Harnessing AI Risk Initiative

Overview

Version of May 9th, 2024

An ongoing effort to aggregate a critical mass of globally diverse states to design and jump-start an **open, expert and participatory constituent process for the creation of a new global intergovernmental organization for AI and digital communications** that is suitable to reliably manage their immense risks in terms of human safety and concentration of power and wealth, and realize their potential to usher us in an era of unimagined prosperity, safety and well-being.

Table of Contents

You can click on the Contents below to move to specific sections.

1. Introduction

(the text below is the same text as the Initiative's web page as of May 7th, 2024)

The convergence of a shocking acceleration of AI innovation and unregulated digital communications has brought us to what may be the most critical juncture in human history.

We can still turn the dawn of this new era into the greatest opportunity for humanity, but only if we come together globally like never before to govern its immense risks and opportunities.

The *Harnessing AI Risk Initiative* aims to catalyze the creation of a new global democratic organization that will ensure that AI will turn out to be humanity's greatest invention rather than its worst.

We are facilitating a**n open treaty-making process for AI** based on the model of the **intergovernmental constituent assembly** and on other time-tested innovative democratic processes and technologies.

We believe such approach is the most likely to result in a treaty-organization that will sustainably avoid AI's immense risks, for human safety and concentration of power and wealth, as well as realize AI's potential to usher us in an era of unimagined well-being for all.

Unlike all other ongoing AI treaty-making initiatives, ours is be based on history's most successful and democratic treaty-making model: that of the **intergovernmental constituent assembly** that started with two US states convening the Annapolis convention in 1786 and culminated with the federal constitution of the United States.

Together with a wide network of experts and NGOs, we are aggregating a critical mass of states - through a series of structured summits in Geneva - to design and jump-start a similar process by agreeing on the *Mandate and Rules for the Election* of an *Open Transnational Constituent Assembly for AI and Digital Communications.*

The Assembly will be mandated to draft a treaty for a new organization to develop, regulate, and jointly exploit the most advanced safe AI technologies, while reliably banning unsafe ones.

The Assembly will be guided by the principle of *subsidiarity*, where control rested to the most localised level possible, from global to state, community, and individual.

The Assembly will aim to maximise expertise, timeliness, and agility, while also emphasising participation, democratic processes, impartiality, and inclusivity, to ensure that the resulting treaty-organisation will be widely trusted to:

- Encourage broad compliance with future bans and oversight
- Enhance safety through diversity and transparency in setting standards
- Ensure a fair and safe distribution of power and wealth
- Mitigate destructive inter-state competition and global military instability

Given the inherently global nature of AI's primary threats and opportunities, the mandate of the *Assembly* will include the following:

- Setting global AI safety, security and privacy standards
- Enforcing global bans for unsafe AI development and use
- Developing world-leading or co-leading safe AI capabilities via a public-private $15+ billion Global Public Benefit AI Lab and supply chain
- Developing globally-trusted governance-support systems

A sweeping AI treaty has been called for by hundreds of AI expert signatories of the AI treaty, by the UN Secretary General, by Open AI and explored in fine detail by Google Deepmind. Sam Altman even suggested last March the US Constitutional Convention of 1787 as a "platonic ideal" treaty-making model for AI.

Despite its ambitious nature, we are optimistic that even the AI superpowers, the US and China, will eventually join the Initiative due to a few compelling reasons. Firstly, preventing the proliferation of catastrophically dangerous AI will be much more challenging than nuclear weapons. Secondly, the immense abundance that AI is almost certain to deliver, if the risks are properly addressed, reduces the incentives for competition. Thirdly, 77% of US voters support a comprehensive international treaty for AI. Lastly, how can the US be against an initiative that will replicate verbatim, globally and for AI, that the democratic process which led to its constitution?

A Better Treaty-Making Method

Regrettably, the prevailing approach to treaty-making - characterized by unanimous non-binding statements, and unstructured summits largely co-opted by a few powerful states - has proven to be both undemocratic and inefficient, as evidenced by the outcomes in areas like climate change and nuclear disarmament.

To address this, the Initiative will adopt, specifically for AI, what is widely considered **the most successful and democratic model of intergovernmental treaty-making in history**. This model began with two U.S. states calling a meeting of three additional states at the Annapolis Convention of 1786, which subsequently led to the adoption of a federal constitution at the U.S. Constitutional Convention of 1787 through a simple majority vote. This constitution required ratification by at least nine states, eventually receiving unanimous approval from all 13 states in 1789.

Given the significant disparities in AI capabilities, global power, and literacy rates—with over three billion people either illiterate or without internet access—the *Open Transnational Constituent Assembly for AI and Digital Communication*s will apply vote a weighting based primarily on population size and GDP.

This mirrors the early U.S. Constitution and the ancient Athenian democracy, when only one in eight adult residents were initially eligible to vote. Yet, the mandate of the *Assembly* will ensure that nearly all citizens in participating states achieve literacy and internet connectivity within a specified timeframe, and so progressively reduce the influence of GDP to zero.

Twenty percent of the Assembly's delegates will be elected directly by citizens of participating states through uniform electoral processes, while five percent will be selected by random sample.

The US and China, as global and AI superpowers, are welcome to join at any stage, yet their participation will be held in suspension until the other one also joins. Early-joining states and superpowers will receive significant temporary economic and voting advantages.

Strategic Positioning

The Initiative seeks to fill the wide gaps in global representation and democratic participation left by global AI governance and infrastructure initiatives by leading states, IGOs and firms - including the US, China, the EU, the UN and OpenAI's public-private "trillion AI plan" - and become the platform for their convergence.

The Initiative aims to become the critical enabler of the UN Secretary-General's call for an "IAEA for AI." It aims to build a treaty-making vehicle with the global legitimacy and representativity that is needed, and his office, agencies and boards are lacking - in line with his clarification that "only member states can create it, not the Secretariat of the United Nations." The Initiative will eventually constitute a Caucus within the UN General Assembly and later seek approval by the UN General Assembly to become a part of the UN system while retaining full governance autonomy.

As in 1946, when the US and Russia proposed a new independent UN agency to manage all nuclear weapons stockpiles and weapons and energy research via their Baruch and Gromyko Plans but disagreed, we now have a second chance with AI. We can harness AI's risk to turn it into an unimagined blessing for humanity and set a governance model for other dangerous technologies and global challenges.

Preliminary Designs and Scope of the new IGO

The Initiative is advancing a proof-of-concept proposal for the scope, functions, and character of a new intergovernmental organisation that matches the scale and nature of the challenge, with unique levels of detail and comprehensiveness and the support of dozens of advisors and experts.

We group the required functions in three agencies of a single IGO, subject to a federal, neutral, participatory, democratic, resilient, transparent and decentralised governance structure with effective checks and balances:

- (1) An **AI Safety Agency** will set global safety standards and enforce a ban on all development, training, deployment and research of dangerous AI worldwide to sufficiently mitigate the risk of loss of control or severe abuse by irresponsible or malicious state or non-state entities.
- (2) A **Global Public Benefit AI Lab** will be a $15+ billion, open, partly-decentralised, democratically-governed joint-venture of states and suitable tech firms aimed at achieving and sustaining solid global leadership or co-leadership in human-controllable AI capability, technical alignment research and AI safety measures.
    - It will accrue member states' capabilities and resources and distribute dividends and control to member states and directly to their citizens while stimulating and safeguarding private initiative for innovation and oversight.
    - It will be primarily funded via project finance, buttressed by pre-licensing and pre-commercial procurement from participating states and firms.
    - It will seek to achieve and sustain a resilient “mutual dependency” in its wider AI supply chain - vis-a-vis AI superpowers and other future consortia - through joint investments, diplomacy, trade relations and strategic industrial assets of participant states.

- (3) An **IT Security Agency** will develop and certify radically more trustworthy and widely trusted AI governance-support systems, particularly for confidential and diplomatic communications, control subsystems for frontier AIs and other critical societal infrastructure, such as social media.

Far from being a fixed blueprint, such a proposal aims to fill a glaring gap in the availability of detailed and comprehensive proposals. It aims to stimulate the production of other similarly comprehensive proposals to foster concrete, cogent, transparent, efficient, and timely negotiations among nations leading up to such Assembly and eventually arrive soon at single-text procedure negotiation based on majority and supermajority rule, rather than unanimity.

Momentum and Roadmap

Through our collaborative efforts, we have successfully onboarded 32 world-class experts and advisors to the Association and the Initiative. Additionally, over 39 world-class experts, policymakers, and 13 NGOs are participating in our upcoming Summit.

In March 2024, our organisation conducted high-level consultations with the United Nations missions in Geneva from four states. These meetings included three heads of mission - ambassadors - and specialists in artificial intelligence and digital technologies. These states, located in Africa and South America, collectively represent a population of 120 million, have a Gross Domestic Product (GDP) of $1.4 trillion, and manage sovereign wealth funds amounting to $130 billion. We are currently engaging with three additional delegations.

In early April 2024, we received formal correspondence expressing interest from the Ambassador to the United Nations in Geneva, representing one of the largest regional intergovernmental organisations encompassing dozens of member states. Since December, we have extensively discussed with three of the top five AI Laboratories regarding their participation in the Global Public Interest AI Lab.

On April 23rd, 2024, we launched the Coalition for the Harnessing AI Risk Initiative. This launch led to the creation of an Open Call, an invitation extended to all individuals and organisations to participate, collaborate, and share their expertise.

We plan to host bilateral and multilateral meetings in Geneva in May and June with states, intergovernmental organisations (IGOs), and AI labs. These meetings will coincide with the United Nations International Telecommunication Union World Summit on the Information Society (UN ITU WSIS), scheduled for June 10-13th, and the United Nations AI for Good

Summit, set for May 25-29. Additionally, we will hold our Pre-Summit Virtual Conference on June 12th, leading to our debut 1st Harnessing AI Risk Summit, this November in Geneva.

Learning from History's Greatest Treaty-Making Success

Nine years after the ratification of the U.S. Articles of Confederation in 1781, it became evident to many states that these measures were insufficient to adequately protect their economic interests and security. Therefore, in 1786, two states initiated the Annapolis Convention by inviting three others to participate. This meeting laid the foundation for the U.S. Constitutional Convention of 1787, which established a robust federation.

During the Constitutional Convention, state delegations reached a consensus through a simple majority vote on a draft of the U.S. Constitution. This constitution was set to be ratified if endorsed by the legislatures of at least nine of the thirteen states. In retrospect, this process marked a significant achievement, even though only one in eight adults had voting rights.

Given the historical context and the success of this approach, a similar strategy should be adopted at the global level for Artificial Intelligence (AI). AI represents a pivotal technology with far-reaching consequences for the economy, safety, security, and the very fabric of human existence. Suppose we can initially bring together seven or more globally diverse states. In that case, it will pave the way to easily engage additional nations in a "Global Annapolis Convention for AI" and ensure its success.

More inFormation

For more information on the Initiative and the *Global Public Interest AI Lab*, please review this 90+ page Overview document below, which can be easily navigated via a clickable Table of Contents.

2. The Immense Risks of AI

The acceleration and proliferation of AI capabilities in the next few years and decades pose two main classes of risks, that of **extreme unaccountable concentration of power and wealth** and that of **immense human safety harm** deriving from misuse or accidents, and from loss of human control.

Frontier AIs are expected to keep expanding their capabilities five- to ten-fold annually. And that's based on growth in investments and computing power alone, without accounting for AI's increasing ability to self-improve and multiply the productivity of its developers. Meanwhile, ever more powerful AI technologies are leaked, stolen, released or reverse-engineered in reusable and modifiable formats.

Since hundreds of AI **scientists**, including two of the top three, stated last March that "*Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war*," awareness of AI risks has been mounting. Twenty eight **states**, accruing to 80% of the world population, recognized in the Bletchley Declaration such safety risks, including "loss of control". Over 55% of **citizens** surveyed in 12 countries were "fairly" or "very" worried about "loss of control over AI ". At an invitation-only Yale CEO Summit last June, 42% of **CEO**s surveyed said they believed AI has the potential to "destroy humanity within the next five to 10 years."

On the risks of concentration of power and wealth, the prime minister of Israel Benjamin Netanyahu - not widely known for communist tendencies - recently stated*: "You have these trillion-dollar [AI] companies that are produced overnight, and they concentrate enormous wealth and power with a smaller and smaller number of people. …That will create a bigger and bigger distance between the haves and the have-nots, and that's another thing that causes tremendous instability in our world. And I don't know if you have an idea of how you overcome that?"*..

Investments in AI and AI infrastructure are exploding. If successful, OpenAI's public-private **$7 trillion AI plan** to aggregate states, funders, chip makers and power providers will either (a) create an entrenched dominant global oligopoly under US control or else (b) possibly become the seed of a safe and democratic global governance of AI that Altman has been consistently calling for - as we argue in this article in The Yuan.

Meanwhile, seven years after the Cambridge Analytica scandal and ten after the Snowden revelations, **social media and sensitive communications** are ever more vulnerable to abuse and control by unaccountable entities, stifling fair and effective dialogue, within and among nations, at a time when it is most needed.

Timelines of Safety Risks

The timelines of safety risks are subject to high uncertainty, but some of the greatest world experts believe those could be just years away.

