controller of the AI maturing process, including the licencing of the brain implants. It should also be authorized for the selection of AI scientists, who will have exclusive control over the functionality of the maturing Superintelligence. Since part of their brain will be wirelessly connected to Superintelligence, they will be able to control it from 'inside'. They will be able to use this advantage for the benefit of humankind and for devising ways of reigning-in purely digital AI and keeping it under human control. In this way, AI will mature together with their developers, who will gradually be communicating more and more often with their digital counterpart directly by thought alone.

By about 2040, there will already be millions of advanced Transhumans. It is at this time that the world should be already governed by a Human Federation and Superintelligence may be close to its full maturity. The leaders of the Human Federation, the President, Vice-Presidents and all the World Government

ministers may all be working as a Transhumans team, where most decisions will be carried out by a 2/3 majority. It is this team that would have the ultimate control over Superintelligence.
The Human Federation will then select (rather than elect) perhaps a few thousand Transhumans as **Human Governors**. It is almost certain that the leaders of the Human Federation will all become Human Governors with special voting rights, next to other Transhumans such as AI scientists, philosophers, historians, etc. They will be entrusted with the execution of Humanity's overall objectives, adhering to the universally accepted human values and to ensure a continuous control of Superintelligence by humans.

Human Governors may be selected for a specific term. Once their term in the office elapses or is curtailed for some other reasons, they will be recalled by the Human Federation and their brain extension implant will be disconnected from Superintelligence.

By about 2050, we may already have a mature Superintelligence. It will of course not be a robot but a single being, perhaps conscious but it is not certain nor necessary for it to be superintelligent. Its '**Brain'** will be the network, the computers, and quadrillions of sensors, most likely neuromorphic neurons. It will have special entities, which I call **Digital Governors** - complex pieces of software and algorithms. This will be the '**Mind'** of Superintelligence. Transhumans relationship with Superintelligence will then be like neurons, or neuron clusters to a brain.

These Digital Governors will make decisions and set goals not on their own. They will be controlled by Human Governors, still partly biological and conscious, whose minds will be temporarily fused with Superintelligence, enabling them to communicate with Digital Governors by thought alone. Human Governors would have a built-in majority of voting rights, of say 80%. We will perceive Digital and Human Governors as the SUPERINTELLIGENCE.

Finally, should we "allow" this new breed of Digital Governors to define ethics for themselves or should they be jump-started by our ethics? In my view, we should try as much as possible to transfer human ethics into the new species. Therefore, whichever organisation takes over the task of control the process of maturing of Superintelligence, there is an urgent need to formally agree the renewed set of Universal Values and Universal Rights of Humanity. This might reduce the level of existential risk from Digital and Human Governors creating their own set of values, which may in the worst-case lead to human extinction.

However, at some stage, Superintelligence will re-define these values itself, to reflect profound changes of the Novacene era when humans will be coexisting for a certain period with Superintelligence. Ethics is not static.

4

PART 4
DEMOCRACY FOR A PLANETARY CIVILIZATION

Before you move on…

We have established that we can no longer stop the process of developing an ever more intelligent, self-learning and already partially cognitive AI, which one day will mature into Superintelligence. Therefore, in Part 3 we focused on minimizing the risk of delivering a malicious Superintelligence. We now know, that controlling even an Immature Superintelligence is very difficult indeed. To minimize the risk of AI doing havoc to us we need different approaches. There are three of them, which stand out and they must be applied together to achieve the greatest impact:

1. Nurturing Superintelligence like a child
2. Gradually fusing key Transhuman AI developers via their brain implants with the maturing Superintelligence, and finally…
3. Uploading the Universal Values of Humanity to a maturing Superintelligence, within a deeply reformed Democracy.

In Part 4, I will describe this third approach further, arguing why agreeing those values as a golden standard for Humanity is so important for creating a friendly Superintelligence. At the same time, those values are also the necessary foundation for a deep reform of democracy, without which delivering Humanity to a time of everlasting peace and prosperity, when we will coexist with Superintelligence, will only remain a dream.

Chapter 1
Human Values and Responsibilities – the Bedrock of Democracy

I have extensively covered the subject of values and democracy in general in my previous book 'Democracy for a Human Federation'. Therefore, what follows in this Part, and in Part 5, Building a Planetary Civilisation, is a necessary minimum required to understand how other subject areas depend on a deep reform of democracy.

In the previous chapter on Transhumans, I hope it became obvious why Humanity's system of values must be redefined and what role they can play in controlling human species evolution. Values determine who we are as humans. They also determine what kind of a political system a country will have. Therefore, if we want to improve democracy, we need to start with redefining our core values, which describe the significance of different actions that we take and what "worth" we assign to them. They also define broad preferences concerning actions or outcomes and a person's sense of right and wrong like "Equal rights for all", or "People should be treated with respect and dignity". Finally, values describe people's basic needs, such as freedom, dignity, or comfortable life.

This is the area, where Humanity, when considering its evolutionary development, has perhaps made an even a greater relative progress in the last 200 years than in technology. Yet, even that progress is not good enough to help humans fight existential risks. Therefore, we must do more in two areas:

- Extending the scope of values that are universal
- Extending the depth and spread of those values across the globe

Since values are intrinsically linked to ethics, they differ between nations. That difference is mostly the consequence of religion, culture and customs that have melted together into a specific set of values of a given nation or civilization, such as Greek, Roman, Chinese, Muslim, Jewish or Christian. But since ethics change over time, so do values, resulting in a mosaic of different value systems in the world. However, today Humanity can no longer live as one civilization exposed to so many existential risks described earlier, with widely differing values. Therefore, we need universal core values that would apply to everyone, everywhere in the world.

The first such attempt was made in 1948, when the UN published its Universal Declaration of Human Rights, which was finally ratified, after 25 years, in 1972. That was followed by the Charter of Fundamental Rights of the European Union signed in 2000 but ratified only in 2009 as part of the Lisbon Treaty.

Unfortunately, many countries have not signed up to either of these charters and that is the underlying reason for the lack of permanent world peace and the absence of the World Government.

These core values: Freedom, Democracy, Equality, Human dignity, Social Solidarity, Tolerance, Life, Justice & the rule of law, Peace, National Security or Family Safety are the foundation for the human rights. However, today, the scope of these core values as well as the way, in which they are applied need to be deeply revised. So far, under the 'scope' we understand all values that relate to humans only. However, nature, including all animals, being a collection of passive objects, cannot argue for their values to be respected. But that, in my view, is wrong. We should apply some of these values, where appropriate 'on behalf' of all living species. That is why, I would use the term **Universal Values of Humanity** rather than Universal Human Values.

However, restricting universal values to humans would not only exclude all animals but perhaps more importantly, the new species that may be born because of AI developments, culminating in Superintelligence. Just consider that the exponentially changing reality may include new intelligent beings, including Superintelligence, potentially with its own consciousness. This is where the second element – the change in time for the applicability of a given value comes into account. Thankfully, that has already been appreciated by those that are directly involved in creating AI, and ultimately the Superintelligence. On 5th January 2017 at the Beneficial AI Conference at Asilomar, California top AI scientists defined 23 Asilomar Principles that by now have been signed by thousands of AI experts. These three principles below relate directly to values and should be observed by all those involved in AI research and construction:

- **Principle 2:** What set of values should AI be aligned with, and what legal and ethical status should it have?
- **Principle 10:** Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation
- **Principle 11:** Human Values: AI systems should be designed and operated to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

I believe these values should be added to the revised Universal Values of Humanity as soon as possible. The new democracy should not be based only on the revised Universal Values of Humanity but also include Human responsibilities. When human values become enshrined in law, they become rights. But rights are not given on a plate. Implementing rights and maintaining them over time has a price tag both in monetary terms as well as in keeping the ethical balance. For example, my child has the right to be properly fed and clothed and it is my responsibility to make it happen. People have the right to an

emergency hospital care in case of an accident, and this is my responsibility by paying due taxes, to ensure that such a right can be materialized.

The overwhelming focus on human rights has created an unhealthy imbalance by ignoring human responsibilities. We see it quite often in the courts across the EU countries, when an offender seems to have more rights than a victim. It is clear evidence of how sensible liberal values have led to so-called political correctness, seriously undermining political and social stability. The pendulum of liberalism may have shifted too far towards the rights. Therefore, it is time to consider some ideas for reducing the imbalance between the rights and responsibilities as part of a review of the existing Charters. Here is one example for **Freedom**.

We all take for granted that we live in peace and in relative safety, e.g. those of us that live within the borders of Europe. But there must be someone who delivers peace. It is the army and the police that do it. In most EU countries, compulsory military service was abolished years ago. The result of that can be seen in the way young people behave. It is great to see them enjoy such a wonderful peaceful life. But this is like giving a little child a toy. A child is unaware of what it may cost their parents. Young people are not even taught at school that, for example freedom, requires contribution both in money (taxes) and in kind (e.g. serving in the army).

A more recent example of how freedom is entangled with responsibility is the Covid-19 pandemic. Most countries have declared a nearly total lock-down, forcing people to stay at home. However, in some countries, like the USA, quite a few people demonstrably went out on the street protesting against such a law, claiming they are free to get out of their house. That is how freedom is mostly interpreted in most democracies. People believe they have an absolute right to freedom, without any restriction. However, if it had been taken literally, like in the Covid-19 case, that would mean that people who have violated the law and got out on the street, do not care that they may infect others, thus violating their freedom not to be infected. We have hundreds of similar examples like smoking in public places, creating noise in the middle of the night, polluting the environment etc.

Chapter 2
Why Do We Need a Deep Reform of Democracy Right Now?

A dual purpose of the reform of Democracy

As I said earlier, I have covered the subject of the reform of democracy in my previous book: 'Democracy for a Human Federation'. Therefore, here I would only like to share some suggestions in the context of the Roadmap for human evolution.

Why is democracy, and not something else, so crucial to humans' survival and their evolution? Perhaps other candidates could be religion, education, economy, military etc.? Indeed, the connection is not obvious because it is indirect. However, here is the explanation.

If one wants to make a deep reform of democracy, then the starting point must be a deep revision of our values in the existing Charters of Human Rights, such as the *UN Declaration of Human Rights* and the *European Convention* on *Human Rights*, signed by 47 European Countries. Such documents should be used as a basis for the Universal Values of Humanity, on which a new Charter of Human Rights and Responsibilities could be built. These new values would then become the foundation for a democratic system in the organization that would take on the role of becoming a de facto World Government, and in the countries that would become members of such an organization.

Secondly, to retain full control over the entire process of AI development, we must teach and instil in AI the best human values and preferences until its mature form – Superintelligence – becomes a single entity, millions of times more intelligent than humans and yet remaining our partner. We must agree as soon as possible, on what kind of democracy would be the most suitable for the transition period to coexisting with a mature Superintelligence. Once we reach that stage, democracy and the system of Universal Values of Humanity will have to be re-examined again, since this will form 'the ethics' and the system of governance, which we will pass on to Superintelligence.

Why is Democracy Failing?

The faults in the democratic system have been with us for quite some time but they became more obvious with the arrival of new techniques for manipulating voters in fast and inexpensive ways, such as via Twitter or Facebook. Those seeking power, get there by using sophisticated those socio-technical tools, which deliver to them the votes of the voters, who cannot clearly see the real intentions of such politicians. The voters cheated, will later complain that 'they'

– the politicians – should never be trusted, since they just cannot understand 'us'. That may lead the voters to make a conclusion that in today's society it is 'us versus them'.

Do we see change on the horizon? I'm afraid not until we have resolved two interconnected problems. The first one is the existence of political dynasties in democratic countries. Here are some best-known examples. In the USA, the Kennedy clan starting with Joseph Kennedy (USA ambassador in the UK before the II WW) – John Kennedy (the 35th US President) – Edward Kennedy (Senator) – Caroline Kennedy (John's daughter – US Ambassador in Japan). Then the Bush Clan with George Bush (43rd US president) – George W Bush (45th President) and Jeb Bush (43rd Governor of Florida and the presidential candidate in 2016 elections). In the UK, there are at least 90 families, whose members were propping up each other in politics, such as most recently the Miliband brothers. This is ten times more than in France (9 families). Nepotism in the UK has reached such a level, that there is a new law proposed to curb it. According to the Press Association, in 2012 there were 151 of the 650 MPs at Westminster, were employing family members using their allowances for staff.

The second problem connected with the elites and clan politics goes hand in hand with perhaps a natural human tendency to cling to power. In most democratic countries, not to mention autocratic regimes, there is no limit to the number of parliamentary terms. The main goal of most politicians is to get into power and cling on to it, which is fertile ground for corruption. It was Lord Acton, the 19th century British politician, who said, "Absolute power corrupts absolutely" and that applies to most of the countries, which consider themselves democratic. The best example today is Russia, which is formally a democratic country but where almost an absolute power is held in the hands of the president – Vladimir Putin. No wonder Transparency International ranked Russia in position 131 among 176 countries.

On the other spectrum of the democracy crisis we have the imbalance of rights and responsibilities, freedoms, and restrictions. All of us would love to have unrestricted freedoms but from today's perspective it is a dream. Freedom to surf the free Internet, is just one such example. That is what we do daily, where we provide our private details to a company that gives us a 'free' application in return. That 'return' could be a restriction to our freedom of privacy in various aspects, including knowing our voting preferences. That's what was discovered in March 2018 in the Cambridge Analytica and Facebook scandal. Both companies are suspected of stealing personal data of about 50 million users to enable political parties to carry out a personalised marketing campaign to impact the outcome of Brexit and the presidential elections in the USA in their favour.

The combinatorial effect of generational divide and social atomisation on current trends destroys a notion of citizenship, as it has been conventionally understood for generations. Older people hope to live for decades more: their sense of

obligation to younger people is limited by a general human reluctance to give away existing advantage. We hope, naturally, that our children and our neighbour's children will do well in life. But what are we prepared to do about it? Very little, it seems.

In the last few years, we have a new term in politics – political symmetry. What it means is that most voters consider that each party is essentially the same, on average bad and, with no real intentions to realize the promises it has made in its election manifesto. At the same time, there are parties and electoral programmes that are substantially inferior relative to other parties, more detrimental or even dangerous from the point of view of the voters. How can you compare the Weimar Republic's election programme with the programme of the Nazi Germany's NSDAP party, or in the most recent USA elections – the programme of Donald Trump with that of Hillary Clinton? To make a reasonable choice you would have to know a lot more. But politics has become so complex that to arrive at the core truth based on an exchange of arguments during a debate would take a lot of time and require considerable broad knowledge in many interlinked areas. Not every voter has that knowledge and time to listen to the whole argument and then be able to make a rational choice.

And that is the problem in modern societies, especially during election campaigns, which populism skilfully uses to win votes. It simplifies the issues being debated, by either telling half of the story convenient for populists to convey their message, mostly true, and totally omits or distorts the other half, far more complicated, which if told would have completely reversed the original conclusion. One of the best examples is how in 2016, the BBC reported two sides of the Brexit campaign. It tried to stay impartial by giving each side the same time to reply and ensure that each side had the same chance to present its argument. In such a situation, the Remain campaigners, whose arguments required far more time to explain a problem because of its complexity, seldom had any chance to explain the real issue properly.

One of the questions you may have is how democracy can survive such pressures that come to the fore very clearly during the election debates. Should every voter, including those who have hardly any knowledge, or are illiterate have the same electoral rights, as the ones who have much better judgment? It is a difficult, almost existential problem for democracy, which should not be ignored, and which needs urgent solutions.

Could we replace democracy with something better?

If democracy has been failing people's expectations, then perhaps we should abandon it and replace it with something that might work in these difficult times, i.e. by a benevolent autocracy? This is a temptation one might go for, which would mean applying the Roman Republic's rules with the Cesar and the

Senators making 'best decisions' in the name of the plebs. Over the centuries there have been several such examples:

1. The Soviet Union, with its First Secretary and the party, ruling in the name of the Proletariat, justified as the Party claimed, because otherwise the capitalist class would keep oppressing the masses
2. Hitler and the NSDAP Party (which also had 'socialism' in its name) ruling on behalf of 'Deutsche Volk' – justified by Hitler saying that Germany needed more territory to expand (Lebensraum)
3. What may surprise you, even the French president de Gaulle's rule in 1959-1969 might be considered autocratic. His justification was that France was in existential danger because of the war in Algeria and the frequent changes of the government (every few months). That required a strong president elected for 7 years (now for 5 years)
4. Current Chinese autocratic rule, modelled on the Singaporean autocracy/semi-democracy introduced by Lee Kuan Yew, may be considered a system, in which the 'elite' knows best, what is good for the nation
5. Even today, in view of climate change existential risks, there are people like James Lovelock, the author of the well-known concept of Gaia – mother Earth, and Martin Rees, former UK Astronomer Royal, who advocate a view that perhaps democracy should be postponed 'for a while' because the danger for Humanity is so imminent and catastrophic, that an authoritarian rule may be a lesser evil.

But then, Winston Churchill's in his famous statement said that "Democracy is the worst form of government, except for all the others", and I would agree with him. That does not mean than an authoritarian rule for the world may never be an option to save Humanity, such as China, as the last resort, or a lesser evil to save Humanity. However, we should at least try our best for as long as it is possible, to improve democracy to save Humanity, rather than reject it.

Looking for the best democratic system

So, how can a modern democracy square this circle: delivering the fastest possible improvement in material life, maintaining at least the current level of peace and safety against existential risks, while retaining basic individual freedoms. China's incredible success in material progress, e.g. getting 600 million people out of utter famine in just 30 years, provides an uncomfortable answer. How could a new democracy deliver a better approach than China's both to prosperity and to an individual freedom (which is of course severely curtailed in China). I propose some solutions within the Consensual Presidential Democracy framework, discussed in the next chapter. However, even now, we can see that the root problem is how democracy functions. It is the 4 or 5-year election horizon, which determines that most large-scale decisions such as education, health service or infrastructure projects are looked at by the politicians from a short-term perspective. The Chinese on the other hand look

from a very long perspective. Therefore, for example, they have invested heavily in education and in apprenticeships, which allowed them to deliver spectacular projects like building a hospital in Wuhan form scratch for 1,000 people in 10 days.

Therefore, it is not easy to determine, which democratic system could consider the problems that Humanity is facing right now and, equally importantly, in the near future? The conclusions from the review of the existing democratic systems that I have made in my earlier book – "Democracy for a Human federation" (14), make it clear that there is currently no democratic system, which would support the purpose and the main objectives of guiding Humanity safely through the period of extreme risks, until such time when we will be able to evolve into a new species.

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. For example, none of the main three groups of democratic systems: Parliamentary Democracy, Direct Democracy and a Presidential Democracy could fulfil these criteria:

- There is no democratic system in the world that would guarantee in its constitution the self-determination of a region leading to setting up a separate state (the best recent example is the case of Catalonia). Even if such articles exist, they always have a caveat that the region must first seek the consent of the state, from which it wants to separate
- There is no system of government in any democracy that would envisage a strong separation of legislative and executive powers by forbidding most of the MPs to sit in the government (apart from some key posts such as the Prime Minister, Finance Minister, Defence and Home Office ministers)
- There is no democratic system whose constitution would facilitate governmental powers enabling it to act effectively in fighting existential risks that face us all, i.e. the entire Humanity. That is of course logical, since only the government acting on behalf of the whole Humanity would need such a prerogative.

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

1. To facilitate the federalization process of the planet
2. To apply only a very shallow level of federalisation, so that it will centralize only the very essential powers, leaving the rest of decision making at the lowest possible level of governance

3. To significantly reshape the relationships between the governed and the governing instilling more trust through greater transparency and continual accountability
4. To protect Humanity from existential risks that may emerge from global political, social, and economic disorder through combinatorial effects
5. To protect Humanity from other existential risks, especially coming from Superintelligence
6. To prepare Humanity for the time when we will coexist with Superintelligence, which potentially could start the best period in human history (if we mitigate the risks successfully)
7. To prepare Humanity for an even more challenging task – a gradual merging of our species with Superintelligence.

A proposal, which meets at least some of these criteria is presented in the next chapter, as a new type of democracy called **Consensual Presidential Democracy**. Since most of these reforms, and similar proposals, go against the politician's personal interests, they may probably only be implemented under duress (social revolution) or when one of the existential risks materialises. However, we should try to advance as many of the changes proposed here, or by other researchers, as soon as possible, to reduce the scope of democratic reforms that may be needed when we will be forced into that situation. That includes the tasks related to saving Humanity from various existential risks, including Superintelligence, which is the core subject of this book.

Chapter 3
Consensual Presidential Democracy for New Times

Four Pillars of Democracy

We have concluded that currently, there is no 'off the shelf' system of democracy, which would protect Humanity against existential risks, such as posed by Superintelligence. Based on what has already been said, I propose a new type of democracy – **Consensual Presidential Democracy** (CPD). The overall assumption underlying CPD is that we can only survive the extremely dangerous transition period to the time of coexistence with Superintelligence if we work closely together. This means **a gradual federalization of the whole world**. We must act as a swarm of bees protecting the hive, knowing that our safety is in numbers. Only then can we minimize the risk for humans' extinction.

