Harnessing AI risk: Transforming our Greatest Threat into Humanity's Triumph. (v.1)

This document is a verbatim PDF copy (link) of a Linkedin article and blog post published June 28th, 2023 by the Trustless Computing Association, and authored by its president Rufo Guerreschi with the support of the association's advisors. It describes our open call for a critical mass of states to build new intergovernmental organizations to steer the IT and AI revolutions to the benefit of all humanity, and our foreseen role in it via the Trustless Computing Certification Body and Seevik Net Initiative. We'd gladly receive your comments, likes and shares on the original Linkedin article, as we plan to publish a version 2 of by the end of the Summer 2023.

ABSTRACT

A wide consensus is quickly arising among top AI experts and CEOs, world leaders and citizens alike that AI poses not only a fatal threat to democracy but also a significant human safety and existential risk for humanity on the horizon of a few years or decades.

The EU is drafting its AI Act to mitigate AI threats to democracy, with some accountability and transparency measures. At the same time, top AI experts and CEOs propose to tackle the existential risk by creating new intergovernmental organizations similar to the ones that helped mitigate the nuclear risk so far, the IAEA and the CERN.

Such measures, while needed, essential and encouraging, barely scratch the surface.

The AI Act is likely to turn out very insufficient and delayed, just as the threats of surveillance and social media to democracy have barely been mitigated, seven years after the Cambridge Analytica scandal and ten after Snowden revelations. An IAEA would be a great step to limit proliferation, and it's the right model to start, but far from enough for AI.

Akin to the Baruch Plan proposed by the US in 1946 for nuclear weapons and energy, we must enact a much deeper international cooperation by holding in Geneva an intergovernmental constituent assembly for new inter-governmental organizations for IT and AI that are sufficiently empowered, wide-scoped, competent, multilateral and participatory to truly tackle the risks, be widely trusted, and adequately mitigate the new threats they'd inevitably introduce.

Introduction

The launch of OpenAI's ChatGPT in November 2022, and rapid AI developments since then, have awakened the world to the fact that we stepped into an AI Age, just as we did in 1945 for the Nuclear Age, calling humanity to contend with it.

Leading AI firms and states, and smaller teams leveraging open-source tools, are in a race that has gained a blistering pace as hundreds of billions are being poured into it, and those AIs are being used for their own improvement.

These unforeseen developments have given rise to two main concerns that have rightfully taken center stage in public discourse: the fatal risks of AI for democracy and for human safety.

The first is a fatal **risk for democracy**, as the unregulated deployment of those AIs is bound to have a devastating impact on misinformation, elections, civil rights, hacking of sensitive systems, and further concentrating power and wealth within and among nations.

This risk has been well-known and understood since the 2016 US elections when state and non-state entities scaled up their use of AI, chatbots, troll farms, and deepfakes to influence the democratic process unduly or illegally. Almost nothing has been done about it.

The growing accessibility to those entities and the public of ever more sophisticated AIs is sure to turbo-charge those activities, expanding even more the power and number of entities able to manipulate elections and hack elected officials, journalists and diplomats at scale.

Left unchecked, this will inevitably result in further entrenchment of global power and wealth in a handful of tech firms, billionaires and states via their control of AI and human communications.

The second is a fatal **risk for human safety**. The accelerating advancement and self-improvement of these or future AIs not only pose immediate risks in their application to biological and nuclear weapons, but conviction is quickly growing among AI experts, leaders, top CEOs, and people that, if we follow the current course, we may give rise to Superintelligence.

This term refers to Artificial Super Intelligence (ASI) that - regardless of whether they'll achieve consciousness or sentience - would most likely slip out of human control, eluding the constraints and objectives it was programmed with.

A main concern is that such an entity could determine its own instrumental sub-objectives, such as self-preservation, self-improvement, or resource acquisition, possibly leading to an unprecedented scenario where it will possess the ability to wipe out all of humanity and exercise it for unknown reasons or to permanently prevent humans from switching it off.

An Unprecedented Opportunity for the Betterment of Humanity?

While these concerns are very well founded, it is vital to recognize that this and other looming catastrophic risks also present an unprecedented opportunity for the betterment of humanity, rivaling in positive transformational potential the decade after World War II.