Yoshua Bengio, recipient of the prestigious Turing Award, states that loss of control to rogue AI systems could occur in as little as a **few years** unless appropriate precautions are taken. Anthropic's CEO Dario Amodei believes that the chance of an AI-induced civilizational catastrophe is "somewhere **between 10-25%**", while Amodei has also testified that he believes dangerous AI systems will be created soon: he claimed that systems that enable biological weapons are likely to emerge **within two to three years**. Google Deepmind's CEO Hassabis believes AGI it "could be just a **few years**, maybe a decade away."

The principle of precaution, necessarily given the immense risks of loss of human control, mandates that those warnings be taken very seriously.

Relative Urgency and Win-wins

Although there is near-universal agreement about those risks, **opinions on their relative importance and urgency diverge sharply**. Acrimonious discussions and confrontations have ensued in a moment when unity is literally vital for humanity.

The silver lining is that the **solution for mitigating these diverse risks may be largely the same**: establishing robust intergovernmental organizations designed to manage both the risks and opportunities presented by AI.

Seriously tackling the safety risks requires empowered global institutions as does tackling the risks of concentration of power and wealth. States, in fact, are unlikely to join or comply with global safety bans and oversight unless the benefits and power of AI will be sufficiently shared.

The dynamics are not dissimilar to those that caused many nations to adhere and comply with the nuclear proliferation constraints of the IAEA on the expectation of access to the know-how needed to access the bounties of nuclear energy.

3. The Astounding Opportunities of AI

If we succeed in durably tackling those immense risks of loss of control and concentration of power, we'll have access to extremely capable AIs to advance our own well-being, largely controlled by us collectively and individually according to the *subsidiarity principle*.

Such AIs have the potential to improve human life enormously, not only in practical areas like hunger and poverty eradication, improved health, scientific innovation, cheap clean energy and increased leisure through efficiency, but also in dramatically enhancing human social, mental and emotional well-being.

If applied to our digital and social communications with controls and incentives aimed at promoting users' wellness rather than advancing the goals of political and commercial advertisers, AIs could dramatically improve communications and relationships among persons, groups and nations, promoting peace within and across nations, communities and families.

AIs could greatly improve our ability to master and transform unpleasant mental and emotional habits and patterns, as taught by spiritual traditions. Envision everyone having access a 24/7 personal assistant, advisor, butler, chief of staff, coach, doctor, wellness consultant, and spiritual guide that - interacting with trusted human ones and personal data - provides tailored guidance that surpasses even that of the best and wisest human experts, fostering significant growth in emotional intelligence, happiness and self-awareness.

These benefits could compound over time, radically raising the average happiness and wisdom of humans, advancing dramatically towards realizing humanity's greatest potential, as pursued by our kindest, wisest and most enlightened ancestors, and possibly igniting an ever-expanding happiness movement realizing humanity's greatest potential.

AI's dual potential role in materially enriching life, advancing safety and peace, and fostering better relational and inner well-being can herald an unprecedented and unimagined whole new era of true human advancement.

The potential “AI pie,” if we avoid the immense risks, is so enormous that rich states and people can get richer while the poor can be much better off. But success inevitably requires a fair distribution of the power in shaping our collective future in this Digital and AI Age.

As in 1946, when the US and Russia, with their Baruch and Gromyko Plans, proposed a new independent UN agency to manage all nuclear weapons stockpiles and weapons and energy research but failed to agree, **we now have a second chance with AI**. We can harness AI's risk to turn it into an unimagined blessing for humanity and set a governance model for other dangerous technologies and global challenges.

4. Lessons from the Baruch Plan

As we are facing the enormous challenge of AI, many do's and don'ts can be learned from the ways states managed the risks of nuclear weapons and the promises of nuclear energy, as they suddenly emerged after World War II.

A year after the nuclear detonation over Hiroshima, the United States put forth the Baruch Plan, aiming to accrue the research, development, and management of nuclear weapons and energy within a new UN agency. This included the exclusive stockpiling of weapons and materials.

An even more ambitious Russian proposal, the Gromyko Plan, came just 5 days later, adding to such a proposal a total disarmament. Neither plan was approved, nor a convergence among the two was found in the UN Security Council. **Intelligence agencies** filled the void left by such a diplomatic and political failure via intense coordination to curb proliferation, complemented only in 1957 by institutions like the International Atomic Energy Agency (IAEA, in a secondary role, and only after all 5 members of the UN Security Council had achieved nuclear weapons capability.

While we are still alive and avoided major accidents, many near-disasters ensued. Russia and the US went on researching ever more powerful weapons, arriving 6 years later to test thermonuclear bombs 10,000 times more powerful than Hiroshima. The **nuclear risk has increased**, with more nations acquiring nuclear weapons, nuclear treaties failing one after the other, an ongoing nuclear arms race, and further loss of trust following the Iraq WMD and other scandals.

The depth and granularity of nuclear treaties and their technical and organizational compliance mechanisms **proved insufficiently trustworthy and widely trusted** to detect breaches, prevent breaches and enable the global public to ascertain if a party breached them.

Over the decades, entities like CERN, the Nuclear Energy Agency, and the International Thermonuclear Experimental Reactor successfully facilitated, to a moderate extent, the **sharing of nuclear know-how and capabilities** among nations, only after nuclear energy capabilities had turned out to be less economically transformative than initially expected.

Faced with the fast emergence of AI as a new enormously potent and dangerous technology, whose capabilities are expected to grow 5 or 10-fold per year in the next few years, we should learn from the Baruch and Gromyko Plans, pursuing an even more ambitious new global governance architecture for AI.

While it will be central to develop extremely trustworthy and widely trusted AI safety standards and certifications for assessing dangerous AI development and use, we direly need **better compliance, control and digital communications systems to enable them**.

We'll need to rapidly develop and oversee new standards and certifications for ultra-secure IT systems and socio-technical systems - along with an initial suite of systems that comply with them - aimed to achieve a significant leap in their actual and perceived levels of security, privacy, safety, and democratic-accountability - for two primary uses:

- (1) **control subsystems for human-controllable advanced AIs** and other critical societal systems, to enhance the safety, measurability, accountability, and trustworthiness of inherently hyper-complex, advanced AI "black boxes"; and
- (2) **client and server endpoints for sensitive and diplomatic human communications**, to facilitate fair, effective, high-bandwidth, multi-track, and democratic diplomacy and global cooperation.

Today, almost eighty years after the Baruch Plan, we have a **second chance** with AI to tame and steer extremely powerful technologies for the benefit of humanity[1]. This time, hopefully, we'll avoid half-baked fall-back solutions, "kicking the can" and hoping for the best, as we did with nuclear, and in a more **timely, inclusive and participatory** way, ensuring that all nations and peoples are involved and all benefit.

5. Calls by AI labs for Democratic Global Governance of AI

Several of the leading AI labs, their CEOs and top AI scientists - more keenly aware than citizens and heads of state of the immense risks for safety and concentration of power - have called clearly for strong and democratic global governance of AI and in some cases for using the model of the **global constituent assembly**.

Google DeepMind published last July a detailed "exploration" of the feasibility of creating four new IGOs for AI, including a Frontier AI Collaborative, an "*international public-private partnership*" to "*develop and distribute cutting- edge AI systems, or to ensure such technologies are accessible to a broad international coalition*". Its CEO stated in February he sees in the next few years their governance merging into a UN-like organization as we get closer to "AGI".

As mentioned above, we are literally taking **OpenAI**'s CEO Sam Altman at its word - and holding him to it! - when he **called for a global constituent assembly akin to the U.S.**

[1] As analyzed in 2021 in the paper International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons by Prof. Waqar Zaidi and Allan Dafoe, President of Centre for the Governance of AI, and currently Head of AGI Strategy and Governance at Google DeepMind.

Constitutional Convention of 1787 to establish a federal intergovernmental organization to manage AI, in a decentralized and participatory way, according to the *subsidiarity principle*.

Far from an extemporaneous statement, it was largely confirmed in later video interviews yet pushed "down the road". He stated that control over OpenAI and advanced AI should eventually be distributed among all citizens of the world. He stated that "*we shouldn't trust*" OpenAI unless its board "***years down the road will not have sort of figured out how to start"*** transferring its power to "all of humanity""

He stated if humanity jointly decided that pursuing "AGI" was too dangerous, they would stop all "AGI" development ("We'd respect that"). After OpenAI's governance crisis, he repeated that people shouldn't trust OpenAI unless it democratizes its governance. He then repeated that all of humanity should be shaping the future of AI. On February 24th, OpenAI stated in its revised mission, "*We want the benefits of, access to, and* ***governance*** *of AGI to be widely and fairly shared*."

Given the acceleration in AI capabilities, investment and concentration in recent months, and OpenAI's proposal of a public-private "***$7 trillion AI supply chain plan*****",** we believe that his pledge to transfer such power "*years down the road*" sounds more and more like an empty promise, unless they are turned very soon into precise timelines and modalities for the transfer of power to humanity. Yet, as he appropriately stated at the *World Government Summit*, "it is not up to them" to define such constituent processes, so he called on states, such as the UAE, to convene a Summit aimed at the creation of an "IAEA for AI."

Anthropic's CEO Dario Amodei suggested (7-minutes onwards in this video) that solving the *technical half* of the AI alignment problem would be of no use unless the global *governance half* is also solved and that eventually some global body should be in charge of all advanced AIs companies. It has experimented with Collective Constitutional AI to enable (national) citizens' assemblies to determine the values and constraints of AI, a process that could be extended to world citizens and to states.

OpenAI's Chief Scientist **Ilya Sutzkever** stated, "*it will be important that AGI is somehow built as a cooperation between multiple countries*." **Yoshua Bengio** called for a multilateral network of AI labs, analyzing in fine detail the right balance of global and national authority over them.

Hence, not only the Global Public Benefit AI Lab could attract many top AI talents based on its superior mission, but it could also attract close collaboration or full participation by some leading US AI Labs and other states. In addition, substantial **risks of near-term authoritarian political shifts in AI superpowers**, as warned (1.5 min video clip) by Yoshua Bengio, could further entice top US AI labs to "internationalize" their ventures to avoid the risk of falling largely or wholly under the control of an unreliable, undemocratic or authoritarian power in the near future.

While important and encouraging, those calls have diminished in recent months and do not tackle the all-important issue of the **nature, details, participation and timing of the constituent process** to arrive at such treaties that would overall most likely promote global public interest.

The Initiative also aligns with several open calls, such as for an AI Treaty, signed by many top AI scientists, to create both an "IAEA for AI" and a "CERN for AI", as well as those by The Elders and Future of Life Institute, and by Pope Francis, though none of them calls for a democratic constituent process.

6. The Constituent Process for a new IGO

Why Treaty-Making for AI is Broken

Regrettably, the **traditional approach to treaty-making**—characterized by unanimous non-binding statements, and unstructured summits largely co-opted by a few powerful states—has proven to be both undemocratic and inefficient, as evidenced by the outcomes in areas like climate change and nuclear disarmament.

Leading digital and AI **superpowers** appear locked in a reckless arms race - economic, military and geopolitical - over AI and AI chips, seemingly intent on hegemonizing it or, at best, eventually splitting its global dominance.

On their own, **nearly all states stand powerless in the face of AI**, unable to avoid its immense risks for safety, for the concentration of power and wealth, and unable to realize its astounding opportunities. Even larger states like Brazil, India and Germany, or large confederations like the EU. On their own, **nearly all nations lack the strategic autonomy**, on their own, to table more democratic constituent processes to safeguard their economy, sovereignty, and safety in such all-important domains.

Initiatives for the global governance of AI are apparently led by existing **Intergovernmental organizations and fora**, like the UN, G7, G20, the EU, Council of Europe, OECD, GPAI, and the AI Safety Summits.

Yet, these are structurally unable to lead a *democratic* or effective global constituent process for AI governance due to their lack of a mandate, lack of representativity, closed membership and/or statutory over-reliance on unanimity decision-making. Hence, their initiatives severely lack multilateralism, detail, timeliness, breadth, transparency and global inclusivity and are mostly controlled by a handful of states.

The **prevailing treaty-making models** are bound to result in **severely weak, fragile and undemocratic treaties** - as they did largely in past decades - due to their reliance on loose, undefined, unstructured processes, over-reliant on unanimity.

The real explanation is that those treaty-making initiatives are really smokescreen and distractions from the real negotiations. While intergovernmental organizations and fora engage in hopeless treaty-making processes for AI that are structurally weak, slow and undemocratic, **global governance of AI is really taking shape via competition and negotiations among superpowers** and their national security agencies - with an observer role for selected allies - as it did for all other disrupting technologies in the past.

While those superpowers managed to avoid the realization of the most catastrophic risks of nuclear and bioweapons (so far!), the risks of nuclear and bioweapons are today higher than they ever were and instituted an opaque surveillance apparatus that has undermined democracy worldwide. In addition, mitigating the **proliferation and safety risks of AI is likely to be much harder than nuclear**, and therefore a much wider global compliance, adherence, and participation to common safety rules will be necessary.