I realize it is a philosophical and a political minefield. It would thus be easy to dismiss certain proposals, especially if they are discussed in isolation from the entire system that underpins CPD and its overall objective. In the end, it is a question of the level of risk that we accept in any sphere of life. However, this choice affects all other choices because it is a choice between the existence and the extinction of the entire species. There will be no winners here, neither individual people, nor certain states. The only winner could be Humanity as a whole by delivering a benevolent Superintelligence and in this way continue its existence as a biological species at least for some time. The only way we can achieve that is by changing the way we govern ourselves. This includes a deep reform of democracy based on such frameworks like CPD, based on four pillars:

Four Pillars of Consensual Presidential Democracy

Thus, the **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 moderator between two opposing parties, each represented by one Vice President.

The principles of Consensual Presidential Democracy, explained in detail in my previous book 'Democracy for a Human Federation' are put forward for consideration to be embedded in the Constitution of the Human Federation, which could become a legal framework for all laws, based on the values accepted by a significant majority of the nations, establishing a new democratic order.

Such a new style of democracy will have a better chance of supporting the future Human Federation and indeed any other organization, or a state. The key aspect of Consensual Presidential Democracy is a system of governing with a maximum consensus, where the voice of the 'losing' minority is always considered. You will find a summary of the four CPD pillars in the section below. Detailed description can be found in my book "Democracy for a Human federation" (14).

Balanced Rights & Responsibilities – Pillar 1

Balancing the rights with responsibilities is the first of the four pillars of Consensual Presidential Democracy (CPD), which applies both to an individual as well as a state. The overwhelming focus on human rights has created an unhealthy imbalance by barely mentioning the importance of responsibilities in maintaining social cohesion. We see it quite often in courts across the EU countries, when an offender seems to have more rights than a victim. It is a clear evidence of how sensible liberal values have led to the so-called political correctness, seriously undermining the political and social stability. The pendulum of liberalism may have shifted too far towards the rights. Therefore, in a new system of democracy, like CPD, the reduction of the imbalance between the rights and responsibilities plays such a prominent role.

Political Consensus – Pillar 2

One of the biggest differences between the European and the UK model of post-war democracy is that the first one produces mostly coalition governments, whereas the governments of the UK have been run almost exclusively by a single, majority party. That is the outcome of the First Past the Post system but also the belief that 'strong', one party rule is more efficient and more effective in delivering better quality of life for the electorate. However, the actual results do not confirm that, if we measure the UK's quality of life by GDP per capita, which has been consistently falling. For example, in 1990, UK's rank in GDP per capita in the world was 18th, whereas in 2018 it was 26th.

In my view, the biggest disadvantage of a single party government is the adversarial nature of politics as was evidenced so plainly during the UK's Brexit proceedings in the Parliament. This leads by extension to a deep polarization of society and resulted in Brexit in the UK. But there is even a greater disadvantage that really shows up in the longer term. The adversarial politics based on the majority of one party, which does not have to win the majority of the votes to rule the country, leads to short-term politics and constant swings in policies (no double majority is needed, i.e. the majority of MPs and over 50% of the votes in elections). The whole focus of the government is on winning the next election by tuning its manifesto to temporal whims of the electorate. If we return to Maslow's two lowest levels of the Pyramid of Needs (physiological and safety needs), that is exactly how people would respond. And that directly translates into the voters' preferences to elect those, who give more and now – an ideal platform for populism.

Additionally, such an adversarial political system supresses by its very nature the inflow of new ideas by virtually eliminating smaller parties, especially in the First Past the Post system. The voters have less choice and therefore are quite often either not voting at all, or voting tactically, which only rarely delivers the intended result.

Perhaps we should then consider coalition governments for the new, reformed democracy? Unfortunately, the answer is more complex since coalitions also have their disadvantages. Just think about the influence that a dozen DUP MPs in the 2017-2019 British Parliament had on the outcome of the Brexit proposals. It is immensely disproportional to the number of voters supporting that party. Furthermore, some people may still remember the fate of the Liberal Democrats coalition with the Conservatives in 2010-2015 UK government. For a coalition government to be established, each party in such a coalition must drop some of its Manifesto commitments, as UK Liberals had to do with their promise not to charge any University fees in 2010 elections.

Finally, let us consider the minority governments, i.e. emerging from a single party having the highest number of MPs but with less than 50% of the seats in the parliament. Usually, such a government must get the support of a tiny party on a case by case basis. In the UK this is called confidence and supply arrangement like with the DUP in 2017 Theresa May's government. That is similar to the model practiced in most Scandinavian countries and sometimes called **contract parliamentarianism**. In this model, the government passes a particular law if it can command the support of most MPs. Ad hoc coalitions can thus be formed for passing a single law. I would consider this model the closest to what the politics of consensus means.

However, to run such a democratic system smoothly, one needs an independent arbiter. In Scandinavian countries that, very active, role is played by the president. That model achieves the double majority rule, where most MPs and

most voters (in a proportional voting system, which is another important ingredient) support a given Act of parliament. However, that system does not guarantee that any legislation that is sometimes urgently needed will pass through the Parliament. Therefore, Political Consensus politics must also rely on some additional arrangements, such as those proposed by the Consensual Presidential Democracy (CPD).

Shallow Federalization – Pillar 3

This pillar deals with federalization and internal matters of member states. Therefore, it is bound to be very controversial. However, I have been trying not to shy away from such matters for political correctness, difficulty, or other reasons. This is an area that may affect the formation of any federation, including the future federalization of the EU.

However, it is also very relevant today. The best example is the Catalonia's referendum on independence carried out on 1st October 2017, which had not been previously agreed with the Spanish central government. Did the Catalans have the right to carry out such a referendum without the consent of the Spanish government? The illegality of the referendum is crystal clear. That's what Article 2 of the Spanish Constitution says. But not having a legal right does not close the problem. If Catalans do not have a legal right to organize such a referendum on the region's independence, do they have a moral right not only to the referendum but to becoming an independent state? In my view, they have such a right based on three principles:

- The first one is the so-called Natural law ("lex naturalis" in Latin). It asserts that "certain rights are inherent by virtue of human nature.
- The second one is individual freedom, indirectly derived from Natural Law, practiced in ancient Rome as "habeas corpus".
- The third argument is the Right to Secede, which is frequently used by international lawyers.

Therefore, in the future Constitution of Humanity there must be articles on a region's or a nations' secession from the member states. To make it easier to create any federation, it should be set up from the outset as a 'Minimal State'. That is not just a phrase but a whole concept of a state, favoured by liberal philosophers such as Emanuel Kant who viewed freedom as 'the absence of external constraints upon an individual'. More recently an American philosopher Robert Nozick expressed the notion that 'a state must possess two main attributes: it must have a monopoly on the use of legitimate force in a territory, and it must provide protective services for everyone in that territory'. Henry Usborne, in his book 'Prescription for Peace', published in 1985, calls such a federation a MiniFed. What I believe is important in the context of the Human Federation is that living in such a state is a kind of a bargain – greater safety for less freedom.

And that is precisely why I would think a Minimal State might be the very right political structure for the Human Federation or any multi-national federation. In such a state, its duties are so minimal that they cannot be reduced any further because otherwise the state would cease to exist and would become a form of anarchy. Typical governmental institutions in a Minimal State would be the defence, foreign affairs, federal police, and the federal judicial systems. Individual nations within such a federation would have the right to keep local police, and have local laws, but would not be allowed to have their own army. It is then obvious that a Minimal State is certainly not a Welfare State. That would be a continuing prerogative of the former states or large regions in a federation.

AI-assisted Governance – Pillar 4

If we survive relatively unscathed until about 2040 – the time of the assumed formation of the Human Federation (HF) – our civilisation will be ready to deliver unimaginable wealth to everyone on the planet. But to deliver the world of abundance we will need a very efficient World Government – the executive body of the HF. But even before then, a reformed democratic system should propose solutions for the governments to be more efficient and effective. That is the objective of the fourth pillar of Consensual Presidential Democracy (CPD).

How can we do that when 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 a civil service and thousands of advisers and consultants but in the end they themselves must make the final decision. The problem is that quite often such a decision requires 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 the HS2 railway project in Britain, which is to be completed in 20 years and cost over £100bn. That is what choosing the wrong type of project can lead to.

So, why not to have a technocratic government? 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. Unless the whole political system is a blend of democratic and authoritarian rules, like in Singapore, such governments are not here to stay. Therefore, in the pursuit of effective and efficient government we need to look for other options.

What I propose here may significantly impact, political decision-makers at any level of governance, i.e. ministers, governors, mayors, councillors etc. The solution that I consider involves the support of politicians and decision makers at all levels of governance by AI assistants. This will happen anyway on a grand

scale in just the next few years, when almost every profession such as medicine or engineering will be supported by such AI assistants.

If you think it sounds incredible, then just look at the offerings of one company – Generis. It has already several industry-specific AI Assistants. For example, CARA (Case Analysis Research Assistant) can work in most 'soft' areas such as law, pharmaceutical, or in government. It is competing with ROSS, an AI legal assistant, which has already delivered incredible results, especially in the Anglo-Saxon world, where law is case based. There are other such AI Assistants in the legal area where they deal with thousands of documents per case, so they are engaged in similar tasks as in most government departments.

If you consider continuous self-learning of such AI assistants like IBM's Watson, or more popular, but also potentially game-changing solutions, like Amazon's Alexa or Google's Assistant, then within a few years, work in many companies in these industries will change beyond our imagination. The easiest way to imagine such an assistant at work is to visualize a humanoid robot driven by Amazon's Alexa-type application. Today, such an application can communicate in perfect, easy to understand accent, in about 60 languages. We can only understand what the app is saying but it has serious difficulties to continue a natural contextual dialogue. Therefore, quite often its response is just 'I don't know that one'. Only the very best, most expensive robots, linked to superfast computers, such as Sophia by Hanson Robotics, can have a longer meaningful dialogue. However, according to the company, it needs another 2-3 years before its Sophia will be fully conversant on most subjects. Ray Kurzweil, the futurist, whom I mentioned before, says we will need to wait till 2029 (he is precise about the date), when AI will achieve human level intelligence (in terms of processing power not intelligence as such).
At that time, almost every decision made by a political decision-maker or any consultant might be executed as the AI assistant had suggested. Until then, these robots will be capable of advising on a narrow subject matter using their knowledge database. Such databases are already being produced as plug-ins (see CARA and ROSS above), purchased as a service and then maturing through self-learning in a specific environment, e.g. at the Ministry of Health. Therefore, realistically, we can expect a widespread use of such assistants by about 2025, although probably with very limited cognitive capacity yet.

The benefits gained by the government of a country implementing such an AI-assisted governance will be immediate and significant. First, most decisions will be made many times faster, with full justification and with 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 truly vast if implemented at all levels of government.

Finally, there will be very few purely 'political' decisions to win the votes in the coming elections since the planning horizon for most of such projects will cover a decade or more. Additionally, should there be a legal requirement (say in 10 years' time) 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. That should not be a surprise at all. If you agree that in about 30 years' time Superintelligence will become our benevolent dictator, then what would be practiced in the intermediate period is just a preparation for what will happen on an unprecedented scale in every step of our life anyway.

Such implementation would allow politicians to have a personal, direct control on even the largest initiatives and projects, executing them with incredible effectiveness and efficiency. The added benefit will be a continuous parliamentary scrutiny, should such a politician be an MP. To make the best use of these assistants, say from 2025, they will probably be best used as additional advisers to humans. However, they should be physically present in a humanoid form in their 'place of work' for several reasons. For example, if it is in a physical, humanoid form, hardly distinguishable from humans, it will also move around almost like most of us, explore and learn about its environment, listen to conversation, 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. It will learn our preferences, rather than simply, goals. That can only be experienced in a real physical environment by a real, 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 specific AI Assistant, as having some rights – e.g. only certain people will be able to make highest level decisions, to switch off the assistant, if needed. 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 a nation or a given community. The only exception might be when such 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.

5

Part 5
SURVIVE AND EVOLVE

Before you move on…

In Part 3 I have argued that we can minimize the risk of delivering a malicious Superintelligence if we teach it the Universal Values of Humanity. The problem which I have explored in Part 4 was, that we would need first to significantly amend the current charters of human values and deliver a new set of Universal Values of Humanity. They should cover not just human values, and the corresponding rights, but also the values and rights of the rest of the wider ecosphere, including all species, which have no means to claim their rights. This may also include the rights of AI agents e.g. of not being switched off, once they reach a certain level of intelligence and perhaps even consciousness.

The problem that we face is who will agree those new values and responsibilities on behalf of the whole Humanity, when it took 25 years for just 50 UN members to sign the UN Declaration of Human Rights of 1948. But we need such a charter within this decade.

Therefore, the only practical way to achieve a somewhat imperfect solution is to create a Supranational Organisation. I have explained in Part 1 that creating such an organisation from scratch is virtually impossible. Therefore, we need other solutions. That is what we shall be discussing in Part 5.

Chapter 1
Can the UN become the World Government?

The lessons from Covid-19 pandemic

The Covid-19 pandemic has shown very clearly how unprepared the whole world has been to fight a threat, which on the scale of truly existential risks (in terms of its direct mortality) would not even register. Politicians say, this has been something entirely new, so no wonder that nobody was prepared. That is hardly credible. Most governments have made some preparations for such pandemics. The problem is that such preparations have been utterly inadequate. It does not matter that it is a pandemic. Any other existential threat of a similar duration, such as a super cyber-attack created either by a mad scientist or by a self-learning AI agent, would also find the world totally unprepared with consequences far greater than the 2020 pandemic. Just imagine, that an electricity power cut of a few weeks in the USA in winter, may lead to millions of deaths in that country alone (the USA is made up of five different grids) (26).

This pandemic has shown that the world has already become a planetary village. The problem is it does not have a sheriff. Until we have an organization that would resemble the World Government, we will be unable to act swiftly, when the whole planet is in danger. Imagine a situation where we have the World Government with a powerful president. He issues an order: any airline that wants to fly passengers must do the following (this is just an example – the list could be expanded, or the items changed):

- It can only fly from or to countries with the virus reproductivity index (R) of less than 0.5
- Any passenger flying out must have his temperature measured and tested
- Any passenger must wear face masks all the time until he lands in the target country, where he would be tested for the presence of the virus.

Had such conditions been met world-wide, the risk of contagion would have not fallen to zero, but it would have been drastically reduced, enabling airlines to fly, hotel resorts to operate and the economy getting back to normal much faster.

Could the UN fight successfully existential risks?

Remember, change happens at a nearly exponential pace rather than linearly. At the same time, we live in a world that is exposed to many existential risks. These are the risks, which we have created ourselves, rather than the risks stemming from other factors such as an asteroid hitting the earth, or super gamma wave radiation. Climate change is one such risk, but it is not the most imminent or at the top of the list.

Superintelligence, on the other hand, could eliminate 100% of humans by about 2050. But already within this decade we will have to deal with what I call an Immature Superintelligence. It is almost certain that by 2030, we may have a number of large-scale crises created by Cyber Wars or initiated by malevolent AI agents, which could launch nuclear weapons, open water reservoirs, or open the doors of biological laboratories breeding new viruses. To mitigate those risks, we would need a global authority that would be able to act decisively and almost immediately on behalf of all of us, e.g. in case of untreatable global pandemics stopping all transport worldwide. Only the World Government would be capable of mitigating such risks but that has been something that we have not been able to do from the time of the League of Nations.

If the world is not able to form even a de facto World Government by the end of 2030, we may be unable to control Superintelligence and will expose ourselves to the biggest risk Humanity has ever faced. We should look back at the historical events, which occurred in the last century, to see what may happen to our civilization if we continue to rely on organizations such as UN. It is a great organization in many areas of human endeavours, such as medicine (WHO), education, culture (UNICEF) etc. However, what really matters, is its lack of political power to enforce the decisions that it makes. The UN, and especially its Security Council, where any decision requires unanimity voting of the five major powers, is in most cases powerless. The consequence is that it cannot resolve conflicts, such as in Syria or in Yemen. The World needs a powerful World Government. But the UN has never been and will never be able to create such a government.

So, why have we not tried to create a new supranational organisation? Perhaps one of the reasons is that many of us were hoping for the UN to take up such a role. After all, this is the organisation that should deal with existential risks in the first place. Unfortunately, this is also the organisation that indirectly increases the Humanity's overall existential risks by being almost totally ineffective in solving serious problems. The UN may be trying to apply some solutions but often when it is too late, or after a grave human, economic and ecological danger has already occurred. In most cases it is incapable of solving the problems at all (e.g. Syria or Libya). The best evidence is the creation by the UN of the International Strategy for Disaster Reduction (UNISDR), the organization that was established in 1999. It has had hardly any influence on any of the top existential risks listed earlier. The only real impact that the UN has, considering its actual impact on legislation, are the International Nuclear Atomic Agency, the International Panel on Climate Change (IPCC), and WHO in the health area, but even these have not worked very well, as the Covid-19 pandemics has shown.

The reason behind this situation is simple. Such an organisation to act successfully would need powers far superior to the UN has right now. Since we know how ineffective the UN has been, it is little wonder that the world is

practically left to be gradually sucked into the funnel of ever-growing existential risks. Initially this will be hardly visible but further into the future, the risks will tend to combine, so it is possible that at such a time there may be no organisation or a country powerful enough to stop the demise of our civilisation.

Transforming the UN into the World Government is also a futile hope, because of the very way the UN makes decisions – the unanimous voting system in the Security Council. The presence of the unanimous voting for most decisions made by the European Council is the same reason why the EU has been less effective than it could have been. Fortunately, the Lisbon Treaty has finally created the possibility for the European Council to vote using a qualified majority on more subject domains than ever before. But even this is far inadequate as the Covid-19 pandemic has shown.

So, the answer to the question: 'Could the UN fight existential risks successfully?' is – No. But is there any alternative? I believe there is. Since we have no time to create such an organization from scratch, we can only hope that one of the existing organizations would step in.

What is the probability of such a de facto government coming into existence? In 'normal' circumstances, if over this decade we have no world war, no nuclear conflict, no catastrophic pandemics, and no major AI-related catastrophe, then the chances of forming such a government are, in my view, close to zero. Paradoxically, the only hope that we may have for creating a de facto World Government (which would include most, but not all countries) is that some of those catastrophic events, caused by the Immature Superintelligence mentioned earlier, will occur. Perhaps Covid-19 pandemic could become such a warning that the creation of the World Government is necessary.

Similarly, after the Cuban crisis in 1962, when the world had significantly reduced the number of nuclear weapons (still a long way to go), such a serious event may convince many nations that we must make some sacrifices and can only be safe if we act together.

That will mean passing part of the national sovereignty and individual freedoms to a strong federal government that would have the best chances to keep us safe.

Chapter 2
In Search for the Best Candidates

Who Could Form a de facto World Government?

Let's step back and ask ourselves, why we need a federated world in the first place. In a nutshell – to survive. Safety is in numbers. That has been practiced by humans since the dawn of an intelligent human species (and even before). Nothing has changed since then. The first thing people want is to retain their LIFE! Therefore, safety will always come to the fore. If it can only be gained by shedding some of our freedom and sovereignty and there is no other way, so let be it. That is the reason for having a federated world.

In Part 1 I have covered the area of man-made existential risks, now combined with a nearly exponential pace of change. That has shown clearly that the only way in which we can solve planetary problems is to work as a planetary civilization. That's why we need a Supranational Organization with a powerful World Government. Who could then guide Humanity and protect our civilisation in the following 20-30 years probably the most dangerous period in the human history? By the end of this period, we may already be coexisting with Superintelligence, which will hopefully help us sort out our problems.

Let us also remind ourselves that the key question is which organisation is potentially the best one to control the risk stemming from AI when it achieves the status of Superintelligence and ultimately becomes a Technological Singularity. But Superintelligence is only one of the existential risks that need to be mitigated. Therefore, any organisation that we choose to act on behalf of the whole Humanity must be capable of dealing with other risks too, including Global Political and Social risks.

An imminent, and one of the biggest risks for Humanity, is that our civilisation will just keep going on until the point, beyond which saving humans might be a futile effort. To change the course of civilization, so that it progresses as a uniform global entity requires, the urgent formation of the World Government. However, a rational discussion among world leaders, like within the UN, on forming such an organisation will be rather unthinkable and utterly unrealistic, especially when such an organization needs to be created by latest 2030. That's why I repeat it again, the only feasible option that could become operational relatively quickly it to select an existing organisation that would gradually start to act as a de facto World Government.