One year after the creation of the United Nations, it became clear it could not prevent the spreading of nuclear and bioweapons expertise and capabilities, posing an unbearable risk of catastrophe.

So, the US proposed to the UN members its Baruch Plan, which would have mandated all members to transfer control of all their nuclear weapons arsenals and materials to a new single UN agency, which would then have a global exclusivity in research, development and management of nuclear weapons and energy.

Refusal of such a plan by the Soviet Union led to the nuclear arms race, the Cold War, and only ten years later to fall back solutions to mitigate the spreading of nuclear and biological weapons via the creation International Atomic Energy Agency (IAEA) and the Organisation for the Prohibition of Chemical Weapons (OPCW), and other treaties.

Today, almost eighty years later, in the face of the acceleration and proliferation of a new catastrophically dangerous technology, we have a second chance to tame and steer powerful technologies for the benefit of humanity by finally **extending the democratic principle to the global level** - starting from the all-important domains of Artificial Intelligence and human communications - to establish a solid foundation for long-term human safety, dramatically reduce wealth and power disparities, and turn scientific progress to uplift all of humanity.

Shortcomings of Current Proposed Solutions

The **European Union** is drafting an *AI Act* to mitigate the AI threat to democracy. In summary, it classifies AI use cases by risk level, banning some of them. It requires a pre-launch self-assessment of conformity to to-be-determined standards for high-risk use cases, with some audit capability by regulators.

As for EU privacy and social legislation, it is light years away from adequately mitigating the risk that the combination of AI and social media poses to EU democracy.

Meanwhile, **top AI experts and CEOs** propose to tackle the existential risk, by creating a new intergovernmental organization akin to the one that helped mitigate the nuclear risk so far, the IAEA and CERN. Those are good models to start from, but they are insufficient and with many shortcomings.

On either the democracy or human safety risks, not much is moving in the **United States**, aside from a tiny standardization initiative by NIST, early policy discussions among legislators, and Biden's bland declarations following his encounters with top US AI CEOs and with his UK counterpart to discuss national and global regulation of AI safety risks.

Meanwhile, **China** is approving regulations to bring top AI firms under governmental control, re-establishing dialogue with the US (about AI as well?). Still, it has yet to publicly endorse or propose any wide-scoped international or global initiatives.

The IAEA model: advantages and shortcomings

Advantages. IAEA is the most fitting model, and it would be great if we had an IAEA for AI today or very soon. But most likely, we need something more and different from IAEA, as also suggested by the leading world nuclear scientists' association.

- The IAEA ensured that 70 years later, only a few nations have nuclear weapons and no major nuclear conflict or nuclear weapon accident has occurred, crucially complemented by a worldwide mass and targeted surveillance of intelligence agencies of leading nations, mainly the US, Russia, China and Israel.

- The IAEA was created as an autonomous IGO, though reporting to the UN General Assembly and Security Council, and so, therefore, did not inherit some of the governance limitations of those.

- The IAEA statute mandates its members "to accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world." In practice, in exchange for the non-proliferation of nuclear weapons, signing states that did not have nuclear weapons received a promise to receive access to IP and materials for good uses of nuclear energy, and avoid falling backward in foreseen productivity gains due to super cheap energy, albeit those never really materialize.

- The IAEA statute prescribes that, of the 32 seats of the Board, ten are reserved for the more "knowledgeable" states and ensure that they "represent a stipulated geographic diversity."

Shortcomings. Both the nature of an IGO to reliably enforce a ban on dangerous AI and the constituent processes leading up to it should be substantially different and better than IAEA because:

1. **More transparency in the role of state security agencies**. The essential role of powerful nations' intelligence and security agencies of powerful nations has not been recognized and regulated in the IAEA statute and reconciled with civil freedom and democracy. This has resulted in a huge surveillance apparatus maintained by those superpowers at the expense of personal liberties, the sovereignty of nations and democracy, as even prime ministers, foreign ministers and diplomats are unduly

surveilled by who knows who via broken-by-design IT systems, preventing IT from fostering high-bandwidth, efficient and fair dialogue, as we need it the most.

2. **More globally representative in governance and in their creation process**. The creation and stature of the IAEA were negotiated in 1955–1957 by a self-selected group of twelve countries. Sure, you cannot have 200 states negotiating. Still, there should be a higher number, and most importantly, they should be more globally representative while giving a moderately higher decision-making weight to states that are more knowledgeable and powerful. The UN Secretary-General recently stated that an AI regulating IGO should be "*inspired by what the international agency of atomic energy is today*," implying its creation process and early stages were not a good model.