After the UK AI Safety Summit was convened to foster international cooperation on AI risks of misuse and loss of control, the **United States** and the UK each announced their own AI safety institutes instead. A month later, Guidelines for Secure AI System Development were published last November by the national security agencies of the **United States** and the UK, together with the cybersecurity standardization bodies of 16 allied states. Meanwhile, **China** announced its Global AI Governance Initiative, which calls for a "*United Nations framework to establish an international institution to govern AI*" that will "*ensure equal rights, equal opportunities, and equal rules for all countries in AI development and governance.*", but no action followed, except for some reported discussions with the US on AI.

While recent US announcements that it intends to "cooperate" with China on AI safety are very welcome news, albeit late, control of AI safety by a handful of states, as was done for nuclear after WW2, would not work for two main reasons. It would be unfair and undemocratic, leading to **extreme and unaccountable concentrations of power and wealth**. It would likely not work for safety either because the nuclear threat is higher today than it ever was and because preventing catastrophic AI proliferation will likely require a much wider global adoption and compliance.

While lacking so far in inclusion, transparency and democratic process, such US/UK initiative and US/China talks are highly welcome, given that so much of the relevant expertise accrues in their security agencies, and that several top AI experts think **catastrophic safety risks may be just years away**, including via leaked, stolen or published dangerous LLM weights.

Hence, we need a treaty-making model that reconciles global legitimacy, democracy, expertise and reckoning with the huge asymmetries of power and expertise in AI among states.

Towards a Better Treaty-Making Model

To address this, the Initiative will adopt, specifically for AI, what is widely considered **the most successful and democratic model of intergovernmental treaty-making in history**. This model began with two U.S. states calling a meeting of three additional states at the Annapolis Convention of 1786, which subsequently led to the adoption of a federal constitution at the U.S. Constitutional Convention of 1787 through a simple majority vote. This constitution required ratification by at least nine states, eventually receiving unanimous approval from all 13 states in 1789.

Hence, there is a historical role for a few pioneering states, NGOs and business leaders to play the role of convenors of such assembly, as they did in 1946 with the Baruch and Gromyko Plans, for nuclear weapons and fission energy; in the 80s with the International Thermonuclear Experimental Reactor (ITER) for nuclear fusion energy; in the 90s with the Coalition for the International Criminal Court for global criminal justice and **especially in 1786 with the convening of the US Annapolis Convention by two US states that led to the US federal constitution**.

We aim to act as a catalyst for several states, NGOs and firms, by launching a dedicated Harnessing AI Risk Summit to be held in November, 2024 in Geneva, Switzerland. It will be ideally preceded by preparatory meetings, in cities that are comparatively neutral and/or locus of suitable international diplomatic missions, such as Geneva, Singapore, Vienna, Brussels or Rome.

Learning from the Annapolis Convention

Nine years after the ratification of the U.S. Articles of Confederation in 1781, it became evident to many states that these measures were insufficient to adequately protect their economic interests and security. Therefore, in 1786, two states initiated the Annapolis Convention by inviting three others to participate. This meeting laid the foundation for the U.S. Constitutional Convention of 1787, which established a robust federation.

During the Constitutional Convention, state delegations reached a consensus through a simple majority vote on a draft of the U.S. Constitution. This constitution was set to be ratified if endorsed by the legislatures of at least nine of the thirteen states. In retrospect, this process marked a significant achievement, even though only one in eight adults had voting rights.

Given the historical context and the success of this approach, a similar strategy should be adopted at the global level for Artificial Intelligence (AI). AI represents a pivotal technology with far-reaching consequences for the economy, safety, security, and the very fabric of human existence. Suppose we can initially bring together seven or more globally diverse states. In that case, it will pave the way to easily engage additional nations in a "Global Annapolis Convention for AI” and ensure its success.

Maximizing the Constituent Assembly Quality

To maximize the participation, decrease risks of undue concentration of power, and avoid a thwarting of deliberations towards unanimous declarations - as typical of international IGOs/fora - the process will be framed as an open, **federal constituent assembly** for the establishment of a treaty for the creation of suitable IGOs, with the participation of states, neutral experts, representative civil society and global citizens' assemblies.

Therefore, the first concrete milestones of such a process will be an agreement on the process and mandate of such an assembly, agreed-upon *Scope and Rules for the Election of an Open Transnational Constituent Assembly for AI and Digital Communications* that will design, largely by majority and supermajority rule, the statutes of the needed consortium and intergovernmental organization.

The quality of the design of such constituent assembly is paramount in maximizing the chances that the resulting IGO will be durably in the global public interest.

On the risks side, while the majority of world citizens in nearly all nations are in support of new global democratic federal organizations, many leaders and citizens have legitimate **fears that**

such powerful new global institutions may fail to be, and durably remain, accountable to world citizens and overall beneficial to their safety, wellbeing and liberty.

On the opportunity side, advances in constitutional and constituent process science, detailed studies of previous national and intergovernmental constituent assemblies and constitutional conventions, and **advances in remote digital communication, deliberation and AI-enhanced language translation** point to great opportunities to be much better than their national or federal precedents. Such constituent process - and the resulting "treaty constitution" and organization - could be design to be much more participatory, resilient, fair, high-bandwidth accountable, expert and efficient than past similar experiences, and carefully embed in itself the seeds for its constant adaptation to changing contexts and self-improvement over time, such as through periodic review conventions.

Much can be learned from the analysis of **previous federal constituent assemblies and constitutional conventions** - such as those of the US, Germany, Switzerland and EU, as well as extensive post-WW2 positive non-federal experiences - to best mitigate such risks.

Given what's at stake and the urgency of the risks, such constituent processes should be both very careful and timely. This can be achieved by creating a very **high-bandwidth process**, such as by convening 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. This time however, it'll need to be a much more participatory process, not largely a ratification of previously agreed draft among the winners of World War II as it happened for the UN Charter.

Also, suitable **fair and secure diplomatic communications** infrastructure will be needed, as provided by the IT Security Agency, to enable confidential, accountable and pseudonymous communications for "Track I" negotiation, in addition to **high-bandwidth communications and AI-powered automated translations** for "Track II" negotiation among less critical non-state actors.

Extensive use of **citizens' assemblies** should be ensured to promote participation by informed citizens, in partial deliberative roles and not only consultative, learning from experiences such as the Global Citizens' Assembly for COP26, and many others deployed in and by many states.

In order to ensure global representativity of states that are not initially participants the assembly could include a sort of global **random-sampled former parliamentarians' assembly**. Members from all nations will be able to join and will need to be globally representative. The selection of such members will minimize their risk of being subject to corruption, threat, or blackmail by powerful entities. Members will be selected half from defense, interior, or intelligence oversight committees, and another half from privacy and civil rights committees. (Logistic and scientific collaboration will be sought with the Climate Parliament and Parliamentarians for Global Action). Members' vote will be weighted via scientific methods to maximize global representativity of all major differentiating human factors, such as gender, race, religion, political orientation, and it will be further weighted by 20% according to the size of the members' nations population.

The **locations** for such a constituent assembly, and the resulting organizations, should maximize neutrality and accessibility for state representatives. No state is completely neutral but Geneva and Singapore may be the best fits. Geneva is also the location to missions of 190 states to the UN in Geneva and home to many UN agencies on digital matters.

A **weighted voting** will need to be agreed upon to balance the representativity of more and less populous states, with a preference towards more proportional representation. Voting will be per country, except it will be weighted according to a to-be-determined coefficient based on population size, GDP per capita, and other metrics, as will membership fees. The global-representativity of participating nations will be maximized via the initial selection of nations and IGOs, and by weighting of the voting to maximize global-representativity in respect to political regimes, continents, population size, religion and other key determinants.

Recognizing that leading national security agencies, AI labs currently and a small set of NGOs accrue essential and rare expertise necessary to successfully regulate highly potent, complex, secretive and fast-moving leading-edge AI technologies, **highly globally-representative and neutral scientific advisory boards and committees** should be set up, ensure very extensive knowledge transfer, and have substantial power, both consultative and deliberative, to influence decisions, especially on the enforcement of bans and the setting of AI safety standards.

All states will be free to join the constituent process and the resulting IGO on equal terms, except **AI superpowers will need to join together** to avoid being perceived as controlled or influenced by one of them. Some of the AI or digital superpowers may decide to support the initiative even in its earlier stage. Yet, to maintain mutual trust and neutrality, it will be important that the two AI superpowers join together, when and if they will do so. It is also crucial that all states should be able to participate, regardless of their social and political systems, or geopolitical standing, including Israel and Iran, for example. Such constituent processes could go ahead even **without the initial participation of all or some AI superpowers**, hoping for future participation. If that does not happen, it'd still bring extensive benefits to participants and radically increase their negotiation power towards superpowers to better find a middle-ground solution.

It's crucial to ensure equitable representation of member states in governance structures. This should include mechanisms for civil society's direct participation in decision-making processes, with a rotation of key positions among member states and mandatory public consultations on major decisions.

Vote Weighting and Early Participants' Advantages

Given the significant disparities in AI capabilities, global power, and literacy rates—with over three billion people either illiterate or without internet access—the *Open Transnational Constituent Assembly for AI and Digital Communication*s will apply a vote weighting based primarily on population size and GDP.

This mirrors the early U.S. Constitution and the ancient Athenian democracy, when only one in adult residents was initially eligible to vote. Yet, the mandate of the *Assembly* will ensure that nearly all citizens in participating states achieve literacy and internet connectivity within a specified timeframe, and so progressively reduce the influence of GDP to zero.

Twenty percent of the Assembly's delegates will be elected directly by citizens of participating states through uniform electoral processes, while five percent will be selected by random sample.

The US and China, as global and AI superpowers, are welcome to join at any stage, yet their participation will be held in suspension until the other one also joins. Early-joining states and superpowers will receive significant temporary economic and voting advantages.

7. Scope and Functions of the new IGO

While the governance structure and statute of such a new IGO cannot be pre-designed, as they'll be the outcome for the Assembly, they are ultimately of paramount importance. They should be expected to be properly empowered and resourced and highly federal, neutral, resilient, participatory, democratic, decentralized with highly effective checks and balances, and safeguards from degeneration or undue concentration of power.

Why should we have a single IGO for AI? Given the unique nature and scope of the challenge ahead, it is unfitting to see explore models for global AI governance - such as an "IAEA for AI", an Intergovernmental Panel for Climate Change for AI, a CERN for AI, or an "International Civil Aviation Organization for AI", or a "global AI lab" - as alternative one to the other, as all of them are needed.

Each of such new IGOs for AI would have significant interdependencies, so the analysis of state's advantage in participating in one IGO must be assessed in relation to its participation in other IGOs. For example, the participation of states in the IAEA, with its commitments to non-proliferation and submittal stringent oversight, was largely achieved by offering access and technical support for harnessing advanced nuclear energy technology.

Why are comprehensive preliminary designs of such an IGO needed? While the statute and governance of such IGO will be the result of the specific design and context under which such an *Open Transnational Constituent Assembly* will be held, preliminary and **comprehensive** designs should be proposed and discussed to encourage substantive discussions and negotiations before and during such assembly. Given the absence of a comprehensive proposal, with the partial exception of the Deepmind one mentioned above, we have taken it upon ourselves to create one below.

Given the inherently global nature of those threats and opportunities, the scope of those organizations will necessarily need to include: (1) setting of globally-trusted AI safety standards; (2) development of world-leading safe AI and AI safety capabilities; (3) enforcement of global bans for unsafe AI development and use; and (4) development globally-trusted governance-support systems.

We group the required functions in three agencies of a single IGO:

- (1) An **AI Safety Agency** will set global safety standards and enforce a ban on all development, training, deployment and research of dangerous AI worldwide to sufficiently mitigate the risk of loss of control or severe abuse by irresponsible or malicious state or non-state entities.

- (2) A **Global Public Benefit AI Lab** will achieve and sustain a solid global decentralized leadership or co-leadership in human-controllable AI capability, technical alignment research and AI safety measures. It accrues capabilities and resources of member states and distributes dividends and control to member states and directly to its citizens, all the while stimulating and safeguarding private initiatives for innovation and oversight.

- (3) An **IT Security Agency** will develop and certify radically more trustworthy and widely-trusted AI governance-support systems, particularly for confidential and diplomatic communications, for control subsystems for frontier AIs and other critical societal infrastructure, such as social media.

Far from being a fixed blueprint, such a proposal aims to fill a glaring gap in the availability of detailed and comprehensive proposals. It aims to stimulate the production of other similarly comprehensive proposals to foster concrete, cogent, transparent, efficient and timely negotiations among nations, leading up to such an *Open Transnational Constituent Assembly for AI and Digital Communications,* and eventually arrive soon at *single-text procedure* negotiations.

8. Defining and Measuring Human-Controllable AI

The first responsibility of such an organization is to develop and maintain measures, methods and socio-technical systems to assess and measure the risk levels of advanced AI systems, services and technologies.

It will formulate measurable **definitions of a threshold for what constitutes a safe AI service, system or component**, i.e. one that is "human controllable." This definition should include its internal control subsystems for enforceable evaluation and benchmarks, physical access systems, and related external compliance control mechanisms.