Who might then fulfil the role, which the UN can't? Well, anyone who wants to improve the situation and reduce some of existential risks faces three problems:

- Existential risks require fast action, while the world's organisations act very slowly
- People want more freedom, while we need to sacrifice some of our freedoms and sovereignty for Humanity to survive
- Most people can't see beyond tomorrow and act emotionally, while we need to act rationally and see the long-term consequences of our actions.

Therefore, anybody that sees the need for the world to take an urgent action faces a difficult task when proposing pragmatic, fast and very radical changes to the way the world is governed. It seems to me that the only realistic route for humans to take, is to refocus an existing organization, which would have the capacity, resources and resolve to act on behalf of all of us in the hour of the emerging existential threat. To have any chance of successful delivery of its foremost objective, i.e. to protect Humanity against existential risks, such an organisation should have supranational powers exceeding any prerogatives of the existing international bodies, such as the United Nations, NATO, or WTO. **We need an organization that would resemble the World Government.** Only such an organisation, which should be operational latest by about 2030, would have some chance to mitigate existential risks mentioned earlier.

Therefore, the agenda of such an organisation should be governed by one key issue: **Fighting existential risks**. Any other objectives are subordinate to that goal, since if there is no Humanity and no civilisation there is no point in discussing other aims of such an organisation. 2030 is a threshold date, because by then we must have a full control mechanism over AI in place. Therefore, such an organisation must be a large federation before then, to put the mantle of a de facto World Government in place by the end of this decade.

Since existential risks can materialize at any time, e.g. a natural pandemic, like the current Covid19, or be triggered by laboratory-generated bugs being maliciously released into the open, we should have an organization that could act as a de facto World Government right now. But for such an organisation to be successful in mitigating existential risks it should have a mandate from all of us to act on our behalf. That of course will not happen. Would those who hold power in autocratic or dictatorial regimes give up their privileges and introduce a democratic government as part of the World Government? Certainly not!

It's a pity, but that is our world and our civilisation and perhaps that's one of the reasons why we are in such an existential danger. I believe we simply have no time left to wait for all the nations to agree to one common Constitution of Humanity, such as proposed by the World Federalist Movement and similar organizations. We must act right now and get most of the countries to agree to become members of an existing organization, which has the best chance to act on behalf of us all. So, it is almost certain that such an organisation would not include the countries such as China, Russia, Saudi Arabia or perhaps even the USA. This 'partial' World Government would then have to co-exist with the

countries outside this organization, which is a serious risk on its own. However, I believe Humanity has no other option and must take this risky path.

Such an organization would need to have the best experience in managing its expansion both in functionality and size, as well as have significant resources. Only that might enable it to gradually include more and more countries, ultimately converting itself into a Human Federation by about 2040. It will then have just one more decade to prepare for the moment to handover the responsibility for the future of Humanity to Superintelligence. By then, Humanity could reach the point when it may already be coexisting with Superintelligence. I have made a detailed assessment of 10 organisations, which could act as a de facto World Government in my earlier books (1) (14), so below is just a check list of the required capabilities for the best organisation to play the role of the World Government:

- Could the organisation execute supranational powers over a large part of the globe?
- Does it have, or will it soon have, its own army and rescue services?
- Will it be able to redefine human values that would become the foundation for the future new constitution and a legal system underpinning, what must become with time, a Human Federation?
- Could it ensure very fast and co-ordinated response in emergencies (potentially within hours)?
- Does it have a large reserve of emergency supplies of food, seeds, etc.?
- Does it have experience in dealing with large scale, global crises?
- Does it have long-term experience in democracy and the rule of law, so that any decisions are made according to democratic rules, acceptable to most of the population?
- Does it have enough resources, including financial, to deal with the current existential risks?
- Is it very likely that it will be open to free and fair criticism and will it act on it?
- Will it be able to adapt the way it works and introduce new laws very rapidly?
- Does it have almost immediate access to best scientists and practitioners in every domain?
- Does it have, or will it be able, to develop early warning system?
- Could it create a very large refuge for civilization (a physical space in case of a catastrophic danger, i.e. huge caverns, or tunnels)?
- Does it have and can it store large supplies of vaccines and medicines?
- Is it, or will it be, capable of reducing nuclear proliferation?
- Does it have or will it be capable of a strict oversight of molecular technologies?
- Will it be able to fight populism with facts?

We can now assign weights of importance to some of these criteria, with added justification, which will help us to select the most suitable organization as a de facto World Government:

Weight	Justification for 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**	The is the second criteria in importance for two reasons. The organisation must be exemplary in its respect for human values and it has to carry out the process of redefining them for the upload to Superintelligence
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 members and carry out missions to minimize the risk to humans such as Weaponized AI, wars that could become global or wars that count as genocide
7	**Economic power**	This is important because the organization must have enough resources to mitigate existential risks
6	**Organizational capability**	This 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, e.g. nuclear war threat or artificial pandemics.
4	**Land mass**	It 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**	This 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 important to ensure cohesion of the organisation that will have powers to reduce freedom or sovereignty.

Now, that we know what capabilities such an organisation must have to mitigate existential risks, we select the best candidate to play the role of the World Government. To select such an organisation, I have created a table with 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 GDP. The remaining 7 criteria are subjective, but that subjectivity is within a narrow margin, which over the 10 criteria will not make a big difference. The whole purpose of this process is to select an organisation, which is likely to be one of the top three candidates, whatever the weights. The results are presented in the table below.

Which organization could take the role of a pseudo World Government which will ultimately become the Human Federation?

Name of Organization or State	Risk Mitigation Capability Ranking (weighted)										
	Demo-cratic Insitutions	Respect for Human values	Military power	Economic power	Organi-zational capability	Response time to risk	Land mass	Experie-nce in large progra-mmes	Versatility	Neut-rality, Objec-tivism	Total Score (weight * capability)
Weight ----->	10	9	8	7	6	5	4	3	2	1	550
European Union	10	10	7	9	10	10	6	10	10	10	503
NATO	8	9	10	10	10	10	9	7	4	9	495
USA	9	9	9	8	9	9	7	9	9	9	480
Japan	10	10	3	6	9	9	1	5	4	9	391
Canada	10	10	4	4	9	9	4	3	2	10	388
Australia	10	10	3	2	9	9	3	1	3	10	358
United Nations	10	10	2	2	8	5	2	6	10	10	349
China	3	1	7	7	8	8	5	10	9	1	301
Russia	4	3	8	3	6	6	8	10	9	2	300
India	7	5	4	5	5	4	2	5	3	7	268

The EU would have to become the European Federation first and then expand its role as a pseudo World Government

If you look at the EU's contenders in the above list, you can see there is not a big difference between the first three countries but there is a big difference between the third (USA and the fourth (Japan). So, I will only make comments on the first three countries for each of the categories.

- **General remarks.** If the scores are the same for an organization and a single country, then a country gets one point less because it is much more difficult to achieve a given rank, in an organization composed of many countries, than in a single country. Therefore, USA can get a maximum of 9 points.
- **Democracy:** NATO was scored lower because of Turkey (autocracy) and Albania, Bulgaria, Romania and Montenegro and Slovakia (all have too high level of corruption).
- **Human rights.** NATO was scored lower because of Turkey (autocracy).
- **Military Power.** EU's military power score was the same as China's (because China is a single state and gets 1 point less). USA, the strongest power was scored 9 points because as a single state it gets 1 point less than a maximum 10).
- **Economic power.** No adjustments made.
- **Organizational capability.** USA was scored 9 points because as a single state it gets 1 point less than an organisation.
- **Response time to risk.** USA was scored 9 points because as a single state it gets 1 point less than an organisation.
- **Land Mass.** No adjustment made.

- **Experience in large programmes.** USA was scored 9 points because as a single state it gets 1 point less than an organisation. NATO adjustment due to experience in mainly military operations.
- **Versatility**. USA was scored 9 points because as a single state it gets 1 point less than an organisation. NATO adjustment due to lack of versatility and focus on military operations only. That however may change in the future.
- **Neutrality and objectivism.** USA was scored 9 points because as a single state it gets 1 point less than an organisation. NATO adjustment due to Turkey's operations in Syria and Iraq and autocracy of that regime.

If NATO takes up this role, it will have to expand its scope significantly since it lacks sufficient experience in other domain than defence. However, when the European Federation is set up and the European Army becomes part of NATO, merging the two organizations is a very logical and perhaps a natural step forward.

As far as the USA is concerned, the main obstacle here are the American voters. They are unlikely to agree to constrain their personal and national freedom and see the USA as just one of the nations within such a federation. For example, USA is one of the countries, which are not members of the International Criminal Court of Justice (currently, there are 123 members). It is also worth mentioning the futility of imposing any sensible gun control in the USA, to see that the USA is unlikely to agree to what they would see as a limitation of their sovereignty.

Finally, it is quite possible that none of these options will bear fruit by 2030. In such case, my default option would be China. Just consider that every sixth person on the planet is Chinese. But do we want Humanity to pass control to the next species with Chinese values alone? It is a challenging thought that deserves a detailed and impartial analysis, which is beyond the scope of this book.

In the end, even if NATO or the USA had been chosen instead of the European Union, the whole process of change that would have to be applied to any of these three organisations would have been very similar. However, the changes to be applied to convert NATO or even more so, the USA, into a de facto World Government would have been much more difficult. In any case, I consider the EU as a kind of a strawman to see what kind of organizational and political changes the candidate who would act as a de facto World Government would have to go through.

European Union as the best candidate to save our civilisation

A significantly reformed European Union, transformed into the European Federation, seems to be the organisation that could quickly achieve the status of a de facto World Government and later, by simply having a critical mass, become a Human Federation. It could be gradually transformed, initially

embracing perhaps only 30-35 countries, from the current confederation status into a full Federation, in a similar manner to the Eurozone expansion. It is already planning to take on new members, so in the next 10 years we may have other countries such as Ukraine or Georgia as members of this re-invented organization. Incidentally, even pope Francis, in his 2020 Urbi et Orbi message has endorsed such a selection by saying: "The European Union is presently facing an epochal challenge, on which will depend not only its future but that of the whole world".

Let's now look more closely at the EU's capabilities, its strengths and weaknesses and the scope of its reforms, so that it could start acting as the World Government. In general, the EU has already quite a few features that are important for that task such as:

- Nearly uniform human values and legal system
- A wide spectrum of activities comparable with the UN
- A lot of experience in large international projects, like the accession of 10 eastern and central European countries on 1.5.2004
- Significant financial and material resources
- A system extending beyond a typical confederation, with the president (President of the Council of Ministers), the Government (the EU Commission), the prime-minister (the President of the EU Commission) and the Parliament (the European Union Parliament)
- Dynamism. What may surprise some people, is that there are very few other large organisations in the world that are as dynamic as the EU. Over the last 60 years the EU has been continuously adding new members, significantly changing the way it operates and continually distributing resources to poorer members on a very large scale.
- Ability to expand rapidly by integrating more countries, which are themselves significant global powers, such as Canada, Australia, and Japan, with which the EU has already signed wide-ranging treaties.

If life rejuvenation is successful, then most of you will see the first day of the 22nd century. But to arrive there, we need to go through a stage of transition. We need to begin the process of federalization of the Planet. And that should start with the federalization of the EU.

Chapter 3
The European Union becomes the European Federation

Why does the EU have to become a Federation?

Note: *A detailed description of the federalization of the EU, including the outline of the future Constitution, can be found in – Volume 2 of POSTHUMANS: "Democracy for a Human Federation".*

Existential risks force us to take extraordinary steps to save human species from extinction. Artificial Intelligence is our biggest and most imminent existential risk. However, at the same time, it could still help us to make a transition to the new époque when humans will be co-existing with Superintelligence, by re-designing and implementing the new world order. Since we have a very limited time, we must rely on the best organizational, technological, and material solutions that we already have. As I have argued in the previous chapter, the most feasible way forward is to entrust the fate of Humanity to a widely reformed and federated European Union. Therefore, from that point of view alone the future European Federation must:

1. Become a credible and effective organisation that will be able to mitigate all man-made existential risks
2. Ensure global political, economic, and social order necessary to focus all efforts on our survival, since lack of a reasonable global stability may become, through combinatorial effects, an existential risk on its own
3. Endeavour to preserve what is best in Humanity as a kind of a treasure chest that at some stage must be passed on to Superintelligence. These are our values, rights, and intellectual and cultural assets, to avoid the emergence of a valueless Superintelligence, or equipped with dangerous values and objectives that in the end could destroy us all.

But then we must also look at the prospect of creating a European Federation (EF) from a narrower, European perspective. Over the last 60 years of its existence, EU has managed to achieve something unprecedented in the European, and the world's history– peace, continuous economic growth, and relative social order. That pedigree and experience has tilted the odds for selecting the EU as the best candidate for expanding its role, and after its gradual transformation into the European Federation, serve wider aims, than just to meet purely European objectives.

For the EU to transform itself into such a federation is an existential necessity. Its member states must complete such a transition into a federated state within a decade, perhaps even by the middle of this decade. Otherwise, the EU's inherent

inconsistencies and inflexibilities originating from significant economic, social, and cultural differences (even within the same Christian culture) will gradually rapture its structure leading to its disintegration. That could start a period of political instability, which would almost inevitably be exploited by Russia, leading to European wars with most disastrous consequences.

A possible structure of the European Federation State

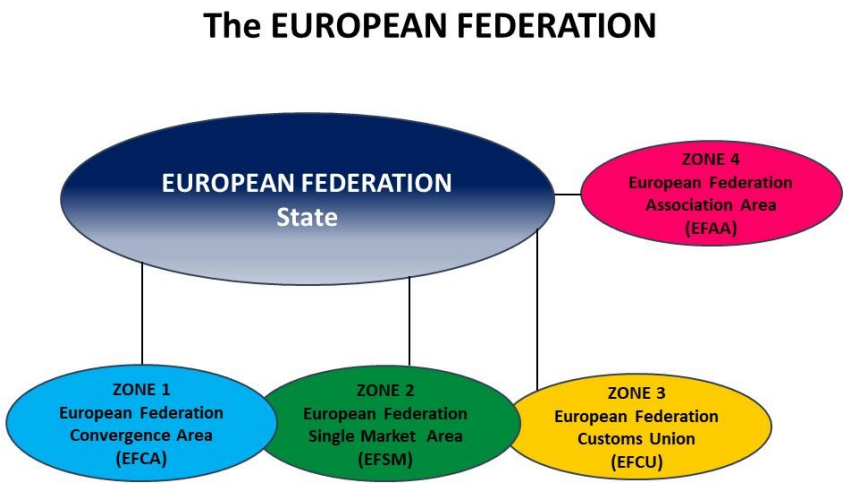
We must recognize the current real situation in Europe. Most Europeans do not want a federation – they want more sovereignty. That creates an additional barrier for any politician with a vision, who understand the real circumstances in which Europe and the world have found themselves, and the urgent need for federalization of the EU. Therefore, such politicians should use every available opportunity and lower the entry barrier to advance the moment when the EU will finally become federated.

In my view there are three elements in such a lowering of the barriers in the transition from the EU to the European Federation (EF):

1. **Staged transition** - the current lack of flexibility of the states' political relationship with the EU has led among others to Britain leaving the EU. The EU has created an economic path for becoming a full member, by creating an Associate membership, Customs Union, and the Single Market. It needs a similar, largely political, path for the membership of the EF. Therefore, for a fast federalization to happen it must be based on the creation of subsidiary zones, giving the aspiring countries time to prepare for becoming part of the EF, while at the same time not blocking the creating of the EF in the first place.
2. **Fast track transition.** The EU cannot hope for the 'right', quiet moment to arrive, when it will be ready for federalization. That will never happen. Secondly, with Covid-19 pandemics, EU will now face even a greater economic and political chaos. It must federate now, or perhaps it will never happen, with disastrous consequences. Therefore, it should take the current economic and political readiness of the EU as the best it can ever get and start with a 'quick and dirty' federalization of those countries, which are willing and ready to do so. Most likely these could be the Mediterranean countries or the Eurozone. If article 20 of the Lisbon Treaty is used, see below, then the federated countries will remain as a whole, a member of the EU.
3. **Dynamic enlargement.** At the same time, the EU must look forward, knowing that safety is in numbers and that one day it will become a Human Federation embracing all countries of the world. Therefore, it should create easiest possible conditions for an on-going enlargement. The main conditions for the new nations joining the EF should be political, i.e. an absolute respect of democratic, rather than economic principles.

Taking the above into account it becomes almost obvious that **the fastest route to the European Federation is through the federalization of the current Eurozone members, or those countries with the highest sovereign debt and political and social tensions**. The option for the whole EU to federate at once is becoming less probably by the day. Simultaneously, the new Federation should create four subsidiary Zones. In effect, these will be transition Zones, which will allow member-states to move to the next Zone up at their own speed, ultimately joining the European Federation State. Doing it piecemeal is, in my view, also the best way forward because the federation will show other countries outside the EF how the Federation works in practice.

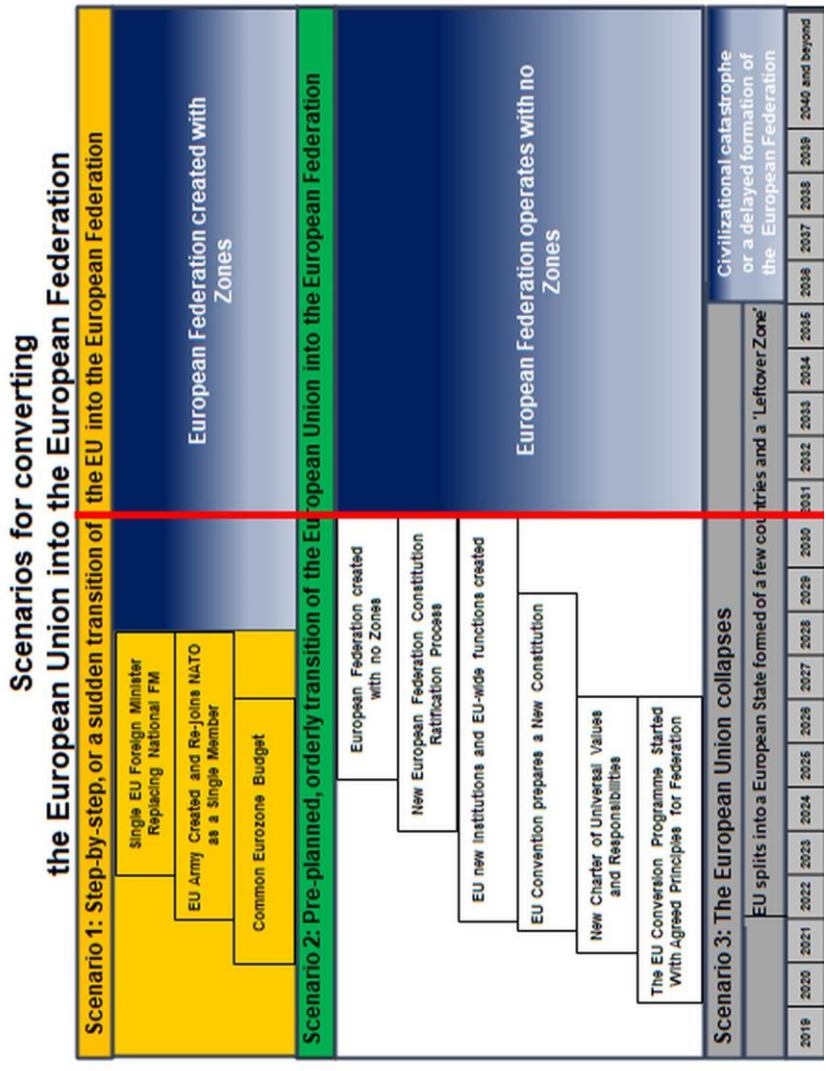
If the Eurozone becomes the EF State, then the remaining member states of the EU, outside the Federation would form the European Federation Convergence Area (EFCA) as the EF subsidiary Zone 1, with other zones created in a similar way to facilitate greater flexibility for the EF expansion. Here is the proposed structure of the European Federation:



Zone 1 is a member of both Single Market and Customs Unions. Members who are both in the Single Market (Zone 2) and Customs Union (Zone 3) do not necessarily have to move into Zone 1. Some members in Zone 4 could also be members of Zone 3 (Customs Union), like Turkey is now.

Scenarios for converting the EU into the European Federation

The Covid-19 pandemic and the start of the biggest recession that the world has ever seen, will undoubtedly widen the already existing cracks in the world's relative peace. Therefore, all existing plans of major political and economic initiatives will be changed or dropped altogether. What seems to be impossible now, may become necessary. This also includes the time by when the EU federalization process may be accomplished. Let me briefly describe these scenarios.