3. **More experts and independent experts**. In addition to experts from state security agencies and Big AI firms, surely needed for unique competence and insight into current techs and future ones, we need more independent experts involved.

4. **More multilateral governance and constituent process**. Needs to avoid too much power in the governments and leading firms of the US and China, as well as Israel (which announced a highly strategic and timely agreement with Nvidia just a few weeks ago). A constituent assembly and process for such a body should be held not in New York as for the IAEA but in Geneva or another nation that can highly guarantee geopolitical neutrality and in-person and digital confidentiality of the highly sensitive negotiations. Negotiations should be powered by highly secure and multilaterally-accountable (while convenient and user-friendly) means for digital diplomatic communication systems and platforms to ensure high-bandwidth, efficient and fair negotiations, which are absent today as we read every day that even prime ministers are continually hacked. Withdrawal by a participating state should come at a very high cost. Funding should not be according to traditional UN principles, which give too much indirect influence to nations with very large GDPs, especially when combined with the ability to pull out at no or low cost, but more evenly distributed.

5. **More citizen participation in governance and constituent processes**. Given the stakes, it is critical to have more citizen participation via citizen assemblies (even OpenAI supports that) and citizen-representative independent civil society (such as large member-funded consumer associations). A great example is the Dartmouth Conference, organized yearly for decades by Russia and the US, where random-sampled citizens were brought together to discuss nuclear risk.

6. **More powers, independence and transparency in breach assessment**. Insufficient levels of all three likely lead to several severe shortcomings. Failure to prevent some states from breaking it, including some to build nuclear arsenals after its establishment, such as North Korea, China, Israel, Pakistan and India, and to prevent Iran from getting very close to it. Failure to prevent the US government via Colin Powell from convincing the UN Security Council in 2003 that Iraq had nuclear weapons of mass destruction,

based on skimpy and faulty intelligence. Failure to enable states to withdraw while claiming another state breached it without clear evidence.

7. **More integration with better non-proliferation treaties** (NPTs). Calls for "NPTs for AI," in addition to an "IAEA for AI," are appropriate. Yet, it is best if constituent processes (and resulting organizations) for "NPTs of AI" and "IAEA of AI" are unified as they are highly interrelated and interdependent. NPTs were not effective enough to prevent cross-accusations and withdrawal by US and Russia from most NPTs with minimal geopolitical cost, as each was able to claim the other had breached it without a clear external adjudication by an entity defined by the NPTs of who was right or wrong.

The CERN model: advantages and shortcomings

The CERN successfully brought together nations to pursue shared research and knowledge transfer on the safety of nuclear weapons and energy production programs and infrastructures, and on advancements in nuclear energy generation.

Yet, a "CERN for AI" should be improved with respect to that model because (1) only EU states creating it and members (except some non-EU associates); (2) greatly fell behind state-of-the-art of technologies with respect to power nations; (3) its wide use of open source licensing and open science is fitting maximizing innovation in AI safety measures, but for AI capabilities it should be replaced with a *translucent* model, whereby transparency of systems code and hardware design, for accountability and security, is reconciled with the needs of non-proliferation.

Towards a Reliable and Durable Harnessing of IT and AI for the Benefit of Current and Future Generations.

Given the inability of a single nation, the UN, the EU and other IGOs to even come close to tackling the challenges and opportunities at hand, and their inherent global nature, the only real solution is the creation of new IGOs that are up to the task.

Even without the initial participation of AI superpowers, a critical mass of states could and should create new inter-governmental organizations (IGOs) to steer the development and use of AI and IT in humanity's best interest.

While it could be a single IGO, we envision three IGOs under a common governance framework, each serving crucial complementary functions, and providing some checks and balances deriving from separation of powers.

The first IGO will accrue a **decisive global *superiority*** in AI capability and safety measures. The second IGO will enforce a **global ban on *dangerous* AI**. The third IGO will build **ultra-secure and transparent IT certifications and systems** to enable effective, high-bandwidth diplomatic and citizens' communications and radically more trustworthy control subsystems to be mandated for all non-banned high-risk AIs.

