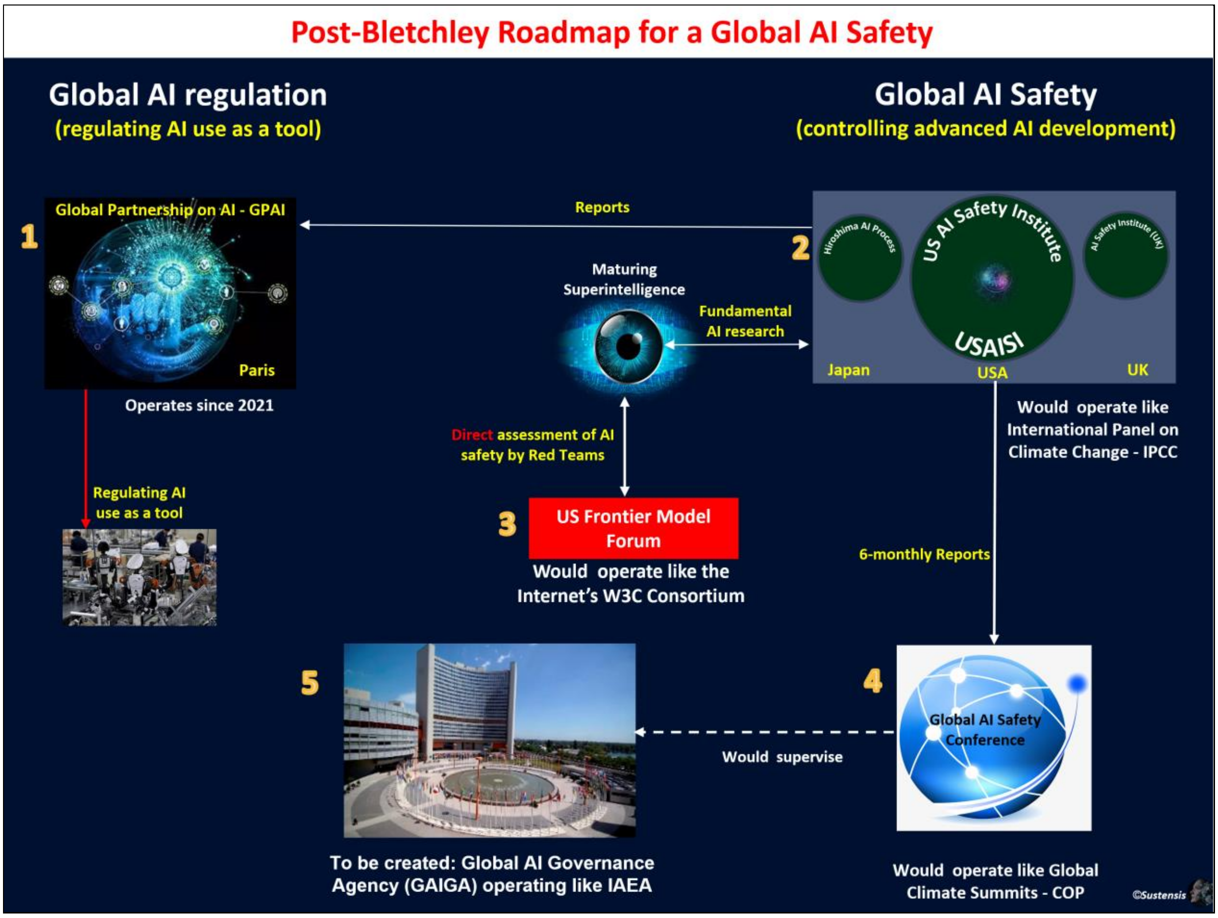
London, 6 November, 2023

What next after the AI Safety Summit?