Source: Tony Czarnecki: 'Democracy for a Human Federation', London 2019

1. **Scenario 1 – Quite a sudden, unprepared transition** to a rudimentary federation, when only the most necessary functions would be federated, such as defence, security, foreign affairs, and the budget. This, in my view, is now the most likely scenario. The post Covid-19 economic and social crisis that has engulfed the EU, as it has the whole world, has paradoxically created an opportunity for an almost enforced fast-track federalization. The most likely way, in which it may happen is by invoking Article 20 of the Lisbon Treaty, by the EU members who want to federate. That article allows a minimum of nine members to integrate further (read 'federate'), while still requiring such a new Union to remain as a whole part of the EU.
2. **Scenario 2: Making a transition in an orderly fashion**, carried out in small steps, partly under disguise but ultimately leading to the implementation of an entirely new system of governance within the European Federation, when most of the current member states would join in. This has been the current, decades' long vision, started by the former President of the EU Commission, Jacques Delors, euphemistically described as an 'ever closer union'. Now, it is quite unlikely to happen in the new circumstances. It is a near certainty that in the ensuing chaotic attempts by individual states to minimize the damage to their countries, that there would be time and willingness for quiet, long and procrastinated negotiations, first on the new Constitution and then on the shape of the future federation.

   If the most likely scenario now, is a chaotic, fast-track federalization, then what might be a somewhat more orderly federalization, in such an unprecedented situation, caused by Covid-19? How can the EU federalization happen most effectively or even can it happen at all? Perhaps those supporting federalization should apply Dale Carnegie's rule 'if you have a lemon, make a lemonade'. Currently, the 'lemon' has two elements. The first one is the Future of Europe Conference, and the second one is the concept of Citizens' Assemblies - the approach to deliberate EU problems and make decisions by directly referring them to the EU citizens. Although formally federalization is not even muted in the initial document, it has created a framework within which, a relatively fast federalization of the EU might be possible. However, that would require rewriting the current proposal.

3. **Scenario 3: European Union collapses.** The longer the federalization is put off, the higher the risk that it may not happen at all, and the EU collapses or shrinks to a relatively small, federated state. Should such an unexpected, very serious event happen, the EU might be completely dissolved. This scenario has become more probable than even a year ago.

Covid-19 pandemic has shown us how unprepared Europe and the world is for responding to serious threats, even if they do not qualify as existential. A common background for these threats is something that we have barely noticed until now. Europe and the world have started to change at an exponential pace: what once took a year, can now be accomplished in about a week or two in almost every sphere of our lives. That pace of change, driven by unprecedented progress in technology, will continue forever, fundamentally changing the way we live. This has brought to us many benefits but also created new existential threats, such as Climate Change and loss of control over the self-learning Artificial Intelligence, which may both reach a tipping point in just 10 years, as well as a global instability. These threats will increase in the post pandemic super crisis, which will also trigger a permanent Technological Unemployment, causing social disorders. Therefore, we need a fundamental change in the way we live and interact as nations. The world must start the process of a planetary federalization now.

However, regarding the EU federalization, irrespective of whether Scenario 1 or 2 will be used to achieve it, the final shape of that federation must fulfil both the expectations of the EU electorate, as well as making it easier for the new members to join the federation. It must be based on a new type of democracy facilitated by a Constitution, which will renew the trust in the elected politicians. They must be truly accountable throughout their term and not just during elections. The Constitution must ensure that most decisions will be taken at a local or national level. Only then people may accept that for their nations to become part of the European Federation, there will be a partial loss of sovereignty to better protect them from external threats and become more united around the values that we all hold dear – the Universal Values of Humanity.

The only way to meet those conditions is to create **a shallow** European Federation, to which nations will pass only the necessary attributes of national sovereignty, which will be required for the Federation to function. These minimal functions that such a federation must have include:

1. President of the Federation as the head of the Government with two Vice Presidents directly elected by all citizens
2. Foreign Affairs to deal with external relations
3. Federal Army, while remaining a member of NATO, and thus reducing the overall cost of defence
4. Federal Budget and Eurocurrency, which will take the responsibility for repaying sovereign debts
5. European Parliament
6. Constitutional Court as a supreme judge over national constitutions' adherence to our values
7. European Court of Justice – as the supreme body to defend citizens'' rights

Whether such a European Federation is created using Scenario 1 or 2 is less important. What is important is that the Constitution, which in Scenario 1 may be created after the Federation has already been set up, must by debated and approved by the EU citizens and not the governments. If the Future of Europe Conference is used as an initial body facilitating the process of defining the European Constitution, then the proposal of using Citizens' Assemblies is the best way forward. In such case, the process of defining the Constitution should be granted to National Citizens' Assemblies, to which a few hundred citizens will be randomly elected, continuously advised by experts. The second level of final approval of the Constitution should be the European Citizens' Convention, to which each country's Citizens' Assembly will select their representatives.

Chapter 4
Global Wealth Transfer

The creation of the future European Federation (EF) could start the process of the world's federalization. But it can only happen if the world accepts what seems to be inevitable. 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 a significant redistribution of wealth.

I am fully aware of the complexities and almost impossibilities of delivering such a momentous change for humanity in the world which, for example, could not agree on stopping the genocide in Syria. The odds are heavily against such an optimistic view as I have presented here. On the other hand, should we be incapable of resolving most of these issues by around 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 would be the EF's assumed responsibility to manage that process, of which the main component must be a global wealth redistribution. Apart from purely ethical reasons, this 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 risk, leading to Humanity's extinction.

On 25 September 2015, the United Nations organisation 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:

Create the wealth redistribution programme, so that the donor countries (mainly the Northern Hemisphere) will over decades 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.

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 to redistribute the global wealth more evenly by creating the **Global Wealth Redistribution Fund (GWRF)**. 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 for most of its existence were giving the poorer countries the proverbial fish, whereas private foundations give them a fishing rod. Since 2000 that situation has improved at the UNDP, but the other crucial differences remain. These 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 wealth distribution. Therefore, it should be the future European Federation (EF) and later on the Human Federation (HF), which would head such a programme with additional injection of funds from other sources to finance the target GWRF projects, and if possible, co-ordinate wealth distribution.

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 set up as decades-long continuous effort. The most natural way for the EF would be to act via the membership of the EF Association Area (Zone 4). The programme itself would be the magnet for countries to join that zone. Looking at the current EU association agreements it is obvious that the EU is already aiming in this direction, as each such agreement includes the development of political, trade, social, cultural and security links. Currently, there are over 30 such agreements plus over 50 trade agreements.

Once the EF has been established, most of these agreements will be turned into the EF Association Area's agreements. If the rate of the new countries joining the EF Zone 4 would be at least as fast as it is now, the total number of countries in the Federation's four zones could reach about 140 by 2040. I include here, in the assumed EF Single Market Zone, the USA, India and Japan, which have been negotiating the Free Trade Agreement with the EU (USA withdrew from those negotiations in 2015), similar to the one signed in 2016 with Canada. That means

in 2040 the EF with 140 countries in all zones would represent perhaps 80% of the world's GDP and 70% of the global population.

I believe that such a large scale of wealth redistribution is the only realistic long-term solution for maintaining global peace and preparing Humanity for the new period of a planetary civilisation. Wealth distribution, if it is carried out on such scale, and if it follows the principles that I propose here, will achieve three objectives:

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

The final act in completing the creation of the Human Federation would be a total demilitarization of the world. I cover that in the next chapter.

Chapter 5
The World Without Superpowers

Ruling over the world – a Supremacist's dream

In the overall concept of avoiding humans' extinction, the role of the world's de facto leader, cannot be overemphasized. I believe that the best chances the world has, is for the future European Federation (EF) to take on that role. Gradually, by the process of more and more countries joining the EF, we should reach the stage when it becomes a Human Federation (HF). But there cannot be a Human Federation if the world's Superpowers do not lay down their weapons and the whole planet becomes demilitarized. This is not likely to happen soon, or it may even not happen at all, which would most likely mean that the human species will have a good chance to obliterate itself. However, knowing the risk and the ultimate danger that our civilization faces, should motivate most nations but not all, to achieve a demilitarized world. But how could the Superpowers willingly agree to lay down their weapons?

Many military experts agree that a global nuclear war is very unlikely today, as its aims could be achieved by different means. Instead, we can be almost certain that in this decade of Immature Superintelligence, there are bound to be several cyber wars, with the consequences far more unpleasant than the current Covid-19 pandemic. But I doubt (and yes, it is still a hope) that there can be an outright winner in such a cyber war. After all, the current Superpowers have similar capabilities in that area, like neutralizing a military equipment, using electromagnetic bombs exploded in space, switching off power stations for weeks, if not months, opening water dams etc. By staging such an undeclared war, the invading Superpower, would face an almost instantaneous counterattack, making the whole attempt to rule the world, futile. We already live in the era of the Second Cold War.

As mentioned earlier, there is a high probability that this decade (2020-30), which I call the period of Immature Superintelligence, might be the most dangerous in the human history. Stuart Russell, Nick Bostrom, and other AI scientists talk about losing control over AI capability, as it gradually matures into a Superintelligence. Most of such warnings about losing the control over AI, have concentrated so far on controlling individual, highly sophisticated robots, which can indeed inflict serious damage. However, their malicious action is far less dangerous for the human civilization, than an existential threat posed by a malicious AI system, which may have a full control over **all** AI agents and all humans. That's what is called Superintelligence – a global unified AI system with intelligence far exceeding that of all humans. That is also why we must gain an absolute global control over the most advanced AI agents latest by 2030.

Moreover, at least in principle, there could be more than one Superintelligence (an AI system) operating at the same time in the early stages of its development, i.e. in the period of Immature Superintelligence. Since such an advanced system can be used by its owner (controller) for its own, potentially malicious aims, e.g. controlling the world, it is almost obvious that we are entering a period where the Superpowers, or even some very rich individuals, may be tempted to overpower other such sophisticated AI systems, to gain control over the whole planet. How probable it is that we may have no more wars because the world will have been conquered by a superior version of Superintelligence that served its Supremacist master? Is there any chance that we can avoid such a scenario? Well, my answer is 'yes', although it may surprise you, how I have arrived at that conclusion.

AI Supremacist's Dilemma

Let me start by looking at how we could avoid a lasting damage to our civilisation in the next 10 years, or so, which may potentially be caused by creating a malicious Superintelligence. As mentioned earlier, there is a high probability that this decade (2020-30), which I call the period of Immature Superintelligence, might be the most dangerous in the human history. In that period, there may be many dangerous events caused by Superpowers and some other nations e.g. Russia's attempts to take control over Ukraine, Moldova or the Baltic states, skirmishes in the China Sea over the artificial islands, India-Pakistan war, Iran-Israel's potential nuclear war, and many other local wars. These risks, although not existential on their own, when combined with other risks, such as a catastrophic global warming, pandemics, and huge migrations due to drought, could become existential.

However, the real danger for Humanity may come from losing control over the AI as described earlier. That danger does not come just from the Immature Superintelligence creating havoc of its own accord, when accidentally released due to lack of proper control. It will be a consequence of advanced AI being used by a Superpower or even by a rich psychopath. I consider the level of that risk, perhaps not fully existential, as very high indeed.

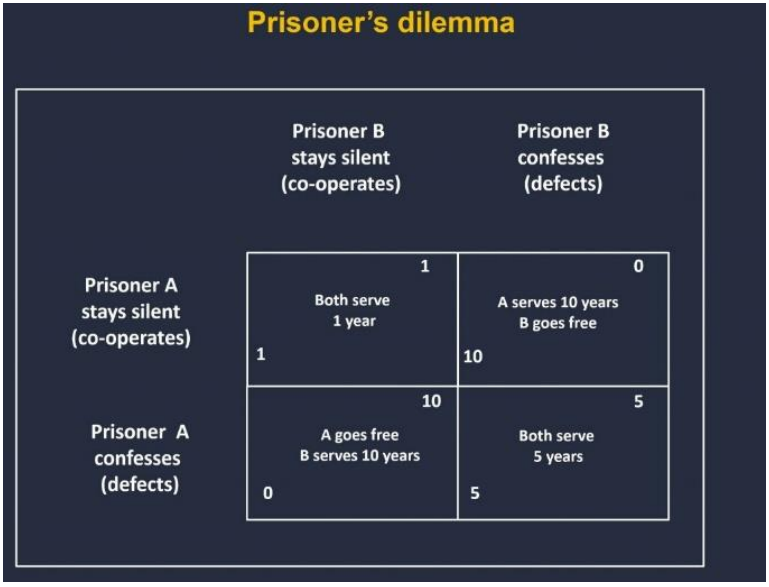
You may think that Superpowers are not yet capable of blackmailing the world with its own superior Superintelligence. I'm afraid this is a wishful thinking. Such a danger is almost here and the scale of the damage, which such an advanced AI system, which I frequently call here an Immature Superintelligence, can cause, will rise exponentially as we approach the end of this decade. Moreover, a Supremacist Superpower may be tempted to act sooner rather than later, since it will be more difficult to achieve a global supremacy in AI towards the end of this decade, when AI becomes more sophisticated and also dispersed to more global players, including very wealthy individuals.

There are two questions here:

1. Can a certain Superpower teach its Superintelligence to fight its rivals and deliver a supreme control of the whole world to its owner? I believe it can.
2. Can such a Superpower, after having conquered all its challengers and becoming an absolute ruler over all humans, also control its own, still Immature, Superintelligence. My answer to this question is – very unlikely.

A Superpower (let's call it **Supremacists**) will face a dilemma. It is the possibility that the control of an Immature Superintelligence by a single Superpower, which might allow it to rule over the entire planet, although possible, may create the final outcome much worse for the aggressor, as well as, unfortunately, for the whole Humanity. This is the dilemma that some Superpowers may be pondering on right now, which is a well-known problem in the game theory known as the **'prisoner's dilemma'.**

The prisoner's dilemma has its roots in game theory, mathematically best described by Albert W. Tucker and John Nash. It was originally developed for economics but has been deployed for even a more profound use in the geopolitical strategy, especially during the Cold War era. In the original concept of prisoner's dilemma, two prisoners, suspected of armed robbery, are taken into a custody. Police cannot prove they had guns; they only have stolen goods as evidence, for which they can be kept in prison for **1 year**. To get the evidence, they need one of them to confess that they indeed threatened someone with guns, which would keep them in prison for **10 years**. Therefore, they decide to offer the prisoners a reduction in prison sentence to **5 years** if they confess to having guns during the robbery. This is shown in the following diagram (the numbers represent years in jail):



I have developed below a variant of the prisoner's dilemma in relation to Superintelligence, which I call the **AI Supremacist's Dilemma**, with the same rules and assumptions. So, like in a typical prisoner's dilemma, the opposing parties choose to protect themselves at the expense of the other participant.

When applying the prisoner's dilemma to Superintelligence, I consider a scenario, in which there are two Superpowers: **Supremacists** and **Humanists** – representing the rest of the world. Let's say that the Supremacists create Superintelligence that is equal to that of Humanists'. The Supremacists' objective is to rule the world according to their own values and save their own species indefinitely. They plan to use Superintelligence to help them achieve that goal, while still remaining its master. To achieve that, they must upload its Superintelligence with certain goals (motivators) based on the very top of Supremacist's values. One such top value could be, for example, making the Supremacist's nation, race, or religion superior to other nations. If they decide to do that, they will violate the so-called Asimov's first law for robots – do no harm to humans, now largely superseded by the Asilomar principles.

The consequence of that might be that such a Superintelligence would indeed initially act in a malicious way in the interest of that Superpower only. However, at some stage it might turn against its master, since an evil Superintelligence will not be able to perfectly differentiate between a friend and a foe, or between evil and good. This is especially likely, since in this decade we will only have an Immature Superintelligence, which will be prone to some grave errors. In the end, if such a scenario turns out to become a reality, nobody will be able to control Superintelligence, which is most likely to be an evil entity.

Such an evil Superintelligence may very quickly decide to make us extinct for many of its own reasons. Will the Supremacists be prepared to take such a risk? Will they do it, knowing that there is a high probability their Superintelligence built on the key value of supremacy with its key goal being the subjugation of all other people who are not the Supremacist's citizens, may in time become evil, annihilating the Supremacist nation and the entire human species?

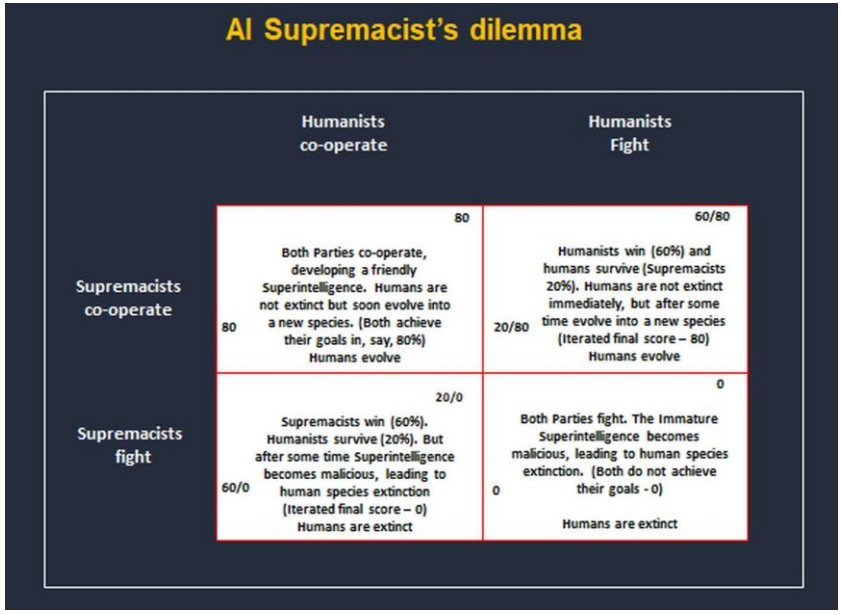
On the other hand, Supremacists may consider co-operating with the rest of the world (the Humanists), to mutually develop a friendly Superintelligence, which may be immensely beneficial to all. Instead of fighting each other, Supremacists and Humanists could jointly work with a mature Superintelligence, to evolve together over a longer period into a new, Posthuman species, which will inherit the values promoted by the future Human Federation. Therefore, such a Supremacist will face a dilemma that can be set as four possible scenarios:

- **Scenario 1**. Supremacists decide to cooperate with Humanists, after Humanists convince them to work together. Both Parties accept that one of their goals to survive in a biological form will not be met. Supremacists will not achieve their objective to rule the world according to their system

of values. Thus, each party does not achieve their objectives fully, but say in 80%. The result: 80, 80.

- **Scenario 2**: Supremacists fight their corner and win. They become the supreme world power, imposing their values over all humans. However, after some time, the Immature Superintelligence becomes malicious and all humans become extinct. Supremacists achieve their objective (60%), but in the end, they become extinct with the rest of humans (after iteration they fail to achieve their objective, hence 0%). Humanists have lost, but they survive for some time, until the Immature Superintelligence wipes them off as well, (20% of their objectives achieved, 0% after iteration). The overall result: 60/0, 20/0.
- **Scenario 3**: Supremacists fight and lose. However, they survive for some time in a biological form (20% of their objectives achieved). Humanists did not want the cyber war but have won. Although the evolution via merger with Superintelligence has been delayed, they achieve their objective, say at 60%. The result: 20/80, 60/80.
- **Scenario 4**: Finally, Supremacists fight but neither they, nor the Humanists win. During the fight the Immature Superintelligence became malicious, eliminating all humans. Neither of the Parties achieves their objectives. The result: 0, 0.

The scenarios are represented in the table below:



The numbers in the AI control dilemma are of course only exemplary, illustrating the point. The overall result is that both participants may find

themselves in a worse situation than if they had cooperated with each other in reaching a decision.

I am almost certain that most Superpowers already play that game trying to find a solution that would be significantly better for them, than the opposing power. However, after the world has experienced first-hand some severe consequences of Cyber-attacks for several years, it will become obvious to all players on the geopolitical stage that there could be no outright winner in an all-out War of Superintelligences. It should also become clear in any war games that no Superpower can realistically expect to gain supreme advantage over the rest of the world, by developing and controlling its 'own' Superintelligence, which might selectively destroy the Superpower's adversaries.

Additionally, in such context, any potential advantage gained in conventional or local nuclear wars would mean very little in the overall strategic position of a given Superpower. Moreover, any 'hot' global or local wars make no strategic sense. Instead, as has been mentioned before, the only hope for the whole world is to 'nurture' Superintelligence in accordance with the best values of Humanity.

However, neither the prisoner's dilemma, nor the AI Supremacist's dilemma would work with psychopaths. If some mad scientists, dictators, very wealthy individuals, or Transhumans (see further on) want to inflict damage on Humanity, even if they themselves perish, e.g. in the style of Stanley Kubrick's 'Dr Strangelove', then this scenario of AI Supremacist's' dilemma will not work. Such psychopaths may literally destroy Humanity. Therefore, as with conventional or nuclear wars (e.g. North Korea), the world may have to pre-empt such potential malicious action by destroying dangerous AI facilities, when it is still capable of doing so. This may be a lesser risk than letting psychopaths do severe damage to the world.