More in detail:

The first IGO will accrue all advanced AI capabilities, IP assets, researchers and proportionate funding from member states and firms into a *Global AI Lab,* to achieve decisive and sustainable global *supremacy* in safe and human-controllable AI and advanced AI safety measures.

- Similarly to the Baruch Plan, such IGO will be endowed by member states with all their *advanced* AI capabilities, assets, intellectual property and talents, both from government and military, to develop capabilities in both AI safety and "human-controllable" AI capability, that sustainably far out-competing any other entity.

- Such a globally-distributed *Global AI Lab* will implement the most extreme, transparent and multilateral security and safety measures to sufficiently mitigate the risk of causing Superintelligence, accidentally or via leaks. Staff selection and oversight will exceed those of the most critical nuclear and bio-labs facilities. It will be located in a neutral nation or on a large nuclear-powered boat that moves between world ports to increase perceived trust.

- It will be tasked to develop the most advanced AI safety measures and only safe and *Human-Controllable AI* (or "HCAI"), a new term we introduce and define here as "*AI systems and deployments whereby the risk of human extinction or severe catastrophe, due to loss of human control or rogue human control over the next 100 years year, is estimated to be less than 1 in 1,000,000, that of a supervolcano eruption, or other similarly miniscule to-be-agreed upon probability*".

    - Member states and their firms will have equal rights on the resulting innovations and infrastructures for their governmental use and commercial exploitation, giving states a strong economic reason to join, in addition to preventing Superintelligence.

    - *Human-Controllable AI* will not be able to find a cure to all cancers or all technical solutions to climate change, and other wonders, as promised by OpenAI that the (purportedly safe) Superintelligence that they are pursuing to build would, if they don't lose control over it.

    - Nonetheless, HCAI will still be able to produce astounding amounts of innovation in robotics, self-driving cars, virtual assistants, generative AIs, health, and much more in nearly every domain. In fact, between 1929, when the FAA introduced very strict safety requirements, and 1938, the number of passengers in US civilian aviation skyrocketed from 6,000 passengers to 1.2 million.

    - Also, when and **if** advances in AI safety measures and the safety of new advanced AI architectures warrant it (as proposed by Goertzel in 2011), the allowed **capabilities may be expanded up until even the unleashment of a**

Superintelligence that will have high enough changes to be durably beneficial to humanity, wise and possibly conscious. An Enlightened AI. A sort of AI God.

    - A strong reason to pursue this option is that the all of three of the leading AI firms, Open AI, Google DeepMind and Meta's Chief AI Scientist have declared recently that, while acknowledging the risks, they consider it better for humanity to pursue the creation of safe Superintelligence, while OpenAI has already given its availability to join a initiative to build it a global initiative.

- Key to enabling such IGO to achieve and retain a decisive AI *superiority* in advanced AI safety and *Human-controllable AI* - especially if the US, China and their leading AI firms do not wish to join as members - is the ability of such IGO to attract and retain top AI talent and experts. Talent attraction in AI is driven by compensation, social recognition and alignment with the employer's mission. So, therefore:

    - Staff will be paid double their current global market value, and their social importance will be highlighted.

    - Member states will be mandated to: enact strong economic and social incentives to retain and repatriate their top AI and IT security; approve laws that will make it a crime to work for AI initiatives (abroad) deemed dangerous, and possibly enact a sort of "conscription."

    - Expectedly, such a Lab will be perceived by most top global AI researchers in non-member states as an initiative that is ethically superior to others by states or private firms, akin to how Open AI attracted top talent primarily due to aiming for "open-source AI."

The second IGO will enforce a ban on all development, training, deployment and research of dangerous AI outside of such a Global AI Lab, to radically mitigate the risk of dangerous abuse of AI by rogue or irresponsible state or non-state actors.

- It will extend such a ban to AI *advanced* researchers, and all sizable data centers, the global surveillance apparatus that has prevented terrorists, rogue states and irresponsible scientists from abusing at-scale weapons of mass destruction, and easily coordinate via strong encryption.

- It will coordinate intelligence and law enforcement agencies of participating and associate states and IGOs (e.g., Interpol, Europol, 5 Eyes, 9 Eyes, 14 Eyes, Club of Berne etc.) that have been absolutely fundamental to enable the IAEA and OPCW to succeed to date in preventing major catastrophes.