The AI Safety Summit held in the UK ended with the Bletchley Declaration signed by the EU and 28 other countries. The UK declared the setting up of a UK-run AI Safety Institute (AISI). But on the eve of the Conference, President Biden signed an Executive Order, establishing U.S. Artificial Intelligence Safety Institute (USAISI). In addition, on 25th October, Japan announced concrete measures made in accordance with the Hiroshima Artificial Intelligence Process, established at the G7 Summit in May 2023. Finally, the G7 Group created in June 2020 Global AI Partnership (GPAI, which could have been seen as the right forum to organize such a Summit.

This is a typical political battle for making countries globally important, as may be the case with the UK, which after Brexit is desperately trying to prove that it is still a global power. Notwithstanding the fact that the Summit was very important and needed, it should have not led to a competition about which country is the most important in AI. That question has been decided some time ago and it is obvious it is the US, where 2/3 of all AI resources are located and where all leading AI models have been created. The only justification for such a duplication of AI Institutes may be acceptable if each country has such an institute *in addition* to one Global AI Research Institute. Great Britain by organising the summit has stepped forward and wants its Institute to have both national as well as a global role. It is now responsible for the preparation of a 6-monthly Report for the next Summit planned in Korea. The UK only assumes such a role since it is not explicitly mentioned in the Bletchley Declaration. But the US Institute may similarly expect that its Institute becomes the centre of global research on AI Safety.

The diagram below shows how the co-operation between various organisations regulating and controlling AI looks after the Summit, why also pointing to the next important step – the creation of a Global AI Governance Agency (GAIGA) with powers similar to the International Atomic Energy Authority (IAEA).



Post-Bletchley Roadmap for a Global AI Safety

Global AI regulation
(regulating AI use as a tool)

Global AI Safety
(controlling advanced AI development)

1 Global Partnership on AI - GPAI
Paris
Operates since 2021

Regulating AI use as a tool

Reports

2

Japan USA UK

Would operate like International Panel on Climate Change - IPCC

Maturing Superintelligence

Fundamental AI research

Direct assessment of AI safety by Red Teams

3 US Frontier Model Forum
Would operate like the Internet's W3C Consortium

6-monthly Reports

4 Global AI Safety Conference
Would operate like Global Climate Summits - COP

Would supervise

5 To be created: Global AI Governance Agency (GAIGA) operating like IAEA

©Sustensis

In my recent book "Prevail or Fail – a Civilisational shift to coexistence with Superintelligence", I have proposed 10 key steps necessary to minimize the existential risk of AI. They have served me as a reference to compare the steps aimed to control AI with those taken at the Summit and just before the Summit. The original wording is in italics.

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

   The Summit has for the first time acknowledged that the risk stemming from developing the most advanced AI is of a different kind than that which is posed by using AI as a tool in most domains of our lives. Therefore, the Summit has indirectly confirmed that regulation differs fundamentally from the need to control the process of development of the most advanced AI.

2. ***Undertake a comprehensive reform of democracy****, as it is a prerequisite for achieving effective AI development control and aligning it with human values. We must rebalance the power of governance between citizens and their representatives in parliament.*

   Unsurprisingly, in such a short conference that subject was not touched at all. Even if there had been enough time, the problem is of such political magnitude that none of the politicians would have wanted to raise it. However, pretending that the reform of democracy is an entirely different subject will have a direct impact on the effectiveness of AI development control, of which one key area is the AI alignment with human values. Who will confirm those values as applying to all humans if there is no Global Government, and democracy is on its proverbial knees? The researchers at the AI Safety Institute (AISI) will raise it anyway but as with setting the global limit for the temperature increase to manage global warming, the problem is almost impossible to resolve. So, there will have to be a compromise. The set of universal values of humanity will not be truly universal but perhaps applicable only to majority of the nations, serving as a reference for AI developers.

3. ***Retain control over AI governance beyond 2030****. While there is no scientific proof that AGI will emerge by 2030, just as there is no proof of the Global Warming reaching a tipping point by that time, we must develop AI as if AGI were to emerge within that timeframe*.