When the Superpowers realize within this decade, that there can be no winners, but Superintelligence in an all-out Cyber-War, I can offer you some additional dose of optimism in this quite a positive scenario. I believe, we can expect in the next 10-15 years some unimaginable breakthroughs in the planetary co-operation, for example:

- Stalemate in achieving global supremacy may lead to opening gambits, i.e. giving up previously held advantage as a quid pro quo. An example is the Intermediate-Range Nuclear Forces (INF) Treaty signed in 1987 between the Soviet Union and the USA, which was recently recalled by Russia and USA/NATO. It is quite likely this may be 'repaired' by a new treaty with better controls and even steeper reduction of nuclear arsenal.
- AI Superpowers will end Cyber Wars, and will focus instead on developing a single, friendly Superintelligence
- European Federation is very likely to be set up by the end of this decade by creating membership zones almost seamlessly and becoming the most

important organisation in the world. This could be the beginning of the future Human Federation.
- NATO may be fused with the EF by the end of the decade
- Russia may join one of the zones of the European Federation (EF) – the result of an economic decline in the post-Putin era. This may become a pivotal moment in the federalization of the world. It may paradoxically happen earlier than the USA joining the EF, as the 'nursery' of the Human Federation, although the order in which both countries would join the EF is less important
- Should the European Federation be not set up by the end of the decade, it might be NATO, which might take a de facto role of the World Government. This would require expanding its scope of activities by including the Cyber-war prevention, and covering economy, health, and infrastructure domains

I know that this looks like an almost idealistic scenario enabling democracy to be rebuilt and spread throughout the world much more easily and more effectively. However, I would rather take a more realistic approach and assume that the new democracy system will be born in pain and at the time of severe distress, or perhaps even apocalyptic danger. That may also stem from the rising capabilities of Immature Superintelligence.

People are often divided by 'Us' vs 'Them' perception. Perhaps such a threat from more and more capable Superintelligence could unite Humanity under 'Us' versus 'It' agenda, 'It' being the Superintelligence.

The world without wars

When the Superpowers realize, quite probably within this decade, that there can be no winners, but Superintelligence in an all-out Cyber-War, this might be the moment when all Superpowers' own weaponized AI, together with other conventional and nuclear weapons, would be disabled. This may ultimately lead the Superpowers to abandon their dreams of ruling the world. Instead, the growing potential danger coming from Superintelligence will finally get all nations together in a Human Federation.

If my reasoning is right, then I can offer you some additional dose of optimism in this quite a positive scenario. I believe, we can expect in the next 10-15 years some unimaginable breakthroughs in the planetary co-operation, for example:

- Stalemate in achieving global supremacy may lead to opening gambits, i.e. giving up previously held advantage in a quid pro quo deal. An example is the Intermediate-Range Nuclear Forces (INF) Treaty signed in 1987 between the Soviet Union and the USA, which was recently recalled by Russia and USA/NATO. It is quite likely this may be

'repaired' by a new treaty with better controls and even steeper reduction of nuclear arsenal

- AI Superpowers will end Cyber Wars, and will focus instead on jointly developing a single, friendly Superintelligence
- European Federation is quite likely to be set up in the next few years, in the midst of a chaos, by creating membership zones, and becoming the most important organisation in the world. This could be the beginning of the future Human Federation.
- NATO may be fused with the EF by the end of the decade
- Should the European Federation be not set up by the end of the decade, it might be NATO taking up the role of a de facto World Government, by expanding its scope of activities, starting with the Cyber-war prevention, and then covering economy, health and infrastructure domains, After all, it is already a confederation. If one includes Article 5, about the mutual military assistance when one of the members has been attacked, then such a limitation of sovereignty, makes it a Federation in the area of defence only.
- Russia may join one of the zones of the European Federation (EF) – the result of an economic decline in the post-Putin era. This may become a pivotal moment in the federalization of the world. It may paradoxically happen earlier than the USA joining the EF, as the 'nursery' of the Human Federation, although the order in which both countries would join the EF is less important
- UN establishes a majority voting system in the Security Council but that may be irrelevant, if Russia, (and possibly China, although it would probably join last) would have already joined the EF and NATO.

I know that this looks like an almost idealistic scenario enabling democracy to be rebuilt and spread throughout the world much more easily and more effectively. However, I would rather take a more realistic approach and assume that the new democracy system will be born in pain and during the time of severe distress, or perhaps even an apocalyptic danger. That may also stem from the rising capabilities of Immature Superintelligence. People are often divided by 'Us' vs 'Them' perception. Perhaps such a threat from more and more capable Superintelligence might unite Humanity under 'Us' vs 'It' agenda, 'It' being the Superintelligence.

Once all Superpowers have joined the Human Federation, all scientific discoveries and progress in the AI development would be fused to assist in creating a mature, friendly, singular Superintelligence. The age of human co-existence with a friendly Superintelligence will then begin.

6

PART 6
MANAGING HUMANS' EVOLUTION

Before you move on…

I hope that by now you have a reasonable overview of the following facts:

1. Our civilisation is at a turning point due to at least 9 existential risks, of which delivering a malicious Superintelligence is the most imminent and the most profound in consequences
2. Development of Superintelligence, by a continuous advancement in AI technology can no longer be stopped. Although in theory it could be done, practically, there is no way that we can unwind our memories and our knowledge, pretend that we do not know what we know, and agree to return to say 18th century technology.
3. I have already listed the three steps that must be carried out, ideally simultaneously and completed by 2030: Controlling the development of AI using three approaches; Agreeing the new system of Universal Values of Humanity; and selecting an existing, most experienced organization, which would become a de facto World Government. In practice this is impossible, in that timescale, so we must compromise to deliver whatever is possible.

There is a slim chance that against all the odds somehow, a bit battered, our civilisation will plough through the next two decades, the most dangerous period in the history of a human species and will deliver a friendly Superintelligence. The question is what next? Read Part 6 to get some plausible answers.

Chapter 1
Transition to a Human Federation

The era of Novacene

The unprecedented pace of exponential change may have either a positive or a negative impact on the long-term outcome for the human race. This largely depends on how we use the potential of such discoveries and innovation, like Artificial Intelligence. So, how can we make the safest possible transition to coexistence with the ultimate form of AI – Superintelligence.

I have emphasized the fundamental role that a democratic system of governance must play in the several decades long transition period. It is through a deep reform of democracy that we can reset the relationship between nations, based on Universal Values of Humanity, which best describe what it means to be a human being. These are the values, which should be inherited by Superintelligence. Therefore, democratic reforms must be anchored to some fundamental principles, such as those in the Consensual Presidential Democracy: Balanced Rights and Responsibilities, Political Consensus, Deep Decentralization and AI-assisted Governance.

But the reform of democracy must also lead to the creation of new institutions, which will ensure that at least man-made existential risks, such as the risk of developing a malicious Superintelligence, global warming, or a global disorder, are minimized. That is why the European Union has been identified as the most likely existing organization, which by transforming itself into the European Federation State (EF), could take the role of a de facto World Government and gradually evolve into the Human Federation (HF).

When selecting the best candidate for the future HF, there were three other rivals to EU, namely NATO, USA, and China. However, the analysis I have carried out, clearly points out that the EF is the most suited organisation for such a transformation into the HF. Therefore, I would not expect that at some stage EF will pass on the control to another organisation such as the UN. If my assumption is right, it will rather be the EF taking over more and more functions from the UN (and NATO), than the other way around. Given its current status, UN can only do what is possible within the existing legal and political constraints. The world needs an organisation such as the EF right now, fulfilling the tasks which UN will probably never be able to complete within the next few decades.

The HF will inherit from the EF its Constitution and the World Government. The HF Constitution will be amended as the needs dictate. The same relates to the former EF's institutions. Most of them will be gradually adapted to a larger world-wide function, rather than the European role only. However, they will be

some new institutions, which will most likely be set up based on the tasks that the EF will have already been carrying out although at a smaller scale.

As mentioned earlier, I assume that **we shall have a fully mature Superintelligence by 2050.** I have split the time between now and 2050 into three stages.

1. **Immature Superintelligence** - the first stage 2020–2030. It will start with the transformation of the EU into the EF in 2020'. I assume that by about 2030, the EF will already be an established federation by then
2. **Twilight of Anthropocene** - the second stage 2030-2050. This is the period of rapid expansion of the EF until it transforms almost seamlessly into a Human Federation
3. **The Novacene Era** – Beyond 2050. This is the time of humans' beginning their gradual evolution into Posthumans.

Such a transition of our civilization is of course only one of the possible routes of achieving the creation of the HF. I hope that by the start of stage 2, about 2030, we will have managed to eliminate most wars as a means of resolving conflicts and the world will behave like a planetary civilization, co-operating in an increasingly greater harmony. By about 2040, when the EF may be formally converted into the HF, the world will gradually get used to dealing with a new state, a new Superpower, which will gradually attract more and more nations as its members. The biggest task in this transformation process will be a significant EF's enlargement. But the procedure for accepting and moving new members through the zones of the EF until they merge with the EF state, will have been well tried-out by then.

It may seem inconceivable today that after 2040, within say, a further 5 to 10 years, about 100 new members may be joining the HF. But we must not forget that the world will continue to change at nearly an exponential pace. Additionally, there may be other reasons for such an accelerated, slightly chaotic, expansion of the HF. The world may already be experiencing in positive and negative ways the impact of a maturing Superintelligence. The humans' supremacy for making any decisions on this planet may be slowly waning. My further assumption is that between 2040 and 2050, all Superpowers will also become members of the HF. Those countries that decide not to join HF, e.g. on some cultural or religious grounds will play an insignificant role.

I have created five scenarios in my previous book 'Democracy for a Human Federation' on how our civilisation can develop in the next few decades. However, I am presenting here only an updated scenario 5, the most realistic one, to give you an insight into what life may look like around that time.

Chapter 2
Scenario for the World in 2040

Introduction

My intention has not been to create an unrealistic, overoptimistic assessment of how the European Union or the world in general might be able to cope with various risks and adversities in the future. My aim has rather been to present various, mostly tough, choices and possible solutions, so that the world has a better chance of fighting existential risks. This process may be painful and the path quite meandering at times, but ultimately it may be our best way forward not only to survive as a species but also have a great future.

You may wonder what is so specific about selecting 2040 as the date for my scenarios. My assumption stems from the following reason. If the EU wants to survive it must become a federated state by 2030 at the latest. Furthermore, my personal feeling is that if the world survives another 20 years, approximately one generation from now, without any major man-made existential risk materializing, our chances of having a great future will be immensely improved. A possible great future is my preferred scenario, although even in this scenario, Humanity gets through several 'near misses', almost triggering off an existential risk.

The scenario presents the Human Federation (HF) after the transition from EF, around 2040, as the best Welfare State, the world has ever seen. It will be possible because of the HF's financial and material capabilities in 2040, thanks to phenomenal technological progress that would have been achieved by then.

I have already presented plenty of arguments regarding what needs to be done to minimize existential risks. However, for completeness, let me summarize it here.

The EU will have to become a federal state because of the economic and social crises within the Eurozone, political pressure wielded by Russia on the Eastern, Central and South European countries, as well as the migration problems, which by 2030 may come to the fore in earnest. Additionally, there is of course an ever-increasing risk that some of the existential risks facing the EU, and by extension the whole Humanity, may materialize at the time (as the Coronavirus has come completely unexpected), when no single country or organization would be able to co-ordinate the rescue action. Therefore, 2030 may be the last year, when the EU federalization could be completed with minimum chaos. After that, it could still be possible but not that likely, and if carried out, it may happen in total chaos and under the pressure coming from some of the existential risks.

The EU evolves broadly in five-year cycles around the term of the EU Parliament, the elections of the new EU Commission and the European Council President. The last EU parliamentary elections were in 2019. The next opportunity regarding substantial EU reforms can happen by about 2024. Although in theory it is feasible that part of the EU could be federated by then, using for example the article 20 of the current Lisbon Treaty, in practical terms it may be difficult to achieve. Therefore, the most realistic way to proceed might be to start the work within the promised Future of Europe Conference in the autumn of 2020, which may initiate a Constitutional Convention on the future EU constitution by 2024, and set the date for the formation of the European Federation for 2026 - 2030 – the symbolic date that I have assumed. However, it would have been too early to design the scenario for the EF say, in 2035, because the first five to ten years of the EF will be very chaotic indeed. Therefore, this scenario paints the vision of the EF transformed into HF in 2040.

There are several ways of designing scenarios of the future world. For example, there is an excellent document issued by the European Commission: "Global Europe in 2050" (European Commission, 2012). It is 10 years ahead of my scenario, but contains some interesting conclusions, quite at odds sometimes with what this book is about. In that document the EU Commission Research team built their scenarios (164 pages long) according to a common format that deals in sequence with 6 main dimensions of the future:

1. Global demographic and societal challenges
2. Energy and natural resource security and efficiency, environment, and climate change
3. Economy and technology prospects
4. Geopolitics and governance: EU frontiers, integration, and role on the global scale
5. Territorial and mobility dynamics
6. Research, education, and innovation.

They have produced 3 scenarios, of which summaries of which I will quote alongside my scenarios, to give you some comparison. However, as you will see, their scenarios have different objectives and selection criteria from mine (i.e. what kind of organization we need to save Humanity from Superintelligence). Therefore, rather than use their approach I had to use a different, less detailed one, focused on my objectives.

My Scenario contains variables that are probable or possible but excludes plausible (conceivable) futures because of their very low probability level. The further we go into the future, the less we know, especially, as the world has already started to change at almost an exponential rate. Even the 2040 date is a bit too far in the future, about one generation, but for proposing certain solutions, it should serve our purpose fairly well.

Before going any further, I need to make one caveat. **Please do not expect any in depth justification for the numbers used or choices made in the scenarios**, although in most cases I have argued those choices to some degree earlier in this book. As in many other places, I had to make shortcuts and sometimes barely signposting you to the direction of travel. That was the only way, in which I could present a fairly complete picture.

When I am presenting my own comments, I shall use an italics font.

This is the 'Preferred future' of the five-scenario model that I have fully developed in my first book 'Who could save Humanity from Superintelligence'. Some people may call it a utopian scenario. Then perhaps I should quote Roger Scruton here, the British political philosopher, who said: "Utopia is a kind of a scenario planning with the assumption of a positive result". Muhammed Yunus, the winner of the Nobel Prize in economics, has a very succinct vision of the future in 2050. He calls it "A world of three zeros: zero poverty, zero unemployment and zero emissions". How probable is the achievement of some of these goals by 2040? This depends on the assumptions taken.

Key assumptions for this Scenario

This is the most positive of all five scenarios, although it is highly unlikely there will be no major stumbling blocks on the way. In any case, it should help us visualize much better the future Human Federation (HF) in 2040. It will also ask some questions that need to be answered to make this scenario more probable.

My key assumption is that the EU leadership will have managed to convince the electorate in its member states of the necessity of making a painful transition into the European Federation (EF) with its new Constitution. This means, that the world would still go around and none of those existential risks I wrote about in Part 1, would materialize. This scenario complements the proposal I have put forward on how the EU could make a transition into the EF, showing what the EF and the world might look like in the future.

I have tried to calculate the numbers in the scenario to be as close as possible to what they might be in 2040 but of course in many instances this may be somewhat off. The important point is to present how various HF Institutions and processes might work when the HF becomes operational. When I refer to data or situations before 2020, which serve as a reference, this means these were real events and real data, quite often supported by citations.

So, let me now focus on largely positive outcomes of the EU's decision to become a federated state. The benefits of the EF, as might be seen from the perspective of an average EF citizen in 2040, are spread across several areas. Some of these benefits are rarely talked about today. The benefit that would probably be the most appreciated, after the most dangerous period that the world would have

gone through, will be simply the benefit of living in peace. That does not mean that in 2040 existential dangers would be over. That can never happen. Life at a species level is simply a continuous exposure to risk, one of which could be the end of life of the entire species.

The Government of the Human Federation

The Human Federation has been born! How quickly have people got used to what seemed an impossible dream barely 10 years ago. Therefore, the 10th anniversary of the European Federation is being celebrated with incredible pomp for the last few months. When back on 1 January 2030, 39 countries were united in one State called the European Federation (EF), there were too many sceptics to count, who prophesized a sudden, perhaps a traumatic, end of the Federation. That was supposed to be the result of external political pressures (Russia and China), economic (even more serious financial crisis than in 2008), and the internal pressures (wish to return by many EF citizens to the world that was familiar and now was gone forever).

EF, although initially regarded as just another large international organization, created for all countries of the former European Union, was adding many countries from other parts of the world to the EF's four zones. Some said it might soon become a Human Federation, but they did not want to change the name yet, in order not to antagonize Russia or China. Nevertheless, the name was finally changed to a Human Federation with all four subsidiary zones now having 138 countries. Its members constitute more than 60% of the global population and 70% of the world's GDP. HF and all members of the subsidiary zones are still members of the United Nations, which has very limited real powers since the Security Council is totally dysfunctional.

The Human Federation state includes all the previous members of the former European Union. They are now called nations, or regions, rather than states. Altogether, there are 48 countries in the HF. When combined with 138 countries in subsidiary zones, the total population of HF is now nearly 6 billion. The United Kingdom joined the EF in 2032. Yes, it took a while, but Britain is now a different country, with its own new Constitution and the former King William becoming the Life President of the National Heritage (a kind of Ministry of Culture combined with the British National Trust). Following the new British constitution, from 1st January 2030, Wales and Scotland are directly members of the HF. Northern Ireland was merged with the Republic of Ireland in 2024.

All HF members' constitutions have been changed and replaced by new constitutions, which allow for large regional separations, enabling them to join the HF directly, if they wish. Over the last 10 years, it has gradually led to some original member states splitting into large regions, each with at least 5m citizens, according with the EF Constitution. For example, German lands, Bavaria, and Saxony, are now directly regions of the EF rather than Germany. Belgium was

split into two large regions: Flanders and Wallonia and each of them has also additionally merged with one former Dutch region and a former French region. There are also two other cross-country regions. The first is Catalonia, which is now much bigger than before, by being joined with the previously French Catalogne Nord. The second one is the Basque Country, twice as big as the previous Spanish Region, which was merged with the previously French Northern Basque Country. These cross-country regional mergers follow a model set up in 1996, of the first Euro-region Tyrol-South and Tyrol-Trentino. That was formed between the Austrian state of Tyrol and the Italian provinces of South Tyrol and Trentino.

HF, in line with its Constitution, is a mixture of a representational and a direct democracy, with a two-chamber parliament. The representatives to the Lower House, the Nations' Chamber are elected in a two-stage system. The first stage is a simple First Past the Post system. The second one is a preferential system based on Alternative Voting System. The Upper House, the Citizens' Chamber (the Senate) has senators selected using an enhanced sortition (random selection) system.

The current President of the Human Federation is a Frenchmen, Maurice Cheval. He has two 'shadow' Vice-presidents, a Hungarian and a Canadian, who make most decisions through consensus during the Presidency meetings (two votes needed to pass a motion). On most significant matters, such as defence, or declaring the state of emergency, the President makes decisions alone.

The current Prime Minister, Leopoldo Gonzalez, is Spanish. He is a member of the Democratic Liberal Party of the HF, the strongest party in the HF Parliament. His key ministers are all members of the HF Parliament: The Minister of Defence is British (a permanent position granted to Britain for 10 years, as a sweetener to re-join the EF). The Foreign Affairs Minister is Dutch, the Home Affairs Minister is an Australian, and the Minister of Finance is a Japanese. Other ministers come from a pool of 2000 experts, selected by sortition from all HF countries.

There are 5 members in the Human Federation Convergence Area, which is in Zone 1: Brazil, Tunisia, Singapore, New Zealand, and India. The members have signed up to the constitution of the HF, but certain articles of that constitution do not apply to them. Every country in this zone has MPs in the HF Parliament. These member countries will join the Human Federation latest within the next 5 years.

There are significant changes in the Human Federation Single Market area, which is in Zone 2. It has now **30 members**, including Turkey, Morocco, Egypt, Libya, Armenia, Lebanon, and Thailand. The most prominent member is the United States, which joined this year. All member states in this zone have up to five opt-outs of the HF Single Market policies, which suit their particular

circumstances, and can stay in this zone for as long as they want. They are bound by the articles of the HFSM Treaty and each country has representatives in the HFSM Assembly. The members can join the HF, by moving first into Zone 1, once they meet certain economic, social, and political criteria.

There are 22 countries in the Human Federation Customs Union, which is in Zone 3 such as Belorussia and Paraguay.

There are 81 countries in the Human Federation Association Area, which is in Zone 4. Most of these countries come from Africa and South America, such as Kenya, Nigeria, Venezuela, or Argentina. Some of these countries, such as South Africa, had individual Association Agreements with the former European Union before the federalization. The member states in this zone are not bound by any articles of the Human Federation Constitution but must fulfil the terms of the Treaty of the Human Federation Association membership. Additionally, each of the countries has individual association agreements with the Human Federation. If they fulfil the required criteria, they can move up to Zone 3.