- The creation of this IGO, an "IAEA for AI," has recently garnered support from leading AI scientists, OpenAI, and the UN secretary-general.

- AI researchers will not be treated as criminals but will operate under heightened scrutiny, akin to the oversight endured by nuclear, bioweapons, and encryption scientists.

- States or firms developing dangerous AI or other harmful science will face severe embargos, or even surgical joint military action akin to those enacted on facilities of rogue states engaging in dangerous nuclear activities, as recently suggested by Eliezer Youdkowsky on Time.

The third IGO will build a new IT security certification and governance body, and initial compliant IT systems that will ensure radically more secure, democratic and accountable IT systems for human communications and for AI control subsystems, to enable fair, effective, high-bandwidth and "multi-track" diplomacy, as well as to increase the trustworthiness of inherently hyper-complex advanced human-controllable AI "black boxes."

- It will create and govern a new IT security certification body for human communications and for the most critical subsystems of society-critical systems - and an initial complete set of IT compliant with it - that aim to achieve levels of actual and perceived confidentiality, integrity and democratic/lawfulness accountability, that are substantially or radically beyond state-of-the-art - while ensuring national and international *legitimate* lawful access, via in-person procedural mechanisms[1].

- For AI control subsystems, such IT and certifications will be mandated for all internal and compliance-related critical subsystems of providers of *society-critical* advanced hyper-complex AI "black boxes," like, for example, AI chatbots like ChatGPT or AI running the feeds and control systems of dominant social media platforms.

    - They'll be mandated to adhere to technical and socio-technical security, transparency, and interpretability requirements, that combine the highest military security standards with the highest levels of transparency and public inspectability.

    - Such requirements, standards and certifications will apply to all integrity- and confidentiality-critical control subsystems, both internal and external to their digital infrastructures, including:

        - Firmware upgrades, security monitoring, and compliance monitoring systems.

        - Pre-deployment Controls (e.g., Adversarial Testing, Red Teaming, Automated Validation of Updates), Runtime Controls (e.g., Safelists and Blocklists, Real-time Monitoring, Supervised System) and

[1] In this 2018 academic paper, titled Position Paper: Case for a Trustless Computing Certification Body, we detail how new standards and certifications to validate both ultra-secure IT and the (in-person, procedural) legitimate lawful mechanisms can be created, in such a way that results in compliant IT services and systems that overall reduce the risk of privacy abuse of its users by anyone to levels that are radically (or at least substantially) lower than any alternative secure IT systems – commercially available today or knowingly under development – which do or do not offer such voluntary processing.

Post-deployment Controls (Feedback Loop, Regular Audits, Ongoing Learning) of core "black box" components of advanced AIs and LLMs.

    - Systems that will automatically scan and flag user interaction logs for egregious illegal activities, which will be as auditable as possible in an anonymized manner, and with limited data retention (as it is done today in most western democracies anyhow).

- For human and diplomatic communications, such IT and certifications will provide much higher levels of confidentiality and integrity for the sensitive communications of scientists, diplomats, heads of state, journalists, C-level executives of leading AI firms, and other influential persons, while ensuring their accountability.

    - It will build multi-national cloud and ultra-thin barebone mobile devices - carried in custom leather wallets or embedded in the back of any smartphone for ease of use and wide adoption - compliant with such certifications.

    - It will initially be available for diplomats, heads of state, scientists, activists and journalists to enable them to engage in the *high-bandwidth, fair, "unbiased," and effective digital diplomacy* crucially needed to swiftly but cautiously build and administer those IGOs. (Exemplificatory of this need is the fact that the Ambassador of Liechtenstein to the UN stated that Covid delayed 2 years the approval of the UN Veto Initiative, evidently because remote digital communications (in 2020!) were not able to sustain key diplomatic negotiations).

    - This is critical to maximize the democratic character, consensus, competency, and cautiousness of the constituent processes leading up to those IGOs, and their governance processes, both in-person and digitally.

    - The constituent process should be conceived to maximize a final statute and membership of the IGO that maximizes competency, representativity of nations and citizens, demographic diversity, wisdom, and altruism, and that is trusted by a wide majority of nations and states.