Given that the risk of catastrophic outcomes due to accidental or malicious AI malfunctions can never be fully eradicated—unless we opt for a complete ban, thus sacrificing all potential benefits—we must decide on an **acceptable risk threshold** for AI systems. This could be quantified as a minuscule X% over a specified period of Y years, such as an yearly risk of 1 in 1'000'000 of a supervolcano eruption[2], along with reliable methods for measuring this risk.

Such thresholds will be different for different AI components based on their risk of becoming accessible to unauthorized, irresponsible or malicious actors as a consequence of having been leaked, stolen, released or reverse-engineered.

Such **definitions, evaluations, benchmarks, and enforcement mechanisms** have been identified by some leading AI companies and top researchers, and the Frontier Models Forum, as the most urgent regulatory requirements for current advanced AI models.

We currently lack the systems and socio-technical standards for IT control subsystems that are both sufficiently trustworthy and can be expected to be widely trusted worldwide. These systems are those that will be responsible for implementing compliance, security monitoring, firmware upgrades, and value systems functions, while also being resilient to errors, and internal and external hacks, by humans or AIs. Such a function will be filled by the *IT Security Agency*, detailed below.

Human-Controllable AIs, due to their constraints, may not be able to find a cure to all cancers or all technical solutions to climate change, or at least not initially. Nonetheless, it will still likely be able to produce astounding amounts of innovation in robotics, self-driving cars, virtual assistants, generative AIs, health, and more in main domains. In fact, between 1929, when the Federal Aviation Administration (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.

9. Tackling the Superintelligence Option

The design of governance of the proposed IGO should consider that it may be forced to make choices that will be incredibly impactful for the future of humanity.

Consider, in fact, that three leading US AI labs - Open AI, Google DeepMind and Anthropic - have all repeatedly declared they aim to build AI that surpasses human-level intelligence in all human tasks, without limits, realizing the so-called AGI or Superintelligence.

While recognizing the immense risks for safety, these firms are moving ahead in such pursuit anyhow because the ongoing race dynamics may be unstoppable; and because they each

[2] https://forum.effectivealtruism.org/posts/Z5KZ2cui8WDjyF6gJ/some-thoughts-on-toby-ord-s-existential-risk-estimates

claim that their specific technical approach *may* succeed better than others' in ensuring the *technical* alignment to prevent "loss of control" or to produce an outcome more beneficial for humanity compared to other similar initiatives.

While most of them publically agree that the most positive scenarios would be those retaining a wide human control over AI, many AI scientists privately consider it possible or likely that loss of human control over AI, so-called *takeover AI*, under some conditions, may be overall highly beneficial for humanity.

While this state of things is extremely unsettling, it must be assumed that it is also the intention of the US government, given that it has not stopped nor even questioned such publicly declared plans. This is likely due to a perceived AI arms race with China, confidence in backend safety guardianship of its national security agencies, or other motives.

Hence, under possible future scenarios, advances in frontier AI safety and alignment or an increased risk of other entities releasing more dangerous superintelligences may lead such IGOs to decide, after wide participation and pondering, that it is overall most beneficial for humanity to substantially relax their "loss of control" safety requirements or even intentionally unleash a Superintelligence.

10. Open Source, Translucency and Public Safety

The new organization will need to define its approach to the public availability of source designs of critical AI technologies. The latter can bring huge advantages and immense risks, depending on the circumstances, and needs to be carefully regulated, but is currently framed in the public debate, quite idiotically, as a binary “all open” or “all closed” choice.

A sensible approach to open source, we believe, will be to require it in nearly all critical software and hardware stacks of an AI system or service, as a complement to extremely trustworthy and transparent socio-technical systems and procedures around them.

Yet, open source is insufficient to ensure that the code is both sufficiently trustworthy and widely trusted by users, states and citizens. Therefore, all source codes of critical components will also be required to undergo an **extreme level of security review in relation to complexity**, performed by a diverse set of incentive-aligned experts.

Also, none of the current open source licenses (not even the GNU GPLV3 Affero License) requires that those running open source code on a server infrastructure - such as an AI lab providing its service via apps, web interface or API - to provide publicly a (sufficiently trustworthy) proof that the copy of the code downloaded by an end-user matches that which is being used. This needs to be ensured.

That said, there will be exceptions for components whose proliferation could cause very substantial safety risks, such as dangerously powerful LLM weights, which could not only be published, but also hacked or leaked.

The trustworthiness of such components, as well as the safety of their public availability, should be managed via a very carefully designed "translucent oversight process", similar to the national legislative committee tasked to review highly classified information, but in an intergovernmental fashion and with much more resilient safeguards for procedural transparency, democratic accountability and abuse prevention.

This translucent oversight process will aim to maximize effective and independent review of the source code by a selected and controlled set of expert delegates of states and independent ethical researchers to maximize actual and perceived trustworthiness by states and citizens.

Those and other requirements are described in the Trustless Computing Paradigms.

11. The Global Public Benefit AI Lab

The Global Public Benefit AI Lab (or "Lab") will be an open, democratically-governed joint-venture of states and AI labs aimed to achieve and sustain a solid global leadership or co-leadership in *human-controllable* AI capability, technical alignment research and AI safety measures. It will accrue capabilities and resources of member states and firms, and distribute dividends and control to member states and directly to their citizens, while stimulating and safeguarding private initiative for innovation and oversight.

At A Glance

The **Global Public Benefit AI Lab** will be a $15+ billion, open, partly-decentralized, democratically-governed joint-venture of states and suitable tech firms aimed to achieve and sustain a solid global leadership or co-leadership in *human-controllable* AI capability, technical alignment research and AI safety measures.

The *Lab* is one of three agencies of a new intergovernmental organization being built by the Harnessing AI Risk Initiative, a venture to catalyze a critical mass of globally-diverse states in a global constituent processes to build a new democratic IGO and joint venture to jointly build the most capable safe AI, and reliably ban unsafe ones - open to all states and firms to join on equal terms.

- The Lab will be an open, partly-decentralized, democratically-governed joint-venture of states and suitable tech firms aimed to achieve and sustain a solid global leadership or co-leadership in ***human-controllable* AI capability**, technical alignment research and AI safety measures.
- The Lab will accrue **capabilities and resources** of member states and private partners, and distribute dividends and control among member states and directly to their citizens, all the while stimulating and safeguarding private initiative for innovation and oversight.
- The Lab will be **primarily funded via *project finance***, buttressed by pre-licensing and pre-commercial procurement from participating states and client firms.
- The Lab will seek to achieve and **sustain a resilient "mutual dependency" in its wider supply chain** vis-a-vis superpowers and future public-private consortia, through joint investments, diplomacy, trade relations and strategic industrial assets of participant states - while remaining open to merge with them on equal terms, as detailed in our recent article on *The Yuan*.

Financial Viability and the Project Finance model

The Lab will generate revenue from governments, firms and citizens via licensing of enabling back-end services and IP, leasing of infrastructure, direct services, and issuance of compliance certifications.

Given that the proven scalability of capabilities, value-added and profit potential of current open source LLMs technologies - and the possibility of extensive pre-commercial procurements contracts with states could buttress its financial viability - the initial funding could follow primarily the project finance model, via **sovereign and pension funds, intergovernmental sovereign funds such as the EIB and AIB, sovereign private equity and private international finance**.

The undue influence on the governance of private funding sources will be limited via various mechanisms, including non-voting shares.

Precedents and Model

The initiative could take inspiration from the current governance of the CERN, a joint venture for nuclear energy capability-building that was started in 1954 by EU states and only later opened to non-EU ones, with a current yearly budget of $1.2 billion. The $20 billion international consortium ITER for nuclear fusion energy is also an inspiration.

Size of Initial Funding

Since the cost of state-of-the-art LLMs "training runs" are expected to grow 500-1000% per year, and many top US AI labs have announced billion-dollar LLM training runs for next year, the Lab would need an initial endowment of **at least $15 billion** to have a solid chance of achieving its capacity and safety goals, and then financial self-sustenance in 3-4 years. If such an amount seems high, consider it would likely increase by about 5-10 times for every year this initiative is delayed.

Supply-Chain Viability and Control

Acquiring and maintaining access to the specialized AI chips needed to efficiently run leading-edge LLM training runs will be challenging given a foreseen intense increase in global demand and export controls.

This is a risk that can likely be sufficiently reduced via joint diplomatic dialogue, appealing to the open and democratic nature of the initiative, and by attracting participating states hosting firms owning suitable AI chips designs, or possibly start pursuing its own AI chip designs, and chip manufacturing capabilities, and invest in new safer and more powerful AI software and hardware architectures, beyond large language models.

Ensuring sufficient energy sources, suitable data centers, and resilient network architecture among the member states, would require timely, speedy and coordinated action for the short term and careful planning for the long term.

Hence, the Lab will seek to achieve and **sustain a resilient “mutual dependency” in its wider supply chain** vis-a-vis superpowers and future public-private consortia, through joint investments, diplomacy, trade relations and strategic industrial assets of participant states - while remaining open to merge with them on equal terms, as detailed in our recent article on *The Yuan*.

Talent Attraction Feasibility

Key to achieving and retaining a decisive superiority in advanced AI capability and safety - especially if or until AI superpowers and their firms have not joined - is the ability to attract and retain top AI talent and experts. Talent attraction in AI is driven by compensation, social recognition and mission alignment and would need to ensure very high security and confidentiality.

Staff will be paid at their current global market value, and their social importance will be highlighted. Member states will be mandated to support top-level recruitment and to enact laws that ensure that knowledge gained is not leaked. Staff selection and oversight procedures will exceed those of the most critical nuclear and bio-labs facilities in sophistication.

The unique mission and democratic nature of the Lab would likely have a strong chance of being perceived by most top global AI researchers, even in non-member states, as being ethically superior to others, akin to how Open AI originally, and Meta more recently, have attracted top talent to work with them, or for them, via claims of their "open-source" ethos.

Just as OpenAI attracted top talent from Deepmind due to a mission and approach perceived as superior, and top talents from OpenAI went on to create Anthropic for the same reasons, the Lab should be able to attract top talents as the next "most ethical" AI project. Substantial risks of authoritarian political shifts in some AI superpowers, as warned (1.5 min video clip) by Yoshua Bengio, could entice top talents to join the Global AI Lab to avoid their work being instrumental to an authoritarian regime.

Public-Private Partnership Model

Participant AI labs would join as *innovation and go-to-market partners*, in a joint-venture or consortium controlled by the participant states.

They will contribute their skills, workforce and part of their IP in such a way as to **advance both their mission to benefit humanity, their stock valuations**, and retain their agency to innovate at the root and application level, within safety bounds:

- As *innovation partners* and IP providers, they would be compensated via revenue share, secured via long-term pre-licensing and pre-commercial procurement agreements from participating states and firms.
- As *go-to-market partners*, they would gain permanent access to the core AI/AGI capabilities, infrastructure, services and IP developed by the Lab.
    - These will near-certainly far outcompete all others in capabilities and safety, and be unique in actual and perceived trustworthiness of their safety and accountability.

    - They would maintain the freedom to innovate at both the base and application layers, and retain their ability to offer their services to states, firms and consumers, within some limits.
- Participant AI labs partnership terms will be designed so as to maximize the chances of a steady increase in their market valuation, in order to attract the participation of AI labs - such as Big Tech firms - that are governed by legal conventional US for-profit vehicles that legally mandate their CEOs to maximize shareholder value.

This setup will enable such labs to continue to innovate in capabilities and safety at the base and application layers but outside a “Wild West" race to the bottom among states and labs, advancing both mission and market valuation.

The Superintelligence Option

The design of governance of the proposed IGO should consider that it may be forced to make choices that will be incredibly impactful for the future of humanity.

Consider, in fact, that three leading US AI labs - Open AI, Google DeepMind and Anthropic - have all repeatedly declared they aim to build AI that surpasses human-level intelligence in all human tasks, without limits, realizing the so-called AGI or Superintelligence.

While recognizing the immense risks for safety, these firms are moving ahead in such pursuit anyhow because the ongoing race dynamics may be unstoppable; and because they each claim that their specific technical approach *may* succeed better than others' in ensuring the *technical* alignment to prevent "loss of control" or to produce an outcome more beneficial for humanity compared to other similar initiatives.

While most of them publically agree that the most positive scenarios would be those retaining a wide human control over AI, many AI scientists privately consider it possible or likely that loss of human control over AI, so-called *takeover AI*, under some conditions, may be overall highly beneficial for humanity.

While this state of things is extremely unsettling, it must be assumed that it is also the intention of the US government, given that it has not stopped nor even questioned such publicly declared plans. This is likely due to a perceived AI arms race with China, confidence in backend safety guardianship of its national security agencies, or other motives.

Hence, under possible future scenarios, advances in frontier AI safety and alignment or an increased risk of other entities releasing more dangerous superintelligences may lead such IGOs to decide, after wide participation and pondering, that it is overall most beneficial for humanity to substantially relax their "loss of control" safety requirements or even intentionally unleash a Superintelligence.

12. The AI Safety Agency

The AI Safety Agency will set global safety standards and enforce a worldwide ban on all development, training, deployment and research of dangerous AI, to sufficiently mitigate the risk of loss of control or severe abuse by irresponsible or malicious state or non-state entities.