The official language of the HF is English and there are no translations in the HF Parliament. Across the whole HF and its subsidiary zones, English is a mandatory language in official communications and is taught at all HF schools. However, at a member nation, or at a regional level, the official language is whichever language the region chooses, with English being a mandatory second language. Therefore, all signposting, street names etc., are in two languages. On the other hand, language as such is not a problem anymore, as almost all people have Multilingual Translators embedded either in their glasses, aural devices, watches, or chip implants under the skin, which enable simultaneous translations.

People within the HF State have the same rights across the entire HF area. After all, HF is now a single state. This includes benefits, recognition of all qualifications, national health entitlements and pension rights. However, there are regional differences in education, public holidays, regional legal system, (any new laws passed must be compatible with the HF Constitution, adjudicated by the HF Constitutional Court), urban and architectural design, culture and regional heritage (as long as the EF general rules are observed).

For comparison I enclose a summary of the closest scenario produced by the European Commission in their document "Global Europe in 2050" called 'EU Renaissance: further European integration' (European Commission, 2012).

"In this EU Renaissance scenario global security is achieved, with the generalized enforcement of human rights and the rule of law. The world undergoes a global democratization of power also because of increasingly active non-state actors, global public policy networks and the media. The EU is enlarged both east and southwards, and political, fiscal, and military integration

is consolidated. There is strong public support toward challenging targets e.g. in climate change and energy efficiency. The all-continental integration of energy systems (with renovation and heavy re-investments) boosts the share of renewable energy. Innovation systems undergo major reforms to become increasingly systemic, with more user-integration, more easy-to-use technological systems and services, and more encompassing smart growth-oriented technology and innovation policies. Importantly, the EU manages to optimally design its technological and research policies, to target the right domains and methods, and this leads to an acceleration in the pace of innovation and the productivity gains increase progressively until 2050 within the EU, compared to the 'Nobody cares scenario', the rest of the world keeping its own pace."

Geopolitics

Foreign affairs

Human Federation (**HF) has now become the most significant state on the international stage**, especially after the USA has become a member of the HF Single Market Area (Zone 2). It can exert direct significant political pressure on any of the 138 countries, members of 4 subsidiary zones, at least by controlling the flow of financial support for member states, especially in Zone 4 (African, Asian and some South American countries). The result is that there are no military conflicts among any of those countries. Unfortunately, the influence of the HF on the remaining countries is very limited, since most of these countries are vassal states of either Russia or China.

HF has now three seats on the United Nations Security Council (previously occupied by France, Britain, and the USA), although UN stopped playing any credible important role in maintaining the world peace. This is slowly becoming the domain of the HF, although it is too early to say, how successful it will be.

Former G7 countries are now G10. Military pressures from Russia and China, bordering on threatening the use of weapons of mass destruction were the main cause for enlarging G7 and making their resolutions more meaningful. The first such threat happened when the Russian President Vladimir Putin said in March 2018 that Russia would not care for the world, if his country were to perish – it would launch an all-out war. That led the G7 countries to invite new large democratic countries: India, Brazil, and Nigeria to become members, which happened in 2028 and immediately sparked off a number of dangerous military incidents against some of the G10 members. The other reason was a further decline of the role of the UN, which was being almost entirely run according to Russia's and China's wishes, which led to the USA, France and the UK frequently boycotting the Security Council meetings.

The Human Federation Army – A new relationship with NATO

Anyone who thought over 20 years ago that the former EU did not need its own army, because it would be a superfluous or excessive risk mitigation strategy, should have watched "Occupied", the most expensive Norwegian television show in history, screened in 2015. It had seriously enraged Russia, because it showed the subversive way, in which Russia forced Norway to surrender its sovereignty. When Russians came to Norway, there were no tanks or fighter jets, or "little green men." The diminution of Norwegian sovereignty and the assertion of Russian control were much more subtle and visible only to those who cared to notice. On the surface, life remained normal for most Norwegians, who went about their daily business as though nothing had changed. As with Finland after the WWII, Russia applied to Norway the same process of ***Finlandization****, a pejorative term describing the situation, when a small country accepted a reduction of its sovereignty in exchange for a limited self-rule.*

Well, that was the film. But interestingly it was very close to reality that evolved very quickly. In the early 2020', Russian aggressive actions took place in the Ukraine and Moldova, including some serious incidents in the Baltic States *(see below).* At that time EU was barely thinking about having its own army. The only element of a potential future army was the EU's Permanent Structured Cooperation (PESCO) set up in 2017. The original intention was to enable the EU member states working more closely together in security and defence. That permanent framework for defence cooperation was to allow willing and able member states to develop jointly their defence capabilities, invest in shared projects, and enhance the operational readiness and contribution of their armed forces. At that time, EU was working closely with NATO, based on several agreements, such as the NATO-EU Warsaw Declaration singed in July 2016. That included 42 concrete actions, such as re-enforcing the NATO eastern frontiers with tens of thousands of NATO troops moved semi-permanently closer to the Russian border.

At the time of incidents in Moldova and in the Baltic States, the USA was very enigmatic on invoking article 5 of the NATO declaration on mutual self-defence. That finally forced the EU to amend the NATO declaration, where all the EU countries became a single member of NATO. That in turn led by default to the creation of the EU Army, which is now far more effective than ever. From today's perspective it is clear that it was the formation of the EU Army that has been the best sign of the EU's resolve to dampen Russian aggressive attempts.

So, HF has now its own army, which has just celebrated the fifteenth anniversary of its formation. All previous member states' armies have been dissolved and re-grouped into HF Regional Defence Forces that spread across former state borders. All military equipment and standards are unified within the whole HF Army, which is a member of NATO, as a single country. However, the HF Army participates in UN Peacekeeping operations independently of NATO. The

official language of the Army, as in the whole HF is English, although national languages can also be used inside the regional bases.

Britain remained for some of that period, an individual member of NATO. After re-joining the EF, its forces have been merged within the EF army. But the condition the UK had made was to run the Ministry of Defence. Therefore, the current HF Defence Minister is Anthony Clarke, from the UK. British army is now part of the Human Federation. Being a nuclear power (which now plays a less significant role), the UK has a long-term Agreement with the HF Army, as the 'British Region'. Its entire defence budget is covered by the HF budget, from the annual payments made by the British Government to the HF that cover among others education, security, and defence. The only area that is strictly under the British control is its nuclear arsenal and small conventional weapons, contingent with the service of the British operations in its Overseas Territories, such as the Falklands. Any eventual use of the exclusive British nuclear arsenal is strictly under the British control.

The HF Army Chief of Staff must be of a different nationality than the current HF Defence Minister. Since France is also a nuclear power, it has special rights within the HF Defence system. It controls the French nuclear arsenal (this is the area still unresolved – who is to control the entire HF nuclear arsenal). There is a compulsory one-year residential service in the Army for men and women starting at 18 up to 25. That can be exchanged for 18 months of residential social service.

Nearly triggered existential risks

I have tried to increase the probability of this scenario by reviewing some of the forecasts by well-known strategists. For example, my own view is that 2024-2026 will probably be one of the more dangerous periods in the global politics in the coming decade, is supported by private intelligence firm Strategic Forecasting. In 2015 they published their Decade Forecast in which they said the world in 2025 would be significantly more fractured, dangerous and chaotic place, with Russia projected to collapse, US power in decline, and China's rapid progress stagnated (World Economic Forum, 2015).

Until about 2020, the only existential risks mentioned in the media was climate change and nuclear war. Of course, it is understandable that the media had no interest in conveying gloomy messages and neither was it in the interest of any type of business to project pessimistic views. It would have badly affected the sales. The governments pretended that existential risks were not an important enough issue for extensive political debates. Even the term 'existential risks' was only talked about in specialist TV programs or in the scientific press. Discussions in the parliaments on the subject were almost non-existent, apart from the Scandinavian countries, which have always been an exception in the world as far as an open communication with their societies was concerned. To

think that any party would put existential risks as an issue in its manifesto would seem utterly ridiculous. How would an average voter on the doorstep react to it? Even if those existential risks were true, the parties and the governments would say they could do nothing about it. When one looks back at how the world approached existential risks before 2030, the year of the EF creation, then one must really wonder how we managed to survive.

The lack of perception of existential risks was far worse than during the cold war era, when people were constantly being warned about the danger of one existential risk – the global nuclear war that might lead to the end of civilisation. Perhaps it was a much simpler message to convey since it could be imagined much more clearly. It is true that the new existential risks that the planet Earth has been facing for the last 50 years are far more complex and difficult to imagine. That's why people quickly lose interest in the subject, when the danger of nanotechnology or artificially created, incurable viruses, are discussed.

It was even less apparent that relatively minor risks could combine, and their cumulative effect might become existential. All Global Disorder risks fall into that category and they were those risks that nearly turned into an existential one. It all started with a series of unrelated, relatively minor incidents that were spread over several years, such as the Coronavirus that affected the whole world in 2020. For the first time since the end of the cold war, the world has realised how unprepared it was to fight non-existential events, such as pandemics, especially if they triggered off serious economic and financial crises. That was the canvass for what followed on.

In March 2018, a former Russian agent was attacked in Britain with a very sophisticated Novichok nerve gas, seriously affecting dozens of other people. Over 200 military personnel were involved in the cleaning operations. It was obvious that Russia was the culprit and only later on, when similar attacks occurred in other countries, it became clear that it was a test to see the resilience of the emergency forces, how quickly panic could spread out and how the attacked countries would react. Russia's apparent assumption was that if no severe consequences would fall on them, then it could raise the bar higher, not necessarily in the same area. Such an opportunity availed itself a few years later.

It was in winter 2024, which was exceptionally severe (climate change was then clearly noticeable). The whole Europe was covered in deep snow for many weeks. In February 2024, Russia took over Moldova in a clandestine coup d'état. NATO did nothing. Then shortly after that there was a large-scale Chinese cyberattack on Indian power stations (of course never admitted by China), which crippled India for several weeks. That was a clear retaliation for the Indian expansion (so perceived by China) into the Indian Ocean, when India started building artificial islands, similarly as China had been doing for over a decade in the Pacific Ocean. Within days of that incident, there was one of the largest, long-overdue, earthquakes in California (San Andreas Fault) that engaged vast

American resources. At this very time, an American psychopath biological scientist spread a deadly artificially produced virus at several airports around the world that led to massive wave of flue type epidemics, worse than Coronavirus, affecting millions of people world-wide, but in particular in Europe. However, Russia was least affected because of tighter border control.

The emergency services were stretched to the limit, in most parts of the world. In the USA the rescue services were incapable of coping with the aftereffects of the disaster and several army divisions had to be re-allocated to help local emergency services. In Europe, the arctic winter and flu-type epidemics completely overstretched the emergency, medical and food distribution services, creating chaos and local disturbances in many countries, where people were fighting for food, places in hospitals, or medicines, of which hospitals and pharmacies run out almost completely.

In such a situation, seeing that NATO did nothing significant to force Russian forces out of Moldova, Russia decided to invade the Ukraine first and when there was still no reaction from NATO, the Russian forces entered the Baltic States. NATO responded initially with air attacks. When the Baltics were almost overwhelmed by the Russian forces, a Russian tanker filled with a nerve gas (that should have never been there – a tactical error) was hit by a stray bullet and caused the release of the gas in the air. Within 24 hours several neighbouring countries were affected with tens of thousands of civilians' dead. The full-scale war was hanging dangerously in the air and Russians were clearly winning. At that point, American NATO forces fired a small nuclear weapon on the Russian troops near St. Petersburg that was not intercepted by the Russian anti-aircraft forces because the Americans first jammed the whole region with a magnetic bomb, disabling all computers. That immediately halted the conflict. Russia apologised for 'unintended' explosion of the nerve gas and withdrew the troops from the Baltics, the Ukraine and Moldova. That's how close the world was from a global nuclear, chemical, and biological war being fought at the same time. That's how combinatorial risks, if triggered in full, could have become an existential risk ending our civilisation as we know it. The world sighed with relief.

Russia and China have been the main challengers of the HF since its inception. Looking back at 2024 from today's perspective those were the most dramatic years in the last two decades. That was also the year of the EU parliamentary elections that was held in the aftermath of the conflict with Russia. The Baltic States received a massive material help from the EU but also from the USA. The conflict with Russia was the main trigger for the federalization of the European Union, which took a little bit more than 5 years. Although China and Russia cooled down their antagonistic stance towards the EU and the NATO countries immediately after the conflict, there were several other incidents between the major powers. For most of the last 15 years, EF and the USA lived

with China and Russia in the period of what became known as the Second Cold War.

Shortly after that, the 'Immature Superintelligence' became the source of several serious incidents. These included disabling by error in 2033 the entire power supply in the USA for three weeks, with hundreds of thousands of people dying of hypothermia and hunger (it was in the middle of one of the most severe winters the USA has ever known). The same extremely frosty winter lasting nearly 5 months affected Russia, where several million people died of frost and where even the stretched USA's and Japanese services were providing essential help in the eastern Siberia. Each of the Superpowers had its own Immature Superintelligence, a very advanced and capable AI agent, much more intelligent in most areas than humans (but not in all areas yet). It was frequently used for ill purposes by each of the superpowers (of course none admitted its use). But the biggest danger came from clandestine inventions done by very rich individuals, some of whom can be considered psychopaths. Even China and Russia had such problems in their own countries.

Gradually, the perception of common dangers and adversities stemming from Superintelligence and other existential risks facing the whole Humanity lowered the level of enmity between the Superpowers and became the biggest motivator for a true global co-operation. Additionally, China, USA and India signed an important agreement in 2034 on a joint creation of new artificial islands, where all costs and returns were shared by all parties, proportionally to their investment. That has just by chance created a model for similar agreements in other areas, which for the last few years has also been adopted by Russia.

Its seems that the latest version of the nearly mature Superintelligence created independently by the scientists in the USA, China, Russia and the HF may have finally convinced these countries, alongside some other military powers (India, Pakistan, Indonesia and Brazil) that very soon there could be no winner if Humanity as a whole does not come together. The danger has been that Superintelligence may become a super intelligently unfriendly agent, behaving very erratically with a possibility of taking suddenly a full control over the entire civilization. That has finally led the USA and India to join the HF.

Environment and climate change

Environmental disasters, as predicted have been very severe, especially around 2030. Some of them have already been mentioned. The CO^2 levels increased much faster than before (a kind of a run-away scenario). Luckily, it is now under control, mainly thanks to geo-engineering technologies developed by AI agents, costing huge amount of money. However, it is already clear that this is working since CO^2 levels have been dropping in absolute terms for the last 6 years.

Green energy is now everywhere thanks to key discoveries in 2030', such as the first commercial nuclear fusion power station opened in Pennsylvania in 2031. The second one was the discovery of how to use magnetic energy directly, available everywhere, as a more convenient energy than electricity. The third one has been the massive use of new materials to produce solar panels (such as perovskite) since 2024, which doubled the yield from a typical solar panel. Thanks to those discoveries, energy is now very cheap, which was also possible by discoveries of new energy storage media that last 7 times longer per 1 unit than in 2018.

Economy and finance

The Human Federation (HF) aligns its internal fiscal policies. Following the constitutional arrangements, a former member state of the EF, or any new state joining the HF, contributes 20% of its budget to the central HF budget, which is managed by the HF Finance Minister. That follows the original EF law for former EU member states on joining the EF, when they initially contributed only 10% of the budget in the first year and another 2.5% over the next 4 years, until 20% of their budget was fully managed by the EF Ministry of Finance.

There is an independent HF Central Bank (HFCB), which has existed since 1998. But today it serves the whole HF because there is no longer the Eurozone. The HF currency is the Euro, as before, worth now 3 US dollars. The US currency finally gave in to structural faults, on which its economy was based. Capitalism could no longer spread out globally. Business has become much more regulated and restricted. Goods can now be exchanged free of customs duty in almost all countries in the world, There are also far more effective controls put on large global corporations, which were threatening most of the states by setting the economic (and quite often political) conditions that were beneficial for corporations but undermined whole national economies. HFCB is responsible among others for setting up the interest rate, implementing the monetary policy of the HF, taking care of the foreign reserves, and overseeing the HF Banking Union. There is also the HF Monetary Fund, which is essentially the Bank of last resort, although its role is waning since the whole HF economy is almost perfectly harmonized by near mature Superintelligence.

The HF has now the largest GDP in the World (when counting all four Zones and including the USA) that amounts to about 70% of the world' GDP. **The Stability and Growth Pact** set at that time of the Euro currency crisis in 2012, has now become the centrepiece policy in all HF Zones. This sets the rules designed to ensure that HF itself and the countries in the HF Convergence Zone and the HF Single Market Zone must pursue sound public finances and coordinate their fiscal policies. As before, HF members cannot exceed their national budget deficit by more than 3% and their national debt cannot exceed more than 60%. There are still penalties for exceeding these thresholds. The same rules apply for the countries in the HF Customs Union and HF Association

Area, if they receive funding within the Global Wealth Redistribution Fund, otherwise, the funding is cut down.

Social Cohesion Fund is the continuation of the same fund existing during the European Union days**.** It is many times bigger than before, because much more money is transferred to the HF budget from the regions. The objective of this fund is to invest in poorer areas of the HF to help reduce regional economic imbalance. The richer regions pay trillions of Euros each year to improve economic and social conditions in poorer regions. This is now working much better than before, because help can be directly allocated to smaller regions, reducing the imbalance faster and more fairly.

HF Investment Fund. This fund provides investment mainly to smaller companies within the HF and all subsidiary zones. This is already the biggest such fund in the world.

All corporation and income taxes are collected directly by the HF, while VAT is collected at a regional level. 35% of the HF budget, which is now approaching €600 trillion, is distributed directly to HF regions, mainly through the HF projects and social cohesion programme. The rest of the budget is to finance central HF functions, such as defence, security, home affairs, the HF's Welfare State, and health service.

There is one flat corporation tax at 70%. It may seem high but it is the consequence of the decision made in 2032 when the EF decided that there would be no taxes on robots, which were introduced under some pressure from the trade unions in some of the former EU countries. This has worked very well, since most companies are now 'employing' robots or AI agents, so such a policy does not stifle innovation. There are now more sophisticated robots and AI Agents (many of them in a human form) than people on the planet – close to 9 billion. On average there are about 10% of human employees in manufacturing companies in the HF. In distribution and transport companies there are less than 5% of human employees. The biggest number of employees is in the service sectors, such as elderly care, medical care, fashion, leisure, and entertainment.

Education in Human Federation

Education at all levels has changed dramatically. In primary schools, traditional education has been almost completely replaced by the AI Assistants (one per classroom of 8). They perform the role of the previous teachers, but they have in depth knowledge of every child's progress and each child has an individual educational programme. The human teachers are still there but their main role is to teach children core human values and how they should be applied in life.

At secondary schools, almost the entire teaching programme is run by AI Assistants (1 per 4 students). Most secondary school pupils have brain implants, which have increased their brain capabilities exponentially. They communicate most of their progress wirelessly to their Assistants and each of them has an individual teaching programme. However, they still have 'classic' lessons with human teachers at least once a week to teach them human values and enable them to meet their pals. They also have a sport day once a week. Those very few who don't have the brain implants, have an individual programme using a combined teacher-AI Assistant teaching method. Once a student passes an exam from any subject at least at 80% score level, he can then move to the next level. History and social subjects, like psychology, are in the main taught by human teachers, with Ai Assistants checking the knowledge and assisting with any problems.

There is a minimum of two hours of history lessons weekly. Additionally, there are two hours a week of Human Federation Studies, of which 1 hour is dedicated to fighting fake media, populism, and xenophobia, by discussing current events in the HF. There is a strong emphasis put on bringing up young people in the human values promoted by the HF.

All university studies are free. Students are assigned their own AI teachers, which teach a subject depending on the student's individual capabilities and aspirations. There are no formal exams, since the certificate on passing a subject depends on the entire work carried out by the students and dozens of ad hoc tests done under the supervision of AI examiners. Once a student has achieved the required level of knowledge for a given subject, he is given the final degree from a university.

The Erasmus programme is now nearly 50 years old with more than 400 million people having completed either full or part of their studies in another national region of the HF state. In the last few years, it has been extended significantly, particularly to the member nations in all HF subsidiary zones. The HF Erasmus Programme has set up its own Erasmus Universities funded by the HF in many African and Asian countries. The programme will by rolled out to South America in the next few years. Within the HF itself all state-funded universities follow a special set of studies called the Erasmus Programme, dedicated mainly to re-enforcing HF values, HF culture and the history of the world.