   There has been no mention in the Bletchley Declaration of any date when Artificial General Intelligence (AGI) may emerge apart from stating that "*Given the rapid and uncertain rate of change of AI, and in the context of the acceleration of investment in technology, we affirm that deepening our understanding of these potential risks and of actions to address them is especially urgent.*" Not agreeing on the likely date of the emergence of AGI may have similar negative consequences as it was the case with the Global Warming until the Paris Declaration in 2015, which declared 2030 as the tipping point for climate change. To initiate action on the ground, the most likely date for AGI emergence needs to be specified by the AISI before the next summit in Korea.

4. ***Create Global AI Regulation Authority (GAIRA)*** *by transforming the Global AI Partnership (GPAI). GAIRA should be responsible for regulating a global use of AI in society*.

   The intention of renaming GPAI as GAIRA was to make a clear distinction between AI regulation and AI development control process, so not a very important issue per se. However, not even mentioning GPAI, nor Hiroshima AI Process in the Bletchley Declaration can be seen as deliberate, so that the Summit could be proclaimed as the first and the only one addressing the risk of AI as an existential treat. It was a further evidence of how desperate the UK government is to re-instate its global role after Brexit. It may have partially succeeded but from a diplomatic perspective it was a disaster as neither President Macron nor Japanese Prime Minister did not attend the conference.
5. ***Create Global AI Control Agency (GAICA)*** *as a Consortium in the USA, since two-thirds of the AI sector is located there. Gradually expand it to engage non-US companies, including China.*

   This has materialized to some extent but in a different way. OpenAI, Microsoft, Anthropic and Google have created **Frontier Model Forum**, which may broadly fulfil the role of a body directly engaged in the creation of a safe AI. That is to be carried out among others by setting up so called Red Teams, which include top AI specialists trying to use the most advanced models in a devious and dangerous way, e.g., by breaking into them to modify their algorithms, creating a certain bias, or using them in a dangerous way threatening humans. There is an intention to invite other, non-US organizations. However, President Biden went further by issuing just before the start of the Summit, an Executive Order to create **US Artificial Intelligence Safety Institute (USAISI)**. The Summit itself resulted in the creation of the **AI Safety Institute (AISI),** which the UK believes should have a global role. Finally, the **Hiroshima AI Process** also falls into that category.

   Having four organization broadly responsible for the same task may lead to a healthy competition but also to some chaos. However, this is how politics works; every country wants to be seen as 'great'. The way out of that chaos may be to uphold the global role of UK-based AISI and consider the USAISI and other similar national institutes as following the guidance of AISI as the basis for national legislation.

   One of the main difficulties for the AISI Institute will be to identify which Models may have to be banned from a wider distribution. First of all, like with ChatGPT, we still do not know how this model has become so competent, only now discovering that its intelligence works in a somewhat different way than human's. Secondly, the researchers will have to apply their own values to consider the systems bias, as is already happening - AI developers are the judges themselves of 'what is right or wrong'.

In the end, what really matters is that there will now be an objective external organization assessing the most advanced models for their potential existential risk. AISI may soon be playing a role similar to the International Panel on Climate Change (IPCC). That is a significant achievement of the Summit.

6. ***Create Global AI Governance Agency (GAIGA)*** *under the mandate of the G7 Group. GAIGA would oversee both GAIRA (GPAI), responsible for regulating the use of AI products and services, and the GAICA Consortium, responsible for AI development control.*

   The lack of some concrete measurable and fairly easy steps to control AI development, such as non-proliferation of the currently most advanced NVIDIA H100 Hopper processors, was the Summit's failure. To be effective, such measures would require international legislation monitored by an organisation similar to the International Atomic Energy Authority (IAEA) in Vienna. We can only hope that an organisation similar to GAIGA, will be set up at the next Summit in Korea, or in a year's time, in France.

   In the meantime, we should look at the current EU's Artificial Intelligence Act (AIA), which may be strengthened as the result of the Bletchley Summit. Once it is ratified, it could become a model for an organisation like GAIGA. The success of the EU's GDPR regulatory framework which affects us all when we are asked to accept or deny cookies when working on the Internet, can be clearly seen. The huge fines imposed on the large organisations such as Google, META, or Microsoft, have minimized the negative impact that otherwise the applications of these organisations might have had. What will differ the EU's AIA is that it will be a legislation rather than non-obligatory assessment process.