Current Initiatives for international AI safety bodies

Last November, the UK *AI Safety Summit* convened leaders of 28 selected "key" states, including China, to seek international cooperation regarding fast-emerging AI risks of misuse and loss of control.

While all those states signed a declaration recognizing those risks, no new concrete and proportional international initiative or body was announced to globally assess, measure those risks, and set up international mechanisms to manage them worldwide. Surprisingly to many, the US and UK announced instead each their own national AI Safety Institutes.

One month later, on November 27th, the national security agencies of the US and UK, together with the cybersecurity standardization bodies of 16 other Western states, issued high-level but quite comprehensive Guidelines for Secure AI Development.

This is an overall welcome first step, as the national security agencies of leading Western states have come out to play publically in the domain of AI, the essential and critical role that they have played in containing the proliferation of other dangerous technologies that emerged over the last century, such as nuclear, strong encryption and bioweapons.

If nuclear weapons have not to date resulted in massive catastrophes, it is due more to their coordination across geopolitical blocks rather than to the *International Atomic Energy Agency*, which was established only in 1957 and which for decades had limited information sharing with national security agencies.

We have to be highly thankful to those national security agencies for their overall success in filling the gap left by the political and global coordination failures of Russia and the US - and the veto-holding members of the UN Security Council - in finding common ground in 1946 between the US's Baruch Plan and Russia's Gromyko Plan proposals to create a global intergovernmental agency to control all nuclear arsenals, fissile material, research and energy.

While it is desirable and expected that other superpowers will join in drafting a shared version of such *Guidelines*, and turn them into an international agency, such an approach to global governance of AI risks replaying the same errors made with nuclear in 1946.

Pursuing such a strategy risks producing even worse outcomes for AI than it did for nuclear. Firstly, while we are still alive, there were many near-misses of catastrophic nuclear events, and the risk of nuclear catastrophe is higher today than it ever was. Second, adjunct innovations in nuclear weapons, like AI and supersonic missiles, weaken the "secure second strike" key to nuclear stability.

A Better Approach

For these reasons, we call on all and each of the signers of such Guidelines to decide to turn them into a seed for creating an inclusive and participatory *Open International AI Safety Agency*, that will be effective, expert and timely, and accountable to the elected bodies of their nations and all nations.

Surely, coordinating 193 countries would be difficult, and many have low technical expertise in hugely complex, secretive and fast-moving technologies. Still, methods can be devised to reconcile the maximization of **expertise, timeliness and agility** on the one hand and **participation, democratic process, neutrality and inclusivity** on the other.

For example, a major effort could be promoted to speed up the technical knowledge of all nations. Critical decision-making bodies of such an Agency could include both a permanent representation of few expert nations, without a veto, and a larger number of less-expert nations, selected randomly or through election by the UN General Assembly, as it happens for non-veto-holders members of the UN Security Council.

We believe, in fact, that pursuing a much more globally participatory approach to the global governance of AI is crucial to ensure that the resulting organizations will be sufficiently trustworthy and widely trusted to:

- Encourage broad adoption and compliance with mandatory bans and oversights;
- Enhance safety through global diversity and transparency in safety standards;
- Achieve a fair and safe distribution of power and wealth; and
- Effectively mitigate the risks of global military instability.

No intergovernmental agencies created to set IT and AI security and safety standards have been involved in drafting or signing such *Guidelines*, not regional ones like ENISA, ETSI and

OECD, nor global ones, like ITU, ISO, IEC. This is not surprising if we consider that the setting of global IT standards that affect national security has been handled at the level of military alliance or their leader, such as NATO or NIST, with advice from the NSA.

The involvement of global international standard-setting organizations like ISO and IEC, globally trusted because they are largely governed similarly to the UN General Assembly, would also add actual and perceived trust to the process.

Given the enormously impactful decisions that such a body will be called to make, other participatory elements should be added, such as global citizens assemblies and an international body representing the world's major spiritual traditions and outlooks.

Hence, we call on the signatories of the Guidelines to convene in Geneva in an open summit with their counterparts from the rest of the world and representatives of their ministries of foreign affairs, to work towards the construction of such an open international AI Safety Agency.

Bans, Oversight, and Oversight of the Oversight

Such an agency will extend a global ban on all sizable data centers and advanced AI chips to train dangerous AIs. This will be enforced via the same kind of global surveillance apparatus and 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 prevented terrorists, rogue states and irresponsible scientists from abusing at-scale weapons of mass destruction, bio weapons or easily coordinate via strong encryption, and via the help of the IAEA and the OPCW.

This was done, however, at **very high costs in terms of privacy and democracy,** due to their inability so far to reconcile civil freedom and public safety via some kind of trustworthy front-door mechanism, which led to an unaccountable or poorly accounted all-encompassing surveillance and systematic weakening of all IT and IT security standards by leading security agencies.

A critical challenge will be the fact that this ban may well soon need to be extended to *advanced* AI researchers, unless the risk of leaks or open source releases of large powerful LLM "weights" cannot be otherwise successfully contained worldwide. In such a case, AI researchers worldwide will need to operate under heightened scrutiny, akin to the oversight endured by nuclear, bioweapons, and encryption advanced scientists for many decades.

We, therefore, will need to face a conundrum that has been left unsolved since the 90s, of finding ways to reconcile via win-win mechanisms the need for oversight and public safety, on the one hand, and the need for accountability and civil freedoms on the other.

This will require that we apply unprecedented levels of trustworthiness and accountability to oversight and compliance systems as well as to secure communications, by applying the same extreme, battle-tested technical, socio-technical and organizational safeguards, as detailed in our 2018 Position Paper on Trustless Computing Certification Body.

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

Considering the deep oversight requirements that will likely be needed to prevent, and those extremely extensive and unaccountable that already exist today - 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 much more accountable to all global citizens and states and reconcile it with civil rights and democratic principles. In this regard, radically more trustworthy, widely trusted and accountable IT socio-technical systems would be critical, as described in the IT Security Agency chapter below, to be applied to such oversight and surveillance systems.

13. The IT Security Agency

The IT Security Agency will be responsible for developing and certifying radically more secure and trusted IT "governance-support systems," particularly for confidential and diplomatic communications, as well as control subsystems for frontier AIs and other critical societal infrastructure, such as social media.

Need for "Trustless" Organizations and Technologies

While the majority of world citizens in all nations are in support of new global democratic federal organizations, with the notable exception of the UK and the US, many have legitimate fears that sweeping powerful new global institutions - such as those suggested in this text - may fail to be, and durably remain, accountable to world citizens and overall beneficial to their wellbeing and liberty.

In addition to designing proper constituent processes and statutes for such global organizations, mentioned above, the key to mitigating such risk will be to ensure extremely trustworthy safety requirements and enforcement mechanisms for the most advanced AIs. This requires **AI control and compliance sub-systems and human digital communications** are much more trustworthy and widely trusted in their safety, security, privacy and democratic accountability than they are today.

To ensure technical safety compliance mechanisms and constituent and governance processes that are sufficiently trustworthy and widely trusted by nations and citizens, the initiative also needs to develop suitable standards, protocols and technologies based on existing globally-supported, neutral, open-source, (ultra) high assurance, battle-tested systems, such as some derivatives of Risc-V chips and Sel4 operating systems, and certain open encryption protocols and algorithms.

This mainly requires, we believe, exceedingly transparent, resilient, trustworthy, democratic and decentralized approaches - in one word: "trustless" - to the engineering and assessment of their critical, compliance and control components, both organizational and technical.

This is needed to instill the needed trust, prevent abuses and accidents, avoid the lack of indisputable mechanisms for the assessment breaches that contributed to the failure of many nuclear treaties, and to mitigate the power of state and non-state actors to exploit such weaknesses to sway public opinion, political leaders and diplomats via their control of digital communications.

The availability of digital tools for sensitive communications and deliberations that are truly trustworthy and trusted in their integrity and confidentiality - including off-the-record, pseudonymous and anonymous communications, and together with tremendous advances in real-time translations and decentralized trust technologies - would be a game changer in the enabling and sustaining accountable digital diplomatic negotiations, global democratic constituent processes and global governance.

This agency will develop and oversee **new standards and certifications for ultra-secure IT**, along with an initial suite of systems that comply with these standards. The aim is to achieve a significant leap in both actual and perceived levels of security, privacy, safety, and democratic-accountability for technical and socio-technical IT systems and endpoint platforms - while ensuring legitimate lawful access, national and international.

These platforms will be used for two primary purposes: (1) **control subsystems for human-controllable Frontier AIs** and other critical societal systems, with the goal of enhancing the safety, measurability, accountability, and trustworthiness of inherently complex, advanced AI "black boxes"; and (2) **client and server endpoints for sensitive and diplomatic human communications**, to facilitate fair, effective, high-bandwidth, multi-track, and democratic diplomacy and global cooperation.

Security Problem of control subsystems for advanced AIs.

Minimized and ultra-secure technical and socio-technical IT systems and standards, referred to as high-assurance, are utilized today to maximize the security, privacy, safety and accountability of control subsystems of critical infrastructure that are society-critical and inherently complex, such as Frontier AIs and their development infrastructure.

While there has been an enormous increase in the foreseen global cost of failures of those systems due to accident, hacking or misuse, even top national security agencies have shown a failure to safeguard their most critical data, as shown by Shadow Brokers, OPM and Vault 7 hacks.

Security Problem of sensitive digital communications.

Heads of states, heads of opposition, ministers, elected officials, journalists and top scientists do not have access to interoperable computing devices and services that enable them to protect the confidentiality of their sensitive and off-the-record communications against innumerable state and non-state hacking entities - and so are forced to rely only or mostly on in-person meetings to further global cooperation and diplomatic initiatives.

Exemplification 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 (even in 2020) were not able to sustain key diplomatic negotiations.

Why a Joint End-point Platform

It is therefore imperative to make those systems both much more trustworthy and widely trusted by people and nations of the world, introducing a new class of ultra high-assurance systems and standards that we call trustless computing.

The solution of both those problems and use cases primarily requires a new modular endpoint platform for client and server-side use and access control systems that have much higher resistance to advanced state-grade technical and organizational hacking attempts based on open, battle-tested, globally supported, and neutral technologies.

We estimate it will be possible to have a single modular platform for both problems and uses cases, that satisfies performance and security and can be used both in (A) a small form factor and minimized endpoints, suitable for ultra-thin basebone mobile devices, as well as (B) to run efficiently as critical control and compliance sub-systems or servers, both alone or in parallel computing and cluster architectures.

Key IT Security Paradigms

Key to the solution of those joint problems is to ensure (1) higher transparency of technical designs and processes; (2) more expert and varied security reviews in relation to complexity, (3) trustworthy procedural legitimate lawful access mechanisms, and, most importantly, (4) higher global participation and neutrality in the standardization and certification governance processes.

- **Higher Transparency of Technical Designs and Processes**: This paradigm focuses on the need for openness in developing and implementing IT systems. It involves making the technical details, such as source code, architecture, and algorithms, available for scrutiny. The rationale is that transparency allows for broader community engagement, from independent researchers to other industry players, leading to the identification and rectification of potential vulnerabilities. It also builds trust among users and stakeholders, as they can verify the security and privacy features of the systems they rely on.

- **More Expert and Diverse Security Review in Relation to Complexity**: This approach emphasizes the importance of comprehensive and diverse security audits for IT systems, proportionate to their complexity. It suggests engaging a wide range of experts from different backgrounds and specializations to examine the systems, ensuring a thorough assessment from various perspectives. Such multifaceted reviews can uncover a broader range of potential issues, from technical vulnerabilities to broader systemic weaknesses, and propose more robust solutions.

- **Trustworthy Procedural Legitimate Lawful Access Mechanisms**: This principle addresses the need for lawful access to encrypted data and systems by authorized entities while ensuring accountability of the process. It involves creating clear, transparent, and well-regulated processes that allow legal access under strict conditions, ensuring a win-win solution between privacy, security, and legal enforcement needs. The challenge lies in designing mechanisms that sufficiently protect from unauthorized access and abuse, thereby maintaining public trust in digital systems against abuse on both sides of the fence.

- **Higher Participation and Neutrality in Standardization and Certification Processes**: The final paradigm calls for inclusive and impartial participation in the processes that set standards and certify IT systems and services. It advocates for a global approach where stakeholders from different regions and sectors have a voice in shaping the standards, ensuring that they are well-rounded, equitable, and applicable across different contexts. Neutrality is key to preventing the dominance of specific interests and ensuring that the standards serve the broader global community’s interests, thus enhancing the universal reliability and security of IT systems.

More details and inspirations can be found in the Trustless Computing Paradigms.

Control subsystems for advanced AIs: Seevik Controls

Such IT and certifications, called **Seevik Controls**, 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 global transparency, accountability 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 for example:

- Firmware upgrades, security monitoring, and compliance monitoring systems.
- AI-specific controls, such as:
  - *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)
  - *Post-deployment Controls* (Feedback Loop, Regular Audits, Ongoing Learning) of core "black box" components of advanced AIs and LLMs.
  - *Value Systems* (or "Constitutions" in *Constitutional AI* -based systems).
- Systems that will automatically scan and flag user interaction logs for severely 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).