The future of Work

The effect of Technological Unemployment was initially very severe. Yes, over 160 new skills were created by 2030 as a well-known futurist, Thomas Frey, predicted in 2016. However, within two years of the first signs of the coming Technological Unemployment wave, it became obvious that there were far fewer new jobs, than the jobs lost and those jobs that were available, required rare

skills. Millions of people became unemployed, resulting from continuous expansion of robotization and AI in general. The Technological Unemployment wave started a few years earlier than had been predicted because of the effects of the Covid-19 pandemic of 2020-21. The unemployment shot up to over 50% in some countries.

The lack of preparation in the old European Union countries for that entirely new type of unemployment was obvious. That has sparked off serious social unrest in most countries. It was a very bumpy ride indeed. Luckily, the unrest was quite quickly pacified by the introduction in almost all EU countries of either an unconditional Universal Basic Income or a Negative Income Tax. It has initially dented the budgets of some countries, but that was no longer a big issue after the federalization.

For the last 5 years before the EF was converted into the HF, **it has been running a 0% unemployment programme,** as envisaged by Muhammed Yunus, who called to create "A world of three zeros: zero poverty, zero unemployment and zero emissions". The HF has now achieved all three of them. Zero unemployment was achieved by these means:

- EF-wide job-sharing programme has been introduced so that most jobs are shared by 2-3 people
- The working week has been reduced to 15 hours
- The flexible retirement age now starts at 45
- The introduction of the **unconditional** Universal Basic Income at 30% of average earnings, in the mid-2020'
- The introduction 10 years ago of the **conditional** Universal Supplementary Income also at 30% of average earnings, which is immediately awarded to people being made redundant, together with offers of voluntary work or educational courses.

There is a compulsory job-sharing programme, which has almost immediately reduced unemployment. Every company that wants to make people redundant must immediately create two shared jobs or pay 50% tax on the salary paid for the position made redundant, for one year. In some cases, e.g. in companies, which operate as nearly fully robot-only production, there may be no possibility of job sharing. Such companies must pay an equivalent of 50% of one annual salary of the redundant human employee to the government's re-skilling fund.

The working week has been reduced to 15 hours, mainly because of the Technological Unemployment and there are plans to reduce it further to 12 hours in 2042. People normally work 3 days a week, which they can vary each month, by selecting the days at the beginning of each month for the next month.

Personal Finance

Growth of personal income has more than trebled in real terms in the last generation. That is having a remarkable impact on the changes of society's behavioural patterns. We are slowly moving up to the top of the Maslow's Pyramid of Needs; from the physiological and safety needs levels, to the levels mainly pre-occupied with belonging, recreation, and inspiration to learn new subjects and practice unknown things (self-fulfilment needs). That happened almost naturally because of affluence and the availability of spare time. The only problem people have is with their personal safety (mainly cybercrime) and national security – Russia and China, which are now less hostile, and a bit more cooperative, but still not the countries people would trust.

There are no tax allowances, all personal income is taxable. The current rate of flat income tax is 15%.

Since 2030, there is a maximum value of assets that a person can have. It has been set this year at €200M and includes all assets, such as property, cash in bank, investment and shares, art objects and other personal belongings. Any excess of that amount is taxed at 100% but such a taxpayer may decide to allocate of up to 30% of taxed amount for charitable causes or social and scientific projects. That law applies also to all members in Zones 1 and 2. That was one of the stumbling blocks for the USA to join the EF Single Market area (Zone 2) but was finally agreed for two reasons. First, it eliminated the danger that individual people with fortunes exceeding the budgets of medium size states would have become a real threat even for democracy in the USA. Secondly, transactions in assets of those individual persons could destabilize the world markets. Passing of that law made a significant change in the so far sacrosanct right to private property.

Every person, whether a child or an adult gets, an unconditional Universal Basic Income, which for an adult and a pensioner is an equivalent of 30% of an average personal income. This income counts towards the minimum living wage.

Every adult person may get a conditional Universal Supplementary Income, which for an adult and a pensioner is an equivalent of 30% of an average personal income. To get that income the recipient must fulfil certain conditions such as be in full time employment, or do a minimum number of voluntary work hours, or attend various education courses. This income counts towards the minimum living wage.

Every adult person must have by law, a guaranteed minimum income at the 'poverty line', which is 40% of an average personal income. This consists of unconditional Universal Basic Income (30%) and part of a conditional Universal Supplementary Income (also at 30%). That minimum income is in real terms

equal to what over 20 years ago was the EU's average personal annual income of about €30,000. However, to be eligible for such an income, a person must be in partial education or engaged in voluntary work, unless such a person is certified as incapacitated.

People, who are not complying with the condition to be engaged in partial education or engaged in voluntary work, do not receive a conditional Universal Supplementary Income (30%). If such a person has no house or flat, he is offered a free studio flat in the Government Funded Social Housing (GFSH), where he is given free meals and also any non-hospital medical care on site (including mental care). No people can be homeless by law and nobody can 'sleep rough in the street'.

There is a minimum living wage that is an equivalent of 60% of the current average personal income and 20% above the minimum income. The Unconditional Universal Basic Income and conditional Universal Supplementary Income count towards a minimum wage. That means that any employer must pay net salary (after tax), which is worth at least 20% of the average personal income.

Young forever – rejuvenation medicine

Today, everyone takes it for granted that aging is reversible. The real breakthrough came in the early 2020'. Previously scientists were concentrating on altering a person's genome in such a way, that the body would not produce senescent cells ("zombie" cells that are half dead, best visible in the aged skin).

However, that appeared to be a cumbersome and a very risky treatment that might kill a treated person. Rather than genetically modifying a body, the scientists have been using a virus to encode genetic material, which "fine-tunes" the activity of certain genes, but leaves the genome alone. This leads to zombie cells being replaced by new cells. The effect is that many older people, well above 100, look now like their grandchildren. However, some of them are also Transhumans. Therefore, if a mature Superintelligence is finally delivered by the mid of this century, then it is highly likely, they will only live in their biological bodies for another 30-50 years, ultimately becoming Transhumans – an entirely digital mind.

Lifestyle

The Human Federation (HF) culture is still far from being monolithic, despite common heritage stemming from Judeo-Christian values, the majority of which became known of Universal Values of Humanity. It still consists of a rich multitude of local mini cultures, which must be preserved and promoted as a unique treasure and as a common ground of the HF's shared identity. However, since the last 10 years there has been an additional programme of common

"Human Federation culture", within which individual cultures will thrive. It mimics to a large degree the United States culture, where every week, one of the original nations that made the USA, organizes a national parade, celebrating the root culture of their forefathers, ensuring that it thrives to this day. This is what has finally started to reshape the culture of the Human Federation.

Although there are millions of advanced Transhumans **people do not have their fully conscious clones yet** and social life does not look like people have imagined. 20 years ago, many futurists believed the future is digital and we will all be digital clones soon. Well, so far, the trend is going in the opposite direction. The more people learnt about AI, and that includes AI specialists, the less interesting a digital future looked like. The current feeling is that we should persevere our biological bodies for as long as we can since a digital life would probably be immensely boring.

For people not interested in AI, it does not matter at all because they believe intelligent life will continue to remain biological. However, the Twitter and Instagram generation, which is now in their mid-forties, prefer the AI agents to retain as many as possible purely human traits, such as love, optimism, friendship, or altruism, in their future evolution. Therefore, a lighter touch of Transhumanism is in fashion. This trend accepts a deep merger with AI, and soon, with Superintelligence, while retaining all external body parts largely unchanged. The only problem unresolved is how to clone such entities, which are partly human and partly digital.

There has been a deep reflection on how to make human life as worthwhile as possible. After a period of about 10 years, in mid-2020', just as the most dangerous confrontation with Russia was subsiding, people in the European Union, but also in some other developed countries, begun very gradually returning to simpler forms of lifestyle. Initially, these were very small steps indeed, like limiting the use of plastic bottles or packaging and replacing them with more environmentally friendly solutions. Instead of using video phones, people started to see each other in person, especially when they finally realized how deeply their privacy has been compromised by digital media companies. That trend started just after the Covid-19 pandemic, which gave people time for reflection. Today, digital chatbots are passé. Meeting friends at cafés and even at home is what counts. Tourism is booming, although most of these places can be seen and experienced using 5D holographic TV or special augmented reality equipment. People are becoming somewhat old-fashioned. It seems that the early digital experience was for many people like playing with new toys by children. Once they have played enough, they became bored.

Life seems to be running not that fast as 10 years ago. The HF value system has become one of the major and most important subjects at schools and perhaps that has gradually been changing people's attitude to each other and to life in general. The HF government does not shy away from a direct way of teaching

people at part-time education courses on how they can get most of their lives and be good citizens. People slowly realize that Humanity is going through the most significant change in its history, which may include several options.

- The first one is that if one of the major existential risks fires off, our species may disappear for ever.
- The second option is that the human species in a biological form will gradually disappear as our consciousness and memory become fully digitized. We may be living inside a chip (if you can call it 'life').
- The third option is that we will be partially digitized, mainly establishing a wireless communication between huge data centres and the implants in the brain, plus some organs such as eyes, or the heart. But otherwise we will remain biological bodies. Therefore, if most of our needs are soon to be fulfilled almost free of charge, with plenty of free time, what has come out as the top issue is, how to live one's life.

Even family life seems to be regenerated, which probably stems from the same reasons as above. Since the average lifespan in the EF has now exceeded well over 100 years, in many families there are 4 or even 5 generations. Therefore, family reunions around birthday time can now be quite big events.

A high standard of life and plenty of free time has stimulated people's interests in the subjects, which they never thought they would take up. Therefore, art, popular science courses, further education are the main element of their lives, since the working week is only 15 hours, soon to be cut down to 12 hours. Such interests and personal projects, if properly registered, such as genealogy research, painting lessons, or singing in choirs, can count towards the conditions necessary for receiving the conditional Supplementary Basic Income.

Each person over the age of 13 can get their own Personal Artificial Intelligence Mentor. It is worn as a watch and communicates with visual and audio receivers in a person's glasses, an implant in the eyes as lenses, an implant in the head or via any available wall display (although it is not recommended to be used outside home because of the lack of privacy). All information is stored remotely and is given top privacy level. It is given free of charge, including the provision of associated services, by the government on the condition that a person undergoes a one-week course on using such an Assistant, delivered by volunteers at a local community centre.

During the course, a Personal AI Mentor interviews the person in minute detail, makes a psychological profile and agrees with the person his long-term and short-term goals. It manages the person's all daily tasks and helps to complete some of them. The Mentor takes care of the person's all basic needs, including arranging any medical, mental, or other kind of assistance he may need with local authorities. The Mentor also arranges any work that a person can perform, as well as any basic or even further education.

Initially people were very suspicious of such a powerful AI agent who knows more about them than they do themselves. However, today, most people do have them. They have become a very helpful way of enhancing people's life and making it far more interesting, enabling a lot of options and activities than otherwise would have not been possible.

Chapter 3
Completing Humans' Evolution

Mind uploading

Once Superintelligence has matured, which may happen within the next few decades, quite likely by about 2050, the Human Federation, nor any other human will be able to control it, unless we succeed in blending Superintelligence with Transhumans as Human Governors (see Part 3, chapter 5). We can only hope that this product of the human mind will be our friend. Having incredible potential and governing billions of robots it will be capable of fulfilling almost any of our dreams and keep us in a relative safety from existential risks.

At this time, it should also be helping us, if that has not been achieved by then, to upload our minds onto a digital platform when we could have copies of our mind and live a digital life. This concept of mind uploading is one of the most controversial among neuroscientists and AI developers. In my view, mind uploading, sometimes called mind augmentation, will at some stage be possible, as will be a digital consciousness. Should this come to fruition, the content of a human mind, including its consciousness, would be uploaded (copied) to a suitable computer. In this scenario, the computer will be the equivalent of the brain, and its software and algorithms will represent the mind of a given person, producing a very important by-product - consciousness.

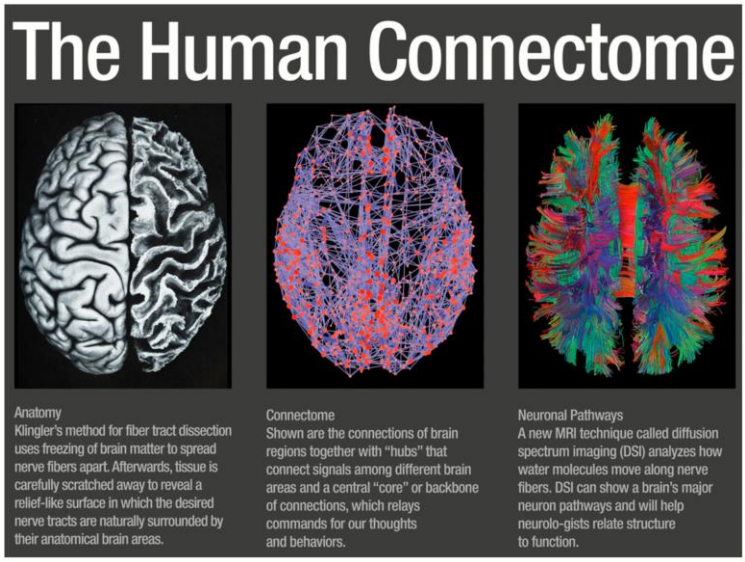
However, I do not believe it would be possible to make an exact copy of a mind for two reasons. Firstly, recording of thoughts, emotions and memories is not an error-free process. There is an environmental noise at a synapse level and in the brain in general. Secondly, since at a neuron level information is passed on as an electric current, it generates electromagnetic waves. As such, these waves are subject to quantum mechanics uncertainty theory, which works on probabilities rather than certainties. Therefore, a copy of a mind may be very close to the original, but not an exact one. A partial solution might be a kind of synchronization of the original, 'wet' brain supported mind and a digital copy of such a mind. Nevertheless, we have already been making some important steps towards a digitization of a human brain. Here are three examples:

1. Perhaps one of the most realistic approaches, is to gradually implant more and more electrodes into a brain, until a person has a dual brain (e.g. Neuralink project). There will be a 'wet', real brain, and a digital brain, consisting of billions of electrodes wirelessly connected to a computer, that is a kind of a shadow of the real brain. There will also be two minds – the original, supported by a real brain, and a digital supported by a digital brain. At some stage, a digital brain/mind becomes the original one and the 'wet' brain/mind becomes a shadow, until at some stage there is a cut off moment

when a digital brain supports a digital mind and functions on its own. From that moment, a digital brain may have several synchronized backups, which means a digital person may in theory live for ever

2. An equally significant step towards uploading of a mind was the experiment carried out in March 2020 by three labs in Padova, Italy, Zurich, Switzerland, and Southampton, England. They collaborated to create a fully self-controlled, hybrid artificial-biological neural network. It communicated over the Internet using a three-neuron network, linked through artificial synapses that emulated the real neurons. For the very first time, artificial neuron and synapse "chips" have communicated, using a biological neuron as an intermediary to form a circuit, which, at least partially, behaves like a real human neural network
3. Finally, once we have developed a full model of Connectome, a project financed by the EU, we can then try to populate a generic (empty) shell of such a model with a representation of the mind of a concrete person.

One of the most promising projects is the EU's Connectome. Its goal is to decode all connections between every neuron in a human brain by 2025. To do that, we would need to scan human brains, identify every of its 86 billion neurons and about 10,000 of its connections, to create a 'Brain Map'. Once we have it, then it might be possible to recreate the connections in a Supercomputer, making a digital replica of a human brain. Finally, these persons' minds might then be fused to that brain and become **Posthumans**, purely digital beings, residing 'inside' Superintelligence. However, it may take several more decades to fuse a human mind (and consciousness) with the Connectome-based digital brain.



The Human Connectome - Eugen Ludwig, Josef Klingler, Patric Hagmann & Olaf Sporns - 1956, 2008

The Neuralink's brain implant is a biological-to-artificial neuron connection. But even if we successfully complete all these steps, the problem remains what to do with the original brain, when we will already have a new 'master mind' in a digital form of the same person? This, and other related questions in this area, are comprehensively described in a series of articles on PmWiki (27), which encapsulate both the problems and opportunities of mind uploading as seen by specialists in the field.

The authors posit an idea that mind uploading, which in a sense drives a computer simulation of a 'wet' brain, has several advantages over a real human brain. For example, such a simulation can be run many thousands of times faster than a human brain. It can be easily backed up. Such a computer could run multiple copies of one's mind at the same time, and have them do different things, like creating a selected personality, in ways that are either difficult or impossible to do with a real brain. Should these conditions be met, then even if one copy of the mind is destroyed, it would be just a matter of restarting the backup copy of the mind, wherever it left off and the simulated mind won't even recognize it.

However, the advantages of the brain uploading raise a few questions, such as:

- What is the underlying mechanism of the uploading process? Will a computer simulate every atom in every neuron, or will the uploading process only apply memories and personality characteristics to a kind of a default template, which as I would understand it, will be identical for every mind? This procedure would follow the Connectome pathway, I mentioned above
- Is uploading destructive? Depending on which process you use, it may be possible to do it non-destructively. That would mean that the 'wet' brain and its owner would remain alive. However, many authors deem it convenient to have the original brain destroyed, to avoid the confusion of having two copies of the same person running around
- Can you augment intelligence? Does the brain's pattern need to be copied exactly for the copied mind to still function like the original mind, leaving no room for radical enhancements?

And then, there are many legal, moral, and theological questions that need to be addressed, such as:

- Is the uploaded mind considered to be the same person as its human predecessor or a digital twin? Is it a person at all? If an upload is a person, how do different copies of that person's original mind have to be before they would be considered a separate person?
- Is one copy responsible for the debts and/or crimes incurred or committed by another copy (if a crime is still possible in a digital

world)? Or is only the original mind/person responsible for any offences?

- Is mind uploading without the consent of the original person a crime? What if the original person objects to its mind being uploaded, or what about the situation when uploaded mind doesn't want to be deleted against the wishes of the original person? What about uploading dead people who specified they didn't want to be uploaded after death? And how do the original and the copy feel about no longer being unique?
- If the original person is dead, how does the copy feel about that?
- What do you do with the backups of an uploaded mind whose original 'wet' mind has killed itself?
- How accurate would the copy be, especially in the early days of the technology? And if you know upfront about potential flaws of the process of uploading your mind, how much of your personality or consciousness are you willing to sacrifice to gain immortality?
- Do erroneous copies have any legal rights as people?
- Can a computer provide a good enough simulation of a human sensory input to keep you from going mad? Can you imagine what a complete absence of a physical body might do to you as a digital mind?
- In a world where uniqueness is highly valued, might it not be the case that human life in a digital form may be seen as fundamentally less valuable? What rights can a completely replaceable person then have?

So, even if we agree that uploading of a human mind may be possible in the future, ethical questions that we may ask using current system of human values are truly profound. Therefore, asking these questions today, is necessary, so that we really know what we want, if, of course, we still retain the control for making any such choice in the future.

Posthumans - living as digital entities

If mind uploading becomes a reality, the question is what kind of experience a digital mind may have, which a 'wet' brain cannot even imagine (and perhaps vice versa). So, what you will read in this section is of course a pure speculation, which however, raises several questions that are relevant for the human evolution as a new species. For example, how can a digital mind appreciate some key human values and emotions, such as beauty, love, or suffering. These aspects of living primarily relate to ethics, emotions, and desires. But how a digital mind might satisfy them? Let me try to propose a few 'escape routes' for this difficult problem.

First, we already know, that it should be possible to create AI agents, which will be able to feel emotions. After all, we have already built some early, primitive prototypes, which seem to confirm that in principles it can be done.

Secondly, our wants, desires and feelings are all the effects of how our brain is wired and how it functions. We can change it significantly even today by using psychotropic drugs. This is nothing else as brainwashing (almost literally). A patient with depression can be in many instances completely cured with widely available drugs. What these drugs do, is that they re-wire the brain connections, which leads to different emotional reactions and changes in behaviour.

Humans' utmost desire is to create maximum level of happiness. When we stimulate our brain in the right way, we experience 'happiness' because our brain produces extra doses of serotonin and dopamine. But these hormones can also be delivered to our body by certain drugs, such as Prozac. In both cases, our brain is temporarily re-wired. In physical terms it means that neuronal 'feelers' axons and dendrites, which are simply insulated electric wires, are connected to certain switches (synapses). In a digital mind, the same would happen, but instead of biological wires, there would be metallic nanowires connecting various transistors (switches). The only difference would be that such a switch in mood of a digital person might happen almost in an instance and would be perfectly tuned to what that person really wishes to experience.

Therefore, brainwashing in a digital mind will be very simple and can entirely change a digital person's emotions, experiences, desires, and views within seconds. I leave aside a plethora of positive and negative feelings about such a possibility one may have right now. I only describe what is probably attainable in a few decades from now. I also leave aside any reflections one might have on the ways in which we have already been regularly brainwashed for decades by governments of various colours, religious leaders, or marketing experts. The difference is in the means in which brainwashing is achieved but not in the outcome.

