A Proposal for

A Global AI Governance Agency to mature AI safely into Superintelligence

1. 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.

The problem is that these proposals have remained just that – proposals. Until May 2020, there has been not a single UN resolution in this area. But even if it had been one, it would most probably face the same problem, typical of many other areas of the UN activities – the inability to enforce the UN's decision. Therefore, looking at the success of the Global Data Protection Regulation (GDPR), it is more likely that such a global AI-Governance legal framework may be created using the EU's proposals, implemented in a similar way. **I understand that this Proposal has a similar objective.**

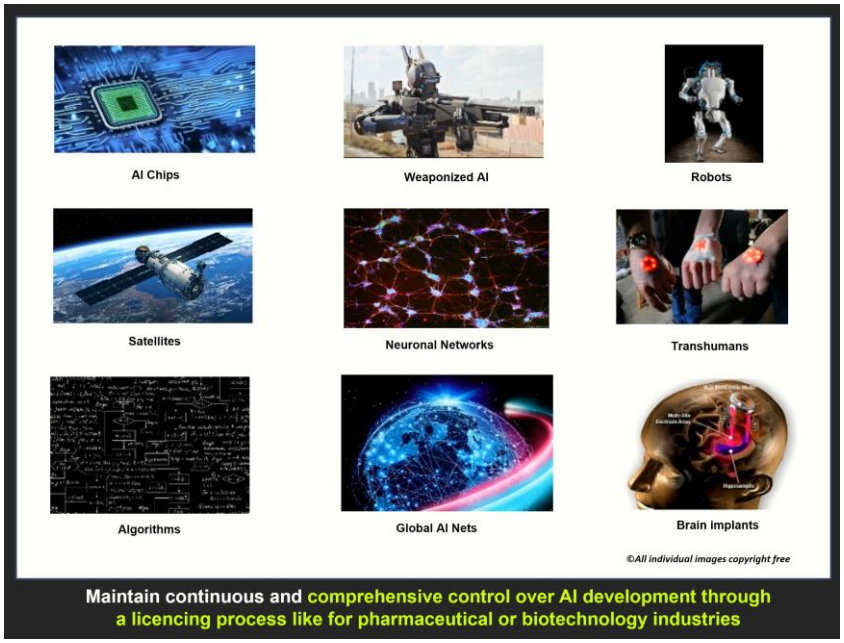
I believe there is a need to establish a single **Global AI Governance Agency**, more urgent in the post Covid-19 pandemic period, because of a likely acceleration of AI development. Additionally, Stuart Russell, one of the most imminent AI scientists, as well as many other top AI specialists, believe that we may not be capable of controlling advanced AI after 2030. Therefore, we must establish a full global control over AI development 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 a global 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) in Vienna, 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 (robots, AI-chips, brain and body implants, visual and audio equipment, weapons and military equipment, satellites and rockets, etc).

It should also cover the oversight of AI algorithms, AI languages, neuronal nets, brain controlling networks and brain implants extending humans' mental and decision-making capabilities – key features of Transhumans. 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.



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

2. 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

3. AI Maturing Framework

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 (nurture AI as a child)
3. Learning Human values from the experience of other AI agents

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 progress 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 one day a Superintelligence. There are a number of 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 then would we further minimize misunderstanding? 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

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

1. In the first one, Superintelligence would work closely with Humanity to re-define those values, while being still under the total control by humans
2. The second option, and I am afraid more likely, assumes that once a benevolent Superintelligence achieves the Technological Singularity stage. At such a moment in time, it will increase its intelligence exponentially, and in a few weeks, it would 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 values that 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 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 an ultimate goal of a controlling human. It is the subject open for a long debate about how such an AI behaviour can be controlled, and how it would impact the working and goals of those Agents, it 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 the future Agency.

4. 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 already see humanoid robots in various roles more frequently. They will become assistants in a GP's surgery, 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 such AI agents have 'graduated' from such a school would they be ready to serve in a community. They will then communicate their unusual

experience of applying values in real environment 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, may 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. At the moment, 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 region 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.

Tony Czarnecki
Managing Partner
Sustensis

19 Raglan Court
London CR2 6N2
Email: tony.czarnecki@sustensis.co.uk
Website: www.sustensis.co.uk

London, 28th May 2020