Recent research by Anthropic and other researchers found that if deep deception capability (a sort of backdoor with preset trigger known to the backdooring entities) are somehow inserted during training, possibly as a result of a hack, there would be no way to remove them. Hence, critical training phases would need to be tracked and secured at the highest level against such threats.

Sensitive digital communications: Seevik Net

Such IT and certifications will provide much higher levels of confidentiality and integrity for the sensitive communications of scientists, diplomats, heads of state, journalists, internal sensitive communication of frontier AI firms, and other influential persons, while ensuring their accountability.

- 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.

- It will build multinational 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.
- 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, which is trusted by a wide majority of nations and states.
- There is no need for magic new insufficiently-tested IT security innovations, but proper uncorrupted good 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 - even at root and at birth by powerful intelligence agencies - is radically mitigated through extreme mitigation measures, down to the chip fabrication processes.
- Given its primary reliance on older, low-performance, battle-tested techs, and its marginal cost at-scale is estimated to be under USD 300 per client device, this infrastructure will eventually be widely commercially available to citizens constituting *the first open global democratic digital communication infrastructure for sensitive human computing* that reconciles public safety with democracy and civil rights. It would complement, and gradually partly 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, the very low target performance levels of such IT systems, and novel chip foundry oversight methods, the cost to initial usable services and infrastructure could be below USD 100 million, and take less than two years.

Legitimate Lawful Access

Seevik Net 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 accountability, that are substantially or radically beyond state-of-the-art - while ensuring national and international *legitimate* lawful access, national and international, via in-person procedural mechanisms.

While a group of highly influential and outspoken UK and US IT privacy experts disagree that the creation of any form of front door access mechanism can be done without further and gravely endangering privacy, we published in 2018 a very detailed and referenced academic paper titled Position Paper on Trustless Computing Certification Body[3] where we argue the opposite view.

As such paper states in its abstract: "*By applying the same safeguards used to ensure ultra-high security, and more, the inevitable added risk will be radically mitigated, resulting in compliant IT services that overall reduce the risk of abuse of end-users by anyone to levels that are radically (or at least substantially) lower than any of the other alternative secure IT systems – available today or knowingly in development – which do or do not offer such voluntary processing*."

Democratic Social Media

Problem

Seven years after the Cambridge Analytica scandal and ten after Snowden revelations, social media and sensitive communications are ever more vulnerable to abuse and control by unaccountable entities, stifling fair and effective dialogue, within and among nations, at a time when it is most needed to tackle challenges to democracy and human safety worldwide.

The digital media platforms that we all use, all day and every day, are led by a handful of tech super billionaires, and their ads buyers and temporary political allies, that deeply shape what we think is true and good, what products we buy, and what candidates to vote for.

Proper national or international media regulations could largely or completely solve these problems. Yet, their approval, and even mere proposal, has been and will be extremely difficult.

[3] https://www.researchgate.net/publication/325011381_Case_for_a_Trustless_Computing_Certification_Body_---_Can_a_new_certification_body_deliver_radically_unprecedented_IT_security_for_all_while_at_once_ensuring_legitimate_lawful_access

Such difficulty is due to the ever-stronger chokehold media powers have on the political process, their repeated threat of pulling out of regulating nations, and the inescapable need of nations to keep all IT systems hackable to fight terrorism. In 2019, Zuckerberg himself testified, saying "*I don't think private companies should make so many decisions alone when they touch on fundamental democratic values*", yet no meaningful regulations have even been tabled so far.

To counter these risks to democracy, nations can't do it alone as they lack the power to impose adequate regulations on semi-monopolistic US and Chinese social media, and now AIs, that their citizens and businesses have grown dependent on.

Governments' role in social media

Any call for a solid regulatory or provisioning role for democratic governments or a number of them is opposed by mainstream media, their owners and therefore most politicians, as an undemocratic and autocratic concentration of power, while we have ample proof that the contrary is the case when such a role is properly designed.

In Western liberal democracies, we've lived in a world that has glorified private media as a guardian of citizens' rights, and government intervention as something evil and authoritarian per se. This is a product of the private media's control of the accepted range of opinions and their crucial interest in promoting this opinion to protect their undue powers over public opinion. There is much truth in the media as guardians of democracy, but the contrary is also true: that "free" media is the primary means by which political and economic elites exert undue power over society.

Given the key societal role of media authors and publishers - from bloggers to newspapers to new app makers - as guardians against abuses from governments, the role of nations should be based on the same principles of *trustlessness* that govern our electoral system - whereby effective check and balances, transparency and citizens' oversight processes prevent the use of such power to stifle their independence - while also prompting their creativity to promote diversity, individual freedoms, and progress.

Democratic Social Media

A critical mass of states joining forces in this AI Safety Agency to set and enforce regulations for **social media and AI applications involved in human communications** - and build a new democratic Global Digital Public Infrastructure, that includes cloud and social media capabilities complaint to it - and so gather enough **leverage** to impose them on dominant semi-monopolistic social media and generative AI platforms.

Such Infrastructure would be not unlike the role of national public TV broadcasters in liberal democracies, designed to uphold all democratic values and rights, while ensuring a role for the private media sector as guardians of power and innovation actors.

Social media regulations in member countries would include, for example:

1. Service providers must require all users, both existing and new, to confirm their identities definitively, similar to the verification process employed by banks or government agencies, while allowing the safe use of pseudonyms.

2. Service providers must allow the 50% less wealthy users to choose and configure their own feed algorithms, free from any commercial or political advertising, while giving such an option as a cost proportional to their wealth to the other 50%.

3. Service providers must adhere to technical and socio-technical security, transparency, and interpretability requirements for their critical and compliance sub-systems, that combine the highest military security standards with the highest levels of transparency and public inspectability.

4. Service providers must require users to disclose unique identifiers for third-party content produced by humans or AI, and its source, included in their posts, if existent unless the content is entirely original - with legal liabilities (...)

This platform needs a new open transnational democratic governance body - uniquely citizen-accountable, competent and resilient - that will define, evolve and govern its technical standards for producers of technologies, and its terms of use for users, authors, publishers, and producers. It will exist in parallel with current platforms, and also include apps that run on mainstream mobile app stores.

All of its critical software and hardware technical layers, including apps and devices, will be based only on open-source or publicly inspectable source designs and will be subject to extreme security review in relation to complexity, and citizen-witness and citizen-jury-like oversight - all the way down to the CPU, chip foundry oversight, and data rooms access management.

Such a body will certify systems created by private providers, including clouds and mobile apps for mainstream mobile stores, for their security and their respect for basic rules for democratic media space.

As discussed above, the Agency will also certify dedicated ultra-secure client devices with a separate app store, for the confidentiality needs of the 1% most politically-exposed, such as diplomats, journalists, heads of state, and persons critically involved in the planned intergovernmental organization.

14. Unprecedented Opportunity for the Betterment of Humanity?

While being concerned and cautious is 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 the positive transformational potential that opened after World War II.

This is essential to stimulate energetic and lucid action by good-will states, citizens and NGOs, focusing not only on the avoidance of a terrible threat but also on the potential for achievement of a much better future.

As mentioned above, 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. The UN Security Council failed in 1946 to agree on Russian and US formal proposals to mandate 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.

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 harness scientific progress to uplift all of humanity.

If successful, it is conceivable and hoped that the resulting governance, constituent process and governance-support systems will become a model for wider IGOs to manage **other dangerous technologies and global challenges**, moving closer to proper global federal democratic organizations that finally realize the promise of the United Nations.

15. Opportunities for States

The Problem

On their own, **states and intergovernmental organizations stand powerless in the face of AI**, unable to avoid its immense risks for human safety and for concentration of power and wealth, and unable to realize its astounding opportunities.

This is the case even in larger states like Brazil, India and Germany, or large regional intergovernmental organizations like Europe.

On their own, **even superpowers stand unable to go it alone** as mitigating the proliferation and safety risks of AI are expected to be much harder than nuclear.

The Opportunity

The Trustless Computing Association (TCA) invites a critical mass of globally diverse states to join our Harnessing AI Risk Initiative and participate in our summits. We envision our assembly as a leading platform for engagement in bilateral and multilateral meetings, aiming to launch the AI treaty-making process. Such a multilateral and participatory treaty-making process will serve as a premise for creating a new open global intergovernmental organisation to jointly build and share the most capable safe AIs and reliably ban unsafe ones.

We invite States to:

- Join the Initiative as a State Member by endorsing a straightforward, non-binding Letter of Interest and contributing CHF 20,000. This engagement also allows us to collaborate as co-lead with one additional Inter-Governmental Organization (IGO), assuming shared oversight of the Harnessing AI Risk Initiative.
- For those interested but not yet prepared to commit, we offer participation in the 1st Harnessing AI Risk Summit this November or the Pre-Summit Virtual Conference on June 12th. These exclusive closed-door bilateral and multilateral meetings in Geneva or via videoconference present an opportunity to gain deeper insights into the Initiative's workings.

A Comprehensive Perspective on Advantages

Unlocking Overall Benefits:

- Enjoy exceptional economic development, sovereignty, safety, and civil rights advantages from the joint control and ownership of the Global Public Interest AI Lab. This ensures reliable long-term access and control over the most advanced safe AI systems for their governmental and private sector needs.
- Radically mitigate the immense risks to human safety and the highly unaccountable concentration of power and wealth posed by AI. Co-lead in shaping AI's ethics, limits, privacy, security, safety and accountability to realise its potential to bring astounding benefits to your citizens and humanity, in a mutually advantageous way for all.
- Increase leverage vis-a-vis other global governance and infrastructure initiatives for AI by leading states and firms.

Summit:

- Engage as a speaking participant in the Summit. “Observer status” or remote participation is also possible. Early state participants will have a guaranteed speaking slot on day one and will be more prominently displayed.
- Participate in co-designing the Initiative and, therefore, in the design of global governance of AI and the constituent process leading up to it. Early participants will be accorded privileged visibility.
- Learn about the current initiative and prospects of global AI governance, AI safety, and the Initiative.

Exclusive Benefits for Pioneers: the first 7 participant states will enjoy unique benefits:

- Economic advantages and discounts. Acquire a "priority option" to become one of a limited number of Founding State Partners of the Initiative. As such enjoy advantages and discounts concerning all fees, quotas, contributions of the first three years of Initiative and Lab, and their sovereign fund participation as Pre-seed Funders of the Lab:
    - 70% for the 1st to the 2nd state
    - 40% for the 3rd to the 4th state
    - 20% for the 5th to the 7th state
- Political prominence. Early participants will be more prominently included in the Summit, its documents, and webpages, and they will be guaranteed a speaking slot on day 1 of the Summit.

Contact us to discuss our collaboration:
partnerships@trustlesscomputing.org

16. Opportunities for Donors

When confronted with the ever-expanding realm of AI, individual states and intergovernmental organisations find themselves relatively powerless, confronting significant risks and challenges beyond national borders. These risks manifest in various forms, including but not limited to the erosion of privacy rights through pervasive surveillance technologies, exacerbating social inequalities due to unequal access to technological advancements, and consolidating power and wealth in the hands of tech giants and oligopolies.

At the same time, AI presents unparalleled opportunities for societal advancement, economic growth, and global connectivity, yet harnessing these opportunities necessitates cohesive efforts and collaborative strategies. Even for nations with considerable resources and influence, the complexity and scale of technological challenges persist. Recognising the interconnected nature of these issues is crucial to the global efforts required to address AI. It underscores the imperative for tightly-knit state and regional intergovernmental cooperation and coordinated action on a worldwide scale.

Trustless Computing Association (TCA) was initially self-funded and managed to achieve significant milestones. These accomplishments were made possible through the volunteer efforts of our advisors, participants, and selected NGOs, along with the great personal sacrifices of Rufo Guerreschi, Founder and Executive Director. With the financial support of external donors, we could reach our critical milestones and expand the scale of our mission, working towards achieving our ambitious goals. As detailed in our Who We Are section, the milestones and momentum we've gained demonstrate our commitment to making a substantial impact. We invite you to join us on this journey to contribute to a cause that promises growth and meaningful change to humanity.

This initiative offers a unique opportunity to stand at the forefront of transformative change. As pioneers in this movement, individuals gain the chance to shape the future, steering the world towards more sustainable and equitable practices, and enabling a safe and controlled use of AI. Your involvement leads to groundbreaking advancements and provides profound moral satisfaction. Being part of such a pivotal effort affords participants a deep sense of fulfillment, knowing that their contributions are helping to forge a better path for future generations. This experience is not just about witnessing change; it's about actively driving it and making a lasting impact on the global stage. As you collaborate with other forward-thinkers and innovators, you become an integral part of a collective endeavor that aims to make the world a better place, securing a legacy of impact and inspiration.