    - Given its marginal cost at-scale of under CHF 200 for the client devices, this infrastructure will eventually be accessible to all citizens to constitute *the first open global democratic digital communication infrastructure* that reconciles public safety with democracy and civil rights, that will gradually replace the current dominant hyper-complex, ad-based, private and insecure-by-design social media, messaging apps and client devices, unilaterally controlled by two states.

- Given the current availability of open, ultra-secure and battle-tested low-level IT designs, wide and dispersed global expertise about them, and the very low target performance levels of such IT systems, and novel chip foundry oversight methods, it will cost below CHF 100 million, and take less than two years.

- Another reason for that is that there is no need of magic new untested IT security innovations, but proper uncorrupted engineering that includes for its critical components only **open, battle-tested, redundant endpoint security, VPN, and encryption**, both classic and post-quantum, where the risk of supply chain attacks by powerful intelligence agencies, at root, at birth, is radically mitigated through extreme mitigation measure down to the chip fabrication.

All states will be free to join these IGOs on equal terms.

In essence, to steer the AI and IT revolution to the benefit of humanity, **we don't need a new or deeper global surveillance apparatus: we just need to deeply democratize the existing one** with the dual benefit of restoring and fostering democracy, promoting innovation and saving us from unbearable risks to human safety. We need to make it accountable to all global citizens and states, and reconcile it with civil rights and democratic principles.

Short Term Actions

Under the best-case scenarios, it will take a few years for each of those IGOs to be established and become operational, because suitable socio-technical standards and infrastructure will need to be built. Urgency must be combined with the need to get them right, especially their governance.

Yet, shortly after it is established, the 3rd IGO - while more stringent technical standards and IT for human communications are being built - could and should implement an initial set of regulations for all providers of AI and social media services to citizens that much stronger than those of the EU or any other nation, to protect and foster democracy.

To start, providers would be mandated to:

- Offer users to **pay a recurrent fee for their service** in alternative to ad-supported free service, if any, as a default option. Under such plans, users should be granted full and user-friendly control over the algorithms that determine their social media feeds, and those that filter or transform the replies they receive from the prompts they submit to generative AIs.

- Require all existing and new users to **confirm their identities** definitively, similar to the verification process employed by banks or government agencies, while allowing the use of pseudonyms.

- Mandate users to disclose **unique identifiers for third-party content produced by AI or other humans,** if any, included in their social media posts, and unique identifiers of its source, with strong legal liabilities for non-compliance.

Precedents

There are precedents of small states and NGOs playing a pivotal role in initiating or steering the creation of successful IGOs. The World Federalist Movement led dozens of NGOs in a Coalition for the International Criminal Court to participate and foster the ratification of such a court. Liechtenstein led over 80 nations to approve the UN Veto Initiative.

The ICC's creation was jump-started by an initiative by Trinidad and Tobago to create an international court for the illicit drugs trade. Major precedents for the wide inter-governmental creation of major technological infrastructure include International Thermonuclear Experimental Reactor (ITER), the International Space Station (ISS) and the Human Genome Project (HGP).

Even if it fails, the proposed initiative will provide leverage for micro, mini, small and medium states that are not AI superpowers, to be used on the table of the negotiations for such new "IAEA for AI" to avoid the current AI and cyber superpowers to set the rules to their perceived advantage.

Why are constituent processes so important?

Creating the IGOs we propose will **introduce numerous and substantial risks**.

They could fail to prevent Superintelligence from being created by some entity. They could turn into a sort of global dictatorship or oligarchy. They could result in a major nuclear conflict, as some powerful nation may oppose them or pursue the creation of more unilaterally or bilaterally orchestrated IGOs with the same aim.

Much is to be done to mitigate those risks, but the key factor will be the actual and publicly-perceived quality of the constituent processes leading to their creation. That is the highest guarantee that those risks will be sufficiently mitigated.

Why are constituent processes so important?

Well, the errors or autocratic mechanisms in governance structures - that are set in place instrumentally to build a new democratic entity - may not be fixable at a later stage. As Martin Buber put it, "*One cannot in the nature of things expect a little tree that has been turned into a club to put forth leaves*."

Let's take OpenAI as an example.

Last week, the CEO of OpenAI, Sam Altman, was asked at minute 20.54 of an interview, "you have an incredible amount of power. Why should we trust you?".