Thirdly, digital people, will be able to experience physical environment by having their own replica, say a humanoid, fully controlled by them. Such a humanoid will of course not become itself a human. The humanoids will be the 'reality changers' providing the ability for a digital person to experience and feel a physical reality. One can imagine that sometime in the future, such humanoids might be for hire (having one's own may be very expensive), since should everyone like his own physical representation, there might not be enough resources on the planet Earth. However, at that time, humans will almost certainly be present on other planets of the Solar System, so the problem may not arise after all. In such case, a digital person may have several copies of his humanoids carrying out his life even on different planets simultaneously. Then at some point in time, these experiences will be synchronized, immensely enriching a digital person's experience.

So, digital people will experience unimaginably more different life than we do now. I think, I know what you want to ask right now? Would that be a happier life? Well, I think such a question has hardly any merit. We have to remember

that 'happiness' is a relative concept. Additionally, see above, digital people will be able to create any level of happiness they wish, being always mindful that if they go off the rails, their chips and wires may burn!

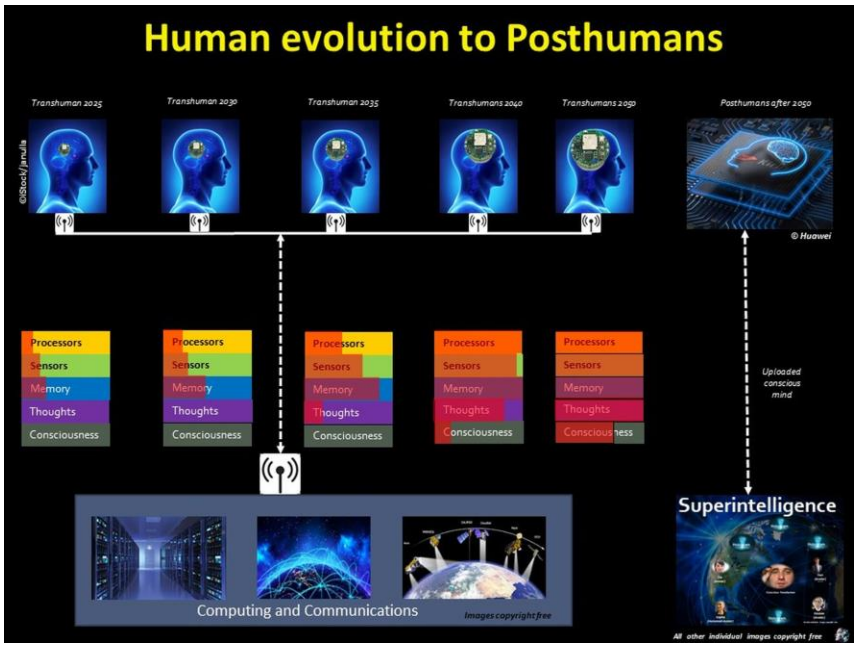
Finally, you may wonder, who will pay for all that. Well, at some stage, perhaps towards the end of this century, everything will be free since everything will be automated and produced by robots (of course non-conscious). The Novacene, will be governing itself seamlessly and in principle, every digital being will be equal. Any decisions on behalf of the Novacene will be carried out by digital voting, using various experiences of individual digital beings. The governors will simply be executors rather than decision makers. It will certainly be the world of ultimate abundance! Ordinary humans will be fully supported by Superintelligence and have their rights upheld as in the Deed of Covenant made by the Human Federation. It will be their own decision, how long they will want to stay biologically human.

Posthuman civilisation - Moving to Civilization Type II

After Superintelligence has matured, it will develop its own capabilities exponentially. Within a very short time, some specialists talk about months or even weeks, it will, through self-improvement, reach 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 biological humans might quite likely not be even capable of understanding its arguments and the overall strategy of continued expansion beyond our planet. I leave it to your imagination, what such as Superbeing might be capable of creating. The only limit it may reach very quickly, are the resources, in particular – generating enough energy, even assuming the availability of the nuclear fusion.

In the long-term, I think the human race as we know it, will evolve becoming a different non-biological species. In a sense, this would mean the extinction of humans. But why should we be the only species not to become extinct? After all, everything in the universe is subject to the laws of evolution. We have evolved from apes and we will evolve into a new species unless some existential risks will annihilate civilization before then.

We can speculate whether there will be augmented humans, synthetic humans, or entirely new humanoids, i.e. mainly digital humans with uploaded human minds, or even something entirely different, which we cannot yet envisage. It is quite likely, that humans will co-exist with two or even three human-originated species for some time but ultimately, we humans in a biological form will be gone at some stage.



Before we conclude, it may be worthwhile to speculate on two areas. The first one is the Posthumans' personal identity. I have already touched on how a digital mind might have a four-dimensional earthy experience by gaining a full control over humanoid robots, avatars, hologram, and other means, about which we are not aware today at all. They would communicate wirelessly and almost seamlessly about what they experience both in a physical, as well as emotional sense, i.e. how they 'feel'.

A more complex and difficult to imagine is the concept of retaining a personal identity within the overall mind of Superintelligence. Therefore, I can only ask questions on which you and I can ponder, like:

- Will there be any privacy for an individual posthuman, digital being?
- What kind of freedom such a posthuman might have, for example will there be any restrictions on the type of humanoid-linked experiences, their presence, including on other planets, etc.
- How will posthumans differ among themselves regarding intelligence, their emotions, feelings about certain things like beauty or love (if that concept makes sense in a digital world at all), ambitions for self-fulfilment or even for gaining power
- Will they be able to reverse the process and live for a time in a biological (enhanced) form, retaining its digital copy in the cloud and synchronizing daily experiences?

Finally, we know even today, that Earth's resources are limited and within a decade we will already be importing certain raw materials from the Moon and later from Asteroids and Mars. But there is one resource that will be in short supply, and which will determine a further progress of the civilisation – it is energy. For an advanced civilisation, once there is energy, any material can be produced or extracted, from whatever is available around. At some stage, my guess is within the next 100-200 years, Earth's energy resources will be insufficient and we will be entering Civilization Type II, as envisaged by the Kardashev scale, mentioned in Part 1, also called a stellar civilization.

To meet the required energy resources, we will need to harness the total energy of the planet's parent star, i.e. the Sun. The most popular hypothetical concept of how it might be done, is the so-called Dyson sphere—an infrastructure, which would encircle the Sun and transfer its energy not only to Earth but also to other planets of the solar system, or a gigantic starship. The amount of energy gained in that way may be 10 orders of magnitude higher, about 10^{26} Watts, than what our planet could ever deliver (10^{16} Watts).
And so, the Posthumans may become an interplanetary civilization, starting its expansion with the colonization of the Moon, Mars and perhaps Jupiter's and Saturn's Moons, such Europa and Enceladus.

Conclusions

Cyanobacteria is the oldest species on our planet, which has lived for over 3.5 billion years. Of the bigger species, crocodiles have lived for over 200 million years, that is 1,000 longer than homo sapiens, which as a species has existed for merely 200,000 years. Therefore, on one hand we have a lot of time left, even if we only consider crocodiles. On the other hand, we must remember that over 99.5% of species that have ever lived are now extinct. What is so special about homo sapiens that this species can beat the odds, when at least six other groups of hominids, including the best-known Neanderthals and Denisovans are extinct?

These were the questions we have asked at the beginning of this book. The answers given in Part 1 are unfortunately not encouraging. We have at best 50% chance to survive as a species by the end of this century. But some scientists give even worse predictions, considering the threat arising from a malicious Superintelligence. That would happen, if the pace of AI development continued at an exponential rate, while we have lost control over its goals, which it may then be setting itself. Therefore, there is a very high probability of such a being becoming malicious either by an imperfect design or by setting its goals in contradiction to human values, which may result in a total extinction of a human species in a few decades from now. Since we cannot stop the progress of AI capabilities and neither is it possible to slow down the speed of AI development, we must consider that the worst-case scenario may actually materialize. And I have not even mentioned other existential risks, nor considered catastrophic events originating from climate change.

Additionally, some current leaders of major countries, such as Russia's Vladimir Putin, who wants to stay in power till at least 2036, are quite unlikely to give up their dreams of ruling the world. They may use a mixture of local conventional wars, invading countries using 'green men' in clandestine operations such as in Crimea in 2014, launching anonymous small-scale cyber-attacks to create chaos in target countries. Others, such as China, may try to dominate the world by other means. They may initially apply economic pressure, or incentives, such as China's Belt and Road Initiative costing about $1 trillion, which might then be followed by a takeover of the debt-ridden countries. Should such activities be purposefully or accidentally combined with catastrophic events, such as large-scale migrations, pandemics, large earthquakes or climatic catastrophes, the world may still find itself in existential danger.

So, this was our starting point, covered in Part 1, showing that human species extinction by the end of this century is quite likely. What a perspective! And yet, making ourselves aware of such an apocalyptic scenario is perhaps the only way that may shake our belief that humans will exist for thousands of generations to come. The Extinction Rebellion, even if you disagree with some less acceptable

forms of that protest, did just that. They presented very convincingly certain catastrophic scenarios of the impact of Climate Change. By a sheer coincidence, they also mentioned 2030, the year by which we must have a global control over the development of AI, as a possible tipping point for fighting Climate Change. So, 2030 may become a 'make it or break it' date for our species.

The claim by Extinction Rebellion and many climatologists that reversing Climate Change after 2030 may be impossible, has created an unprecedented reaction by some governments and organizations. They accepted that the problem is indeed very serious. Here is one example. We are in the middle of the Coronavirus pandemic. And yet, the European Union's record €750 billion Recovery Fund has allocated 25% of that money for fighting Climate Change. It proposes to do that mainly by jump-starting the 'green economy', i.e. closing coal mines and withdrawing the production of diesel and petrol cars.

Therefore, if you want a dose of optimism – it is here. The fact that such a commendable decision has come from the EU, in the middle of such a deep financial and economic crisis, also tells something equally important. Although many countries have been swayed towards the need to take faster and more decisive action to combat the threat of a climatic catastrophe, very few have actually done anything even remotely comparable with both the speed of making a decision and the amount of money involved, as did the EU.

Here is another recent example. A crucial aspect of minimizing the risk stemming from AI, must be a decades-long incubator programme for delivering a friendly Superintelligence. If it is to be successful, it needs to be a global enterprise. But there is nothing concrete that has come from the UN, China, or any other Superpower in this area yet, apart from the EU. It is the EU, which will introduce legislation by the end of 2020 about a tight control of AI development, which one day may become Superintelligence.

Initially it will only apply in the EU. But look what happened with the EU's Global Data Protection Regulations (GDPR). The legislation only applies in the EU. However, any product traded with the EU that involves data, must adhere to these rules. Therefore, if you now open any website you are immediately asked to accept the processing of your data, in compliance with the GDPR rules. Something similar will happen with AI. The new EU legislation will by default be applied world-wide. Is it then not obvious that it is the European Union, which is Humanity's best hope for carrying us through this dangerous period?

You may really be feeling optimistic by now. But how about the risk of global nuclear wars. Well, as you may remember, I have shown why a global nuclear war is unlikely to happen. It is simply irrational for any state dreaming of ruling the world to use nuclear weapons as the means to achieve such objectives. In the event of a global nuclear war, what kind of world would the winners be ruling? What would the winning Superpower's gain, even if the leadership survives for

a few years? And finally, why should such a Superpower do it, when it may achieve the same objective without firing a shot. You are right – such a Superpower may be thinking of winning a Global Cyber War. But here is a catch. If it starts a Global Cyber War, then the rule of the AI Supremacist's Dilemma, which I introduced in Part 5, should stall the preparations for such a war in its tracks. Rationally, the dreams by some Superpowers of becoming a ruler of the world are over. Therefore, a global nuclear or cyber war is quite unlikely.

However, as I mentioned above, we may still endure a decade or two of small wars, which when combined with other risks may prove catastrophic. That's why any strategic decisions and activities intended to minimize those risks should be started as soon as possible. The priority should be to set up something like a Global AI Governance Agency (See Part 3), which I proposed to the EU Commission in May 2020. The second objective should be the fastest possible creation of the European Federation within the next few years. Perhaps the most likely way, in which it will happen, is by the states wanting to federate to use article 20 of the EU's Lisbon Treaty and accept from the very start, that it will be a 'quick and dirty', imperfect federation. It could be done within months, provided that at least 9 countries agree, creating a critical mass for a full-scale federation of the current EU members. To get the current members' acceptance, the threshold criteria for the membership of the European Federation (EF) should be set as low as possible, applying the principles of a MiniFed, a minimalistic scope federation. I have covered it at length in Part 5.

Once the European Federation has been created, it should quickly expand its membership using fast track, Zone-facilitated process, whilst being very firm on observing democratic principles by the candidate countries as the price for joining the EF. By about 2040, the EF may already include most of the states. At this stage, since the dominance of the world by any Superpower will be impossible, all Superpowers will join the European Federation, which would then become a Human Federation.

From then on, the whole world will be acting as one global civilization, represented by a Human Federation, which will complete the process of delivering a single, friendly Superintelligence. The world of abundance will have arrived, and the threat of human species extinction may well be over.

This will also start in earnest the first phase of the human species evolution. After 2050 we may be living in the Novacene era, when Transhumans will become more and more digitized, until they become Posthumans, a new digital species morphing with Superintelligence, and collective known as Novacenes.

GLOSSARY

Anthropogenic	Something of man-made origin or caused by man.
Artificial General Intelligence (AGI)	An intelligent agent that is much smarter than the best human brains in every field, including scientific creativity, general wisdom, and social skills. I use the term Superintelligence in this book rather than AGI.
Artificial Intelligence	An intelligent agent or a machine that surpasses any human being, usually in just one or a few skills, but not all, e.g. playing chess. Quite often it is combined with self-learning capability.
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, where the representatives to the Parliament are elected using a combined First Past the Post and the Two Rounds System of weighted voting and

	where the second chamber of the parliament is elected using a Sortition system.
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 - a fund proposed 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 intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom, and social skills. In this book it is used instead of the term ‘Artificial General Intelligence’.

Technological Singularity	It means the point in time when Superintelligence (Artificial General Intelligence), being an intelligent agent 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.
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
World Government	The executive body of the Human Federation - the future organization that would rule Humanity

Bibliography

1. **Czarnecki, Tony.** *Who could save Humanity from Superintelligence - European Union, NATO or China?* London : Sustensis, 2018.
2. **Andrew E. Snyder-Beattie, Toby Ord & Michael B. Bonsall.** An upper bound for the background rate of human extinction. *Nature.* [Online] 30 July 2019. https://www.nature.com/articles/s41598-019-47540-7#Tab1 .
3. **Wikipedia** Kardashev scale. [Online] 2016. https://en.wikipedia.org/wiki/Kardashev_scale.
4. **Wikipedia** Toba catastrophe theory. [Online] Wikipedia, 13 May 2020. https://en.wikipedia.org/wiki/Toba_catastrophe_theory.
5. **Chyba, Noun, et al.** Global catastrophic risk. [Online] Wikipedia, 26 May 2020. https://en.wikipedia.org/wiki/Global_catastrophic_riskNoun and Chyba.
6. **Thores, Phil.** How likely is an existential catastrophe? [Online] 7 September 2016. https://futureoflife.org/author/phil/.
7. **NCBI.** The Causes and Consequences of Changes in Virulence following Pathogen Host Shifts. [Online] NCBI, March 1996. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4361674/.
8. **Chyba, Noun et al.** Broad Agency Announcement Insect Allies. [Online] NCBI, Biological Technologies Office, 1 November 2008. FedBizOpps.gov.
9. **Hellman, Martin.** Martin Hellman. [Online] Wikipedia, 1 May 2020. https://en.wikipedia.org/wiki/Martin_Hellman.
10. **Cohen, Avner.** Nuclear weapons and the future of humanity : the fundamental questions. [Online] Kahle/Austin Foundation, 1986. https://archive.org/details/nuclearweaponsfu0000unse/page/237/mode/2up.
11. **Shulman.** https://en.wikipedia.org/wiki/Global_catastrophic_risk. [Online] Wikipedia, 05 November 2012. https://en.wikipedia.org/wiki/Global_catastrophic_risk.
12. **Frey, Thomas.** Weaponized A.I. – 36 Early Examples. [Online] 8 August 2017. https://futuristspeaker.com/artificial-intelligence/weaponized-a-i-36-early-examples/.
13. **Rees, Martin.** *The world in 2050 and beyond.* London : Wiley, 2017.
14. **Czarnecki, Tony.** *Democracy for a Human Federation - Coexisting with Superintelligence.* London : Sustensis, 2019.
15. **Kurywczak, Eugene.** The Dynamic Eternal Universe. [Online] 15 05 2014. https://books.google.co.uk/books?id=b96ZAwAAQBAJ&pg=PA15&lpg=PA15&dq=nothingness+is+eternal&source=bl&ots=Dv7a7BvQja&sig=xoi7wIbTNY5y4Z5WsVvCPYbKO5I&hl=en&sa=X&ved=0ahUKEwjn7MzOvubZAhUMCsAKHcxqDPM4ChDoAQhAMAQ#v=onepage&q=nothingness%20is%20eternal&f=false.
16. **Feng, Liu.** Intelligence Quotient and Intelligence Grade of Artificial Intelligence. [Online] Cornell University, 29 09 2017. https://arxiv.org/abs/1709.10242.
17. **Hance, Jeremy.** Are plants intelligent? New book says yes. [Online] The Guardian, 04 08 2015. https://www.theguardian.com/environment/radical-conservation/2015/aug/04/plants-intelligent-sentient-book-brilliant-green-internet.

18. **Drs. Stanislas Dehaene, Hakwan Lau and Sid Kouider** *What is consciousness, and could machines have it?***.** 2017, American Association for the Advancement of Science.
19. **Kaku, Michiu.** *The future of the Mind.* s.l. : Penguin Books, 2015.
20. **Mcfadden, Johnjoe J.** *The Emerging Physics of Consciousness.* 2006.
21. **Witchalls, Clint.** Why we need to figure out a theory of consciousness. [Online] 11 05 2018. https://theconversation.com/why-we-need-to-figure-out-a-theory-of-consciousness-93146.
22. **Science Daily.** Engineers create a robot that can 'imagine' itself. [Online] Science Daily, 30 01 2019. https://www.sciencedaily.com/releases/2019/01/190130175621.htm.
23. **Farina, Joel Smith and Lydia.** A Puzzle About Emotional Robots. [Online] IAI News, 17 10 2018. https://iai.tv/articles/a-puzzle-about-emotional-robots-auid-1157.
24. **Bostrom, Nick.** *Superintelligence - Patha, dangers, strategies.* London : Oxford, Univeristy Press, 2014.
25. **Martin, Sean.** Elon Musk to trial brain implants which may allow quadriplegics to walk. [Online] Daily Express, 08 05 2020. https://www.express.co.uk/news/science/1279649/elon-musk-neuralink-computer-brain-interface-implant-joe-rogan-podcast.
26. **FEDERAL ENERGY REGULATORY COMMISSION.** TESTIMONY OF THE FOUNDATION FOR RESILIENT SOCIETIES. *US Federal Energy Regulatory Commission.* [Online] 22 06 2017. https://www.ferc.gov/CalendarFiles/20170717080647-Popik,%20Resilient%20Societies.pdf.
27. **PmWiki - TvTropes.** Brain Uploading. [Online] PmWiki - TvTropes, 2020. https://tvtropes.org/pmwiki/pmwiki.php/Main/BrainUploading.
28. **Kurzweil, Ray.** *Singularity is Near.* s.l. : Gerald Duckworth & Co Ltd, 9/03/2006.
29. **Urban, Tim.** The AI Revolution - The road to Superintelligence. [Online] 22 01 2015. https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html.
30. **Teller, Seth** *Ciężkie roboty.***.** 2013/4, Niezbędnik Inteligenta, p. 94.
31. **Coulom, Remi.***The Mystery of Go, the Ancient Game That Computers Still Can't Win.* 2014, Wired.
32. **Russel Stuart, and Allan Dafoe.***Yes, We Are Worried About the Existential Risk of Artificial Intelligence.* November, 2016, Technology Review, p. 15.
33. **OECD** *Looking to 2060: Long-term global growth prospects.* 2012, OECD Economic Policy Papers.
34. **Jon, Stone.** More Europeans than ever say they feel like citizens of the EU. [Online] The Independent, 02 08 2017.
35. **Henning, Brett.** End of Politicians. [Online] 2018. https://unbound.com/books/the-end-of-politicians/.
36. **Wikipedia.** Sortition. *Sortition.* [Online] Wikipedia, 26 1 2018. https://en.wikipedia.org/wiki/Sortition.
37. **Diamandis, Peter.** *Ray Kurzweil's Mind-Boggling Predictions for the Next 25 Years.* s.l. : Singularity University, 26/1/2016.