A Comprehensive Perspective on Advantages

Unlocking Overall Benefits:

- Participating in this initiative positions you at the forefront of change, providing the unique opportunity to be a true pioneer in shaping a better world and gaining profound moral fulfillment from driving meaningful progress.
- Your contribution is crucial in advancing global public good, improving safety, and enhancing the well-being of both present and future generations.
- Supporting our Initiative helps tackle what is arguably the greatest challenge humanity has ever faced.
- By helping our organisation we jointly bring solutions to the lack of an empowered, expert, federal and democratic global governance of AI and digital communications. A solution to such a challenge would:
    - Enable humanity to stave off the immense risks posed by AI and digital communications, and realise their astounding potential.
    - Protect the safety and well-being of your children, and their children, for generations to come.
- Affirm your legacy for decades, being recognised as an Early Patron of the Harnessing AI Risk Initiative.
- Join a community and movement of distinguished and accomplished individuals:

    - Receive invitations to exclusive social events during our Harnessing AI Risk Summit, including preliminary gatherings in Geneva and other locations.
    - Engage in and host Patron's meetings in your city.
    - Participate as a speaker at the Summit.
- Acquire an option to participate as a pre-seed funder or investor in the $15+ billion Global Public Interest AI Lab.

Contact us to discuss possible collaboration:
partnerships@trustlesscomputing.org

17. Opportunities for NGOs and Experts

We offer all individuals and organizations - and especially a globally-diverse set of **leading NGOs, scientists, academics and experts in the fields of AI safety, AI governance, state-grade IT security, and global democratic governance** - the opportunity to:

1. **Join as an individual or organizational Member of the Coalition for the Harnessing AI Risk Initiative,** by taking 5 minutes to read and sing the Open Call of the Harnessing AI Risk Initiative v.3.
    1. As a member of the Coalition, you'll be able to apply to join as advisor or as a speaking participant in our 1st Harnessing AI Risk Summit, this November in Geneva, and its Pre-Summit Virtual Conference this June 12th.
    2. As a member, you will help - within your abilities and possibilities, and without any formal or legal obligation - to help us shape the initiative and to attract more relevant stakeholders to the Initiatives, by sharing with them our opportunities for states, IGOs, NGOs, experts, leading AI labs and investors in the foreseen *Global Public Interest AI Lab.*
2. Join as a **partner** of the Initiative in various ways.

Contact us to discuss possible collaboration:
partnerships@trustlesscomputing.org

18. Opportunities for Funders & Investors in the Global AI Lab

We are pleased to present an exclusive investment opportunity for family offices, venture capitalists, angel investors, ultra-high-net-worth individuals, private banks, private investment funds, sovereign wealth funds, and regional intergovernmental funds (such as the European Investment Bank and the Asian Development Bank). We invite you to join as early pre-seed investors in a pioneering initiative: a $15 billion Global Public Interest AI Lab that is partially decentralised and operates as a public-private partnership.

The Lab is a crucial component of the Harnessing AI Risk Initiative, convening a diverse group of states in Geneva to develop and initiate an open global constituent assembly. This assembly aims to draft a binding treaty to establish a new intergovernmental organisation to ban unsafe AIs reliably, and jointly create, regulate and utilise the most capable, safe AIs.

The Global Public Interest AI Lab

- The Laboratory will function as an inclusive, partially decentralised entity, governed democratically through collaboration between states and appropriate technology firms. Its primary objective is establishing and maintaining a robust global leadership, or co-leadership, in human-controllable AI capability, technical alignment research, and AI safety measures.

- The Lab will accrue the capabilities and resources of member states and private partners and distribute dividends and control among member states and directly to their citizens, all the while stimulating and safeguarding private initiatives for innovation and oversight.

- The Lab will be primarily funded via project finance, buttressed by pre-licensing and pre-commercial procurement from participating states and client firms.

- The Lab will seek to achieve and sustain a resilient "mutual dependency" in its wider supply chain vis-a-vis superpowers and future public-private consortia through joint investments, diplomacy, trade relations, and the strategic industrial assets of participant states while remaining open to merging with them on equal terms.

A Comprehensive Perspective on Advantages

- We present investment opportunities ranging from 50,000 to $500,000. These funds will support our initiative to recruit two highly skilled professionals: a senior high-level AI infrastructure architect and a top-level diplomatic official based in Geneva. Investors will benefit from a maximum return cap of 50x, although the investment does not grant voting rights, except for sovereign funds.

- We invite you to participate as speakers at the 1st Harnessing AI Risk Summit to explore opportunities to engage through Letters of Intent or binding agreements as funders or investors in the Global Public Interest AI Lab.

Contact us to discuss possible collaboration:
partnerships@trustlesscomputing.org

19. Opportunities for Leading AI Labs

It is increasingly challenging for even the most well-resourced leading AL Labs to independently compete with a select group of U.S. and Chinese BigTech companies for AGI leadership, secure AI market niches, or influence the direction of an unchecked race for AGIs on a safe and beneficial course for humanity.

The suggested "7 trillion proposed AI consortium," which aims to focus on advancing large-scale artificial intelligence models for scientific discovers and involves a global network of researchers from various prestigious institutions, plans to address the complexities of building, training, and utilising AI models that reach or exceed one trillion parameters. However, the proposal overlooks the necessity of ensuring that it is accessible to all states and firms on equitable and fair terms and is established and governed democratically.

We cordially invite you to join us as an AI Lab Partner of the Harnessing AI Risk Initiative. This collaborative effort aims to unite a diverse array of states, AI labs, and essential supply chain companies from around the globe. We aim to establish a partially decentralised public-private $15+ billion democratically Global Public Benefit AI Lab and transform the IT Security Agency into an International AI Safety Agency to ban dangerous development and use of AI.

The Global Public Interest AI Lab

- The Lab will function as an inclusive, partially decentralised entity governed democratically through collaboration between states and appropriate technology firms. Its primary objective is establishing and maintaining a robust global leadership, or co-leadership, in human-controllable AI capability, technical alignment research, and AI safety measures.

- The Lab will accrue the capabilities and resources of member states and private partners and distribute dividends and control among member states and directly to their citizens, all the while stimulating and safeguarding private initiatives for innovation and oversight.

- The Lab will be primarily funded via project finance, buttressed by pre-licensing and pre-commercial procurement from participating states and client firms.

- The Lab will seek to achieve and sustain a resilient "mutual dependency" in its wider supply chain vis-a-vis superpowers and future public-private consortia through joint investments, diplomacy, trade relations, and the strategic industrial assets of participant states while remaining open to merging with them on equal terms.

A Comprehensive Perspective on Advantages

Role:

- As innovation partners and IP providers, AI Lab Partners will be compensated via revenue share, secured via long-term pre-licensing and pre-commercial procurement agreements from participating states and firms.

- As go-to-market partners, AI Lab Partners will gain permanent access to the core AI/AGI capabilities, infrastructure, services, and IP developed by the Lab. Such capabilities will aim to far outcompete all others in capabilities and safety and be unique in the actual and perceived trustworthiness of their safety and accountability.

- Partnership terms will be designed to maximise the chances of a steady increase in their market valuation, attracting the participation of AI labs governed by for-profit legal forms that legally mandate their CEOs to maximise shareholder value.

Benefits:

- Ability to advance both their mission to benefit humanity and their stock valuations and retain their agency to innovate at the root and application level within safety bounds.

- They should maintain the freedom to innovate at both the base and application layers and retain their ability to offer services to states, firms, and consumers within some limits.

- This setup will enable such labs to continue innovating capabilities and safety at the base and application layers without a “Wild West” race to the bottom among states and labs, advancing mission and market valuation.

The first seven AI labs that will join as participants will enjoy substantial economic advantages in relation to the Initiative and the Global Public Interest AI Lab relative to states that join later. More specifically, concerning all revenue share, IP compensations, decision-making, fees, and co-investments that will be required of AI labs of similar kind in the future by the Initiative and the Lab:

- The 1st to the 2nd lab participant will receive a 45% premium

- The 3rd to the 4th lab participant will receive a 30% premium

- The 5th to the 6th lab participant will receive a 15% premium

TCA Proposal:

We invite leading AI labs to:

- Join the Initiative as an AI Lab Partner by executing a straightforward, non-binding Letter of Interest and contributing CHF 20,000 to reserve your position as one of no more than seven partner labs.
- If interested but not ready to commit:
    - Participate in the 1st Harnessing AI Risk Summit this November or the Pre-Summit Virtual Conference on June 12th - and/or to closed-door bilateral meetings in Geneva or via videoconference - to learn more about the Initiative.
    - Become a Coalition for the Harnessing AI Risk Initiative member by signing our Open Call.

Contact us to discuss possible collaboration:
partnerships@trustlesscomputing.org

20. Opportunities for Regional Intergovernmental Organizations

The Problem

On their own, **regional intergovernmental organizations stand powerless in the face of AI**. They are unable to avoid its immense risks for human safety and for further concentration of power and wealth, and unable to realize its astounding opportunities. Even larger and more integrated ones like the European Union.

The limitations of their **mandates** and industrial capabilities in the **AI supply chain** make it impossible for them, on their own, to (1) achieve and sustain state-of-the-art AI **capabilities** and AI sovereignty for their member states and (2) have a proportional say in the creation of **global governance** institutions to regulate its safety, security and largely shape our future.

On their own, **even superpowers stand unable to go it alone** as mitigating the proliferation and safety risks of AI are foreseen by scientists to be much harder than nuclear.

The Opportunity

The Trustless Computing Association (TCA) invites a critical mass of globally diverse states to join our Harnessing AI Risk Initiative and participate in our summits. We envision our assembly as a leading platform for engagement in bilateral and multilateral meetings, aiming to launch the AI treaty-making process.

Such a multilateral and participatory treaty-making process will serve as a premise for creating a new open global intergovernmental organisation to jointly build and share the most capable safe AIs and reliably ban unsafe ones.

We invite Regional Intergovernmental Organisations to:

- Join the Initiative as a Regional Intergovernmental Organisation Member by endorsing a straightforward, non-binding Letter of Interest and contributing CHF 20,000. This engagement also allows us to collaborate as co-lead with one additional Inter-Governmental Organization (IGO), assuming shared oversight of the Harnessing AI Risk Initiative.
- For those interested but not yet prepared to commit, we offer participation in the 1st Harnessing AI Risk Summit this November or the Pre-Summit Virtual Conference on June 12th. These exclusive closed-door bilateral and multilateral meetings in Geneva or via videoconference present an opportunity to gain deeper insights into the Initiative's workings.

A Comprehensive Perspective on Advantages

Unlocking Overall Benefits:

- Enjoy exceptional economic development, sovereignty, safety, and civil rights advantages from the joint control and ownership of the Global Public Interest AI Lab. This ensures reliable long-term access and control over the most advanced safe AI systems for their governmental and private sector needs.
- Radically mitigate the immense risks to human safety and AI's extremely unaccountable concentration of power and wealth. Co-lead in shaping AI's ethics, limits, privacy, security, safety and accountability to realise its potential to bring astounding benefits to your citizens and humanity, in a mutually advantageous way for all.
- Increase leverage vis-a-vis other global governance and infrastructure initiatives for AI by leading states and firms.

Summit:

- Participate as a speaking participant in the Summit. "Observer status" or remote participation is also possible. Early state participants will have a guaranteed speaking slot on day one and will be more prominently displayed.

- Participate in co-designing the Initiative and, therefore, in the design of global governance of AI and the constituent process leading up to it. Early participants will be accorded privileged visibility.
- Learn about the current initiative and prospects of global AI governance, AI safety, and the Initiative.

Exclusive Benefits for Pioneers: the first 7 participant states will enjoy unique benefits:

- Economic advantages and discounts. Acquire a "priority option" to become one of a limited number of Founding State Partners of the Initiative. As such enjoy advantages and discounts concerning all fees, quotas, contributions of the first three years of Initiative and Lab, and their sovereign fund participation as Pre-seed Funders of the Lab:
    - 35% for the 1st IGO, and their member states
    - 25% for the 2nd IGO, and their member states
    - 10% for the 3rd IGO, and their member states
- Political prominence. Early participants will be more prominently included in the Summit, its documents, and webpages, and they will be guaranteed a speaking slot on day 1 of the Summit.

Contact us to discuss possible collaboration:
partnerships@trustlesscomputing.org

21. Harnessing AI Risk Summit

Bringing together a significant and diverse group of nations to establish a timely, expert-led, and inclusive constituent process, to facilitate the creation of a new global intergovernmental organization dedicated to AI and digital communications.

Working towards ensuring the safety, peace, equity, and democratic integrity of our digital future, ultimately guiding humanity into an era of unparalleled abundance and wellbeing.

At a Glance

We are honoured to invite esteemed representatives from state institutions, intergovernmental organizations (IGOs), artificial intelligence laboratories, distinguished non-governmental organizations (NGOs), and experts affiliated with the Coalition for the Harnessing AI Risk Initiative. This initiative, spearheaded by the Trustless Computing Association (TCA), aims to assemble a globally diverse group of stakeholders from various sectors to decisively address the multifaceted challenges associated with AI, through the democratic method.

Location:

Geneva, Switzerland

Event Date:

TBD November, 2024

Purpose of the Summit

- Achieve preliminary agreement among a number of diverse states to design a timely, expert-led, multilateral and participatory treaty-making process for the creation of an open global treaty-organization. This organization will collectively develop and exploit the safest and most advanced AI technologies, and reliably ban unsafe ones. We are drawing inspiration from the successful and democratic intergovernmental treaty-making process initiated by two U.S. states during the Annapolis Convention and culminating in the ratification of the U.S. Constitution by nine out of thirteen states. Our summit aims to replicate this historical model on a global scale, focusing solely on AI.