"*You shouldn't. ... No one person should be trusted here*", Altman replied, "*I don't have super-voting shares. ... The board can fire me, I think that's important. I think the board over time should get, like, democratized to all of humanity. There's many ways that could be implemented. But the reason for our structure is so weird ... is [that] we think this technology, the benefits, the access to it, the governance of it, belongs to humanity as a whole. You should not trust one company and certainly not one person.*"

"*So, why should we trust OpenAI? Are you saying that we shouldn't*?" the interviewer objected. After a long pause, Altman replied, "*No, I think you should trust OpenAI, but only if OpenAI is doing these sorts of things. If we are years down the road and we have not figured out how to start democratizing control, then I think you shouldn't*".

Altam's intentions to globally democratize what may become the most powerful AI entity in the world is very admirable and may be in good faith, it is "*years down the road,*" and, most importantly, it is not up to him, but to the majority of the seven persons that make up the board of the non-profit arm of OpenAI which controls the group, where he just hold one of 7 votes.

Such a board includes members that are not currently prevented from having direct or indirect conflicts of interest, most of which have not stated their ideas about future OpenAI governance matters, and whose collective backgrounds are definitely not globally-representative. One new member since May 2021, who was a clandestine CIA agent, who turned then Republican politician, called William Hurd, who last June 22nd announced that he was seeking the Republican nomination for president of the United States in the 2024 election.

Given AI's central importance to the future of humanity, and the power these IGOs will have, the constituent processes of such IGOs should be conceived from the beginning with the assumption that it may become a **blueprint for wider transnational democratic federal IGOs** (a UN 2.0?). We need those anyhow to successfully tackle other catastrophic and existential risks, improve justice and wellbeing of all humans, and realize great positive future scenarios for humanity.

Openness to join on equal terms for late-comer states, and the mandatory provision in those IGOs statute for **regular re-constituent assemblies every 10 years or so**, should reduce the risk of late-comers feeling their are joining an entity that does not have their values and interests incorporated, and enable to fix clauses on their statutes that will emerge to be faulty or outdated in respect to societal and technological change.

A good prospect is that a good number or **the majority of the few hundreds of ultra-rich persons that own most of the wealth and power** - who have rarely in the past modified their investment behavior in respect to climate risk - would likely be as interested as all of us to radically mitigate the existential risk for humanity posed by AI, because it would affect them and their kids, just as much as everyone else.

Next Steps

Our challenge is aptly summarized in one sentence by Stephen Hawking, "*Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all*."

A few leading pioneering nations, including possibly Switzerland, should convene next Fall in Geneva an intergovernmental *Conference on the Creation of Intergovernmental Organizations*

for AI and Human Communications, on dates different from those of the announced London AI Summit led by the UK and US.

Its objective will be to determine *Rules for the Election and Administration* of a to-be-convened *Constituent Assembly for Intergovernmental Organizations for AI and Human Communications*.

Given what's at stake, it should be a constituent assembly reminiscent of the 1946 United Nations Conference on International Organization, when 850 delegates from 50 states and 1800 civil society and media members gathered for two months in San Francisco to discuss and approve the UN Charter.

Geneva, and possibly Switzerland, could play a key role. Geneva hosts the UN International Telecommunication Union and over 200 diplomatic missions of states to the UN agencies. Switzerland has historical leadership in IT security for the most sensitive communications and is home to the CERN.

Geneva would not only be a perfect candidate to host such a conference but also to host the 3rd IGO mentioned above, which will create and regulate such a new *open global democratic digital communications infrastructure*.

The Trustless Computing Association, which I am honored to lead, has been engaging over 15 globally-diverse states interested in creating the third of the IGOs mentioned above, for ultra-secure IT for communications and AI critical subsystems, currently named Trustless Computing Certification Body and Seevik Net Initiative.

We'll be discussing such a prospect during our 12th Edition of Free and Safe in Cyberspace, to be held in Geneva next September 20th and 21st, 2023.

In conclusion, we stand at a crossroads in the face of AI and digital communications - one where we can either succumb to fear or seize the opportunity to create a much safer, fairer world.

We firmly believe that, by having the courage to stare boldly and unflinchingly at these risks in their enormous scale, we will acquire the insight, determination and clarity to come together to stand up to the greatest challenge humanity has ever faced.

Join us!