   It is quite likely that the monitoring of the distribution of hardware needed for running the most advanced AI models might be possible by an organisation like GAIGA on a global scale. However, the control of software, especially algorithms, would be practically impossible with the current policy of open source distribution of the advanced AI models. That has already let to the creation of dozens of ChatGPT (GPT-4)-level models by various companies. If Artificial General Intelligence (AGI) emerges, there may be very quickly many instances of this new type of intelligence, each smarter than any human. Nobody will then be able to stop AGI proliferation, including spontaneous merging of individual AGIs, which will then self-improve at a nearly exponential pace leading to the creation of what I call an Immature Superintelligence. Although the emergence of AGI is unavoidable, the later it appears, the more time we will have to prepare for the coexistence with this new type of intelligence.

   Realistically, the only way, in which we may stop the proliferation of the most advanced AI models and consequently the spread of AGIs, would be to develop just one most advanced global AI, which ultimately will become Superintelligence. In that way two objectives may be achieved. The first one is that such an AI system will be immensely more powerful, requiring huge resources such as electricity, highest level performance processors (licensed) and best AI specialists in the world. Therefore, any devious AI system would be no match and could be quickly eliminated. Secondly, developing just one global Superintelligence is the only way to have any chance of controlling it and delaying the time when it will release itself from human control. By that time, it would have already been primed with human values with some experienced of co-existing with humans.

   Implementing such a concept will be a big challenge in the current geopolitical system. Nevertheless, seeing how quickly the Summit has been convened (in just 5 months) it is quite likely it will happen, but only once the dangers of advanced AI become more apparent. Therefore, if an organisation like GAIGA does emerge next year, it may be the start of the implementation of the next 4 steps. None of them could be discussed at the Summit even if the creation of an organisation like GAIGA would have been contemplated. However, I make a few brief comments to illustrate how our civilisation may have to change within a few years to maintain control over AI and reap the benefits that it promises, like the creation of a Global Welfare State.

7. ***Create Global AI Company (GAICOM)**. This would be a Joint Venture company to consolidate the most advanced AI companies into a single organization. Effective control over AI development will be impossible if it remains dispersed among numerous companies.*

   The first step might be a voluntary agreement between Google and Microsoft on developing a joint AI-driven Internet browser, with profits shared between the two companies.

8. ***Create Superintelligence Development Programme** (**SUPROG**) managed by GAICOM, matching China's efforts in the AI sector.*

   Since China signed the Bletchley Declaration, which in itself is a success of the Summit, there is some hope that instead of being a competitor, China may co-operate in developing one common, global Superintelligence, minimizing an existential risk, which if materialized, would have affected everyone.

9. ***Create a de facto World Government** perhaps initiated by the G7 Group, incorporating members from NATO, the European Union, the European Political Community, or from OECD.*

   Those who still dream about the creation of a truly World Federation may be upset but the only realistic way forward is the creation of a de facto World Government. The need is very urgent and it is not important whether it is called a 'federation', as long as it acts like one. We had examples how fast some major changes in geopolitics can happen. NATO was created within one year. The informal anti-Russian coalition supporting Ukraine was created within a month, despite the fateful NATO's withdrawal the previous year from Afghanistan.

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

    There was an indirect reference made frequently during the Summit and in the opening statement of the Bletchley Declaration: "Artificial Intelligence (AI) presents enormous global opportunities: it has the potential to transform and enhance human wellbeing, peace and prosperity." But it also admits that to achieve the world of abundance "AI should be designed, developed, deployed, and used, in a manner that is safe".

My book describes and justifies in detail all those steps, describing various AI control methods, addressing the need to postpone the moment of AGI arrival as far as possible and setting up the framework for the future coexistence with a mature Superintelligence.

Tony Czarnecki
Managing Partner, Sustensis
Email: tony.czarnecki@sustensis.co.uk