- Agree on the Scope and Rules for the Election of an *Open Transnational Constituent Assembly for AI and Digital Communications*. These guidelines should embody robust participation, inclusivity, expertise, and resilience principles. The objective is to facilitate the formation of an intergovernmental body poised to consistently and effectively promote the safety, welfare, and empowerment of all individuals for many generations to come.
- Achieve preliminary agreement among states, AI labs, investors, funders and technical partners on their participation in a democratic, partly decentralized public-private Global Public Benefit AI Lab and ecosystem.

Agenda

Day 1 will feature a combination of 40-minute panel discussions and 5-10 minute "lightning talks" presented by leading experts and NGOs.

Day 2, will include a variety of deliberative working sessions, educational sessions—both one-way and interactive—and both multilateral and bilateral meetings.

Day 1

Each session will include one primary video-recorded track, and may feature up to two additional secondary tracks:

08.30 - 09.00: **Welcome and Introduction**: Hosted by the Trustless Computing Association in collaboration with local, national, and/or international authorities.

09.00-09.10. TBD Lightning Talk

09.10 - 09.45:
AI Risks: Extreme and Unaccountable Concentration of Power and Wealth (democracy, Inequality, civil rights, biases and minorities, unemployment and loss of agency). Human Safety Risks (loss of control, misuse, accidents, war, dangerous science). Risks' comparative importance and timelines, shared mitigations, win-wins and synergies.

10.00-10.10
TBD Lightning Talk

10.10 - 10.45:
AI Opportunities: Abundance, Health, Safety, Peace, Happiness. Can future AI not only bring

amazing practical benefits but even increase very significantly the average happiness and wellbeing of the average human?

11.00-11.10
TBD Lightning Talk

11.10 - 11.45:
AI Scenarios 2030+: (a) Mostly Business as Usual; (b) Global autocracy or oligarchy; (c) Human Safety Catastrophes or Extinction; (d) AI Takeover: Bad and Good Cases; (e) Humanity's Federal Control of Advanced AI.

12.00-12.10
TBD Lightning Talk

12.10 - 12.45:
Preliminary Designs: Federalism & Subsidiarity (global, nation and citizen levels). Checks and Balances. Complexity, Urgency, Expertise, and Acceleration. Transparency, participation, trustlessness and decentralization. Political, technical and future-proof feasibility of bans of unsafe AI. Win-wins for oversight, public safety, civil liberties and democracy. Democracy & monopoly of violence. Role of superpowers, firms and security agencies.

14.00-14.10
TBD Lightning Talk

14.10 - 14.45
Scope and Functions: An AI Safety Agency to set and enforce AI safety regulations worldwide? A Global Public Interest AI Lab, to jointly develop, control and benefit leading or co-leading capabilities in safe AI, and digital communications/cloud infrastructure, according to the *subsidiarity* principle? An IT Security Agency, to develop and certify trustworthy and widely trusted “governance-support” systems, for control, compliance and communications? Other?

15.00-15.10
TBD Lightning Talk

15.10 - 15.50
Constituent Process: Participation. Expertise. Inclusiveness. Weighted Voting. Global citizens’ assemblies. A Global *Collective Constitutional AI*?. *Scope and Rules for the Election* of an *Open Transnational Constituent* Assembly. Interaction with other constituent initiatives.

16.00-16.10
TBD Lightning Talk

16.10 - 16.50
Global Public Interest AI Lab: Viability. Decentralization vs Safety. Subsidiarity principle. Initial funding: project finance, spin-in or other model? Role of private firms. Business models. Safety accords with other leading state private AI labs. The Superintelligence/AGI "option".

17.00-17.10
TBD Lightning Talk

17.10 - 17.50
Setting AI Standards: Technical, socio-technical, ethical and governance standards for the most advanced AIs. Agile, measurable and enforceable methods to assess AI systems, services and components that are safe and compliant.

Day 2

The second day of the Summit will entail:

- To-be-determined close-door, closed and open workshops, working session and self-organized meetings, whereby states and other participants will engage in fostering consensus on key documents detailing the constituent process, and preliminary designs of the resulting IGO.
- Several educational sessions on the technical and non-technical aspects of advanced AI safety, security and privacy and governance. Mainly geared towards state representatives, and run by leading expert NGO participants.

Speaking Participants

- *Confirmed (subject to their availability for the new TBD November date)*
    - Rufo Guerreschi, President of the Trustless Computing Association (TCA).
    - Ansgar Koen, Global AI Ethics and Regulatory Leader at Ernst & Young. TCA Advisor.
    - Robert Trager. Director, Oxford Martin AI Governance Initiative and International Governance Lead at the Centre for the Governance of AI.
    - Kenneth Cukier. Deputy Executive Editor of The Economist, and host of its weekly tech podcast.

- Flynn Devine, researcher on participatory AI governance methods, including research with the Collective Intelligence Project and on 'The Recursive Public'. Co-Initiator of the Global Assembly for COP26.
- Brando Benifei, Member of European Parliament and Co-Rapporteur of the European Parliament for the EU AI ACT.
- Mohamed Farahat, member of UN High-Level Advisory Board on Artificial Intelligence. TCA advisor.
- Kay Firth-Butterfield, CEO of Good Tech Advisory. Former Head of AI and Member of the Exec Comm at World Economic Forum.
- Gordon Laforge. Senior Policy Analyst at New America Foundation. TCA Advisor.
- Marco Landi, President of the EuropIA Institut. Former Group President and COO of APPLE Computers in Cupertino. TCA steering advisor.
- Robert Whitfield, Chair of the Transnational Working Group on AI at the World Federalist Movement. Chair of One World Trust.
- Paul Nemitz. Principal Advisor at the European Commission. Senior Privacy and AI policy expert. TCA advisor.
- Axel Voss. Member of European Parliament and member of the Committee on Civil Liberties, Justice and Home Affairs (LIBE), and the Committee on Artificial Intelligence in a Digital Age (AIDA).
- Akash Wasil, AI Policy Researcher at Control AI.Former senior researcher at Center on Long-Term Risk and Center for AI Safety.
- Muhammadou M.O. Kah. Professor and Ambassador Extraordinary & Plenipotentiary of The Gambia to Switzerland & Permanent Representative to UN Organisations at Geneva, WTO & Other International Organisations in Switzerland. TCA Advisor.
- Jan Camenisch, Chief Technology Officer of Dfinity, a blockchain-based internet computer. Phd researched with 130 paper and 140 filed patents.
- Aicha Jeridi, Vice President of the North African School and Forum of Internet Governance. Member of the African Union Multi-Stakeholder Advisory Group on Internet Governance.
- Beatrice Erkers. Chief Operating Officer at the Foresight Institute.
- Allison Duettmann. Chief Executive Officer at the Foresight Institute.
- Lisa Thiergart. Research Manager at Machine Intelligence Research Institute (MIRI). AI Alignment Researcher.
- David Wood, President of the London Futurists association.

- Chase Cunningham. Vice President of Security Market Research at G2. Former Chief Cryptologic Technician at the US National Security Agency. Pioneer of Zero Trust. TCA advisor.
- Darren McKee. Senior Advisor at Artificial Intelligence Governance & Safety Canada (AIGS). Author of "*Uncontrollable: The Threat of Artificial Superintelligence and the Race to Save the World*"
- Sebastian Hallensleben, Head of AI at VDE, Co-Chair of the OECD Expert Group on AI (AIGO), Chair, Joint Technical Committee 21 "Artificial Intelligence" at CEN and CENELEC.
- John Havens. Exec. Dir. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.
- Philipp Amann. Group CISO at Austrian Post. Former Head of Strategy EUROPOL Cybercrime Centre.
- Ayisha Piotti. Director of AI Policy at ETH Zurich Center for Law and Economics.
- Jan Philipp Albrecht, President of the Heinrich Böll Foundation. Former Greens MEP. Former Minister of Digitization of the German state of Schleswig-Holstein. TCA steering advisor.
- Alexander Kriebitz, Research Associate at the Institute for Ethics in Artificial Intelligence
- David Evan Harris, Chancellor's Public Scholar at UC Berkeley. Senior researcher at Centre for International Governance Innovation (CIGI), Brennan Center for Justice, International Computer Science Institute.
- Richard Falk, professor emeritus of international law at Princeton University. Renowned global democratization expert. Chairman of the Trustees of the Euro-Mediterranean Human Rights Monitor.
- Peter Park, MIT AI Existential Safety Postdoctoral Fellow and Director of StakeOut.AI
- Pavel Laskov, Head of the Hilti Chair of Data and Application Security University of Liechtenstein
- Albert Efimov, Chair of Engineering Cybernetics at the Russian National University of Science and Technology. VP of Innovation and Research at Sberbank.
- Joe Buccino, AI policy and geopolitics expert. US Defense Ret. Colonel. TCA Advisor.
- Tjerk Timan, trustworthy and fair AI Researcher. TCA Advisor.
- Roberto Savio, communications Expert. Founder and Director of Interpress Service. TCA advisor.

Organizations

NGOs

- *Confirmed (subject to their availability for the our TBD November date)*
  - Trustless Computing Association
  - Machine Intelligence Research Institute (MIRI)
  - World Federalist Movement
  - Oxford Martin AI Governance Initiative
  - ETH Zurich Center for Law and Economics
  - Heinrich Böll Foundation
  - Foresight Institute
  - Artificial Intelligence Governance & Safety Canada
  - Institute for Ethics in Artificial Intelligence
  - North African School and Forum of Internet Governance
  - Europia Institut
  - Dfinity
  - StakeOut.AI

States and IGOs:

- *Confirmed:*
  - The Mission of Gambia to the UN in Geneva
- *Engaged*:
  - In March, we conducted meetings with the United Nations missions in Geneva from four states, which included three heads of mission (ambassadors) and three experts in AI and digital domains. We are currently engaging with three additional missions. Collectively, these states, primarily from Africa and South America, represent a population of 120 million, combined GDP of $1.4 trillion, and manage sovereign funds totaling $130 billion. In early April, we received a written expression of interest from the ambassador to the United Nations in Geneva representing one of the three largest regional intergovernmental organizations, which encompasses dozens of member states.

AI Labs

- *Engaged:*
  - Since December, we have been in extended talks with 3 of the 5 top AI Labs about their interest in participating in the *Global Public Interest AI Lab*.

Pre Summit Virtual conference on June 12th, 2024

We are honoured to invite esteemed representatives from distinguished non-governmental organizations (NGOs) and experts affiliated with the Coalition for the Harnessing AI Risk Initiative.

Online Event Date:

June 12th, 2024

Pre-Summit Purpose

- Consolidate and expand a Coalition for the Harnessing AI Risk Initiative, composed of geographically diverse and unbiased non-governmental organizations (NGOs), experts, influential figures, and former public officials. This coalition aims to enhance the initiative's momentum and credibility with states and regional intergovernmental organizations (IGOs).
- Secure agreement on Version 4 of the Open Call for the Harnessing AI Risk Initiative and finalize other related documents.
- Produce and disseminate testimonials, articles, publications, and videos to promote, explain, and advocate for the Initiative.

Pre-Summit Agenda

- 15.30 - Online Panel:
  AI Risks and opportunities: the prevailing science
- 16.00 - Online Panel:
  Treaty-making for technological risks: nuclear, bioweapons, encryption, climate
- 16.30 - Online Panel:
  Treaty-making for AI: the open intergovernmental constituent assembly model
- 17.00 - Online Panel:
  Mitigating the risks of competing AI coalitions, AIs and AI governance initiatives.
- 17.30 - Online Panel:
  Foreseeing and navigating complex socio-technical future AI scenarios
- 18.00 - Online Panel:
  Open Call for the Harnessing AI Risk Initiative (v.4)

Pre-Summit Speakers

A globally-diverse set of NGOs, and experts in AI and global governance:

- Ansgar Koen, Global AI Ethics and Regulatory Leader at Ernst & Young. TCA Advisor.

- Jan Philipp Albrecht, President of the Heinrich Böll Foundation. Former Greens MEP. Former Minister of Digitization of the German state of Schleswig-Holstein. TCA steering advisor.

- (**moderator**) David Wood, President of the London Futurists association. TCA Advisor.

22. About Us

TCA at a Glance

Please explore our Highlights and delve into our Summary, which provides an overview of our Organization, Strategic Approach, Key Initiatives, and more. This section is designed to give you a clear and concise understanding of our core activities and strategic direction.

Summary

Based in Geneva, the Trustless Computing Association (TCA) is an impartial, non-profit organisation committed to enhancing the safety, liberty, democratic oversight, and accountability of digital communications and AI.

Since our inception in 2015, we have focused on building a consensus for establishing new, open, and participatory intergovernmental bodies. Such organisations aim to develop and certify more trustworthy and secure end-to-end IT systems essential for confidential communication and managing critical AI, social media, and other vital societal infrastructures.

Our unique approach combines robust, tested, and time-proven trustless technical, socio-technical, and governance systems inspired by the trustworthiness paradigms of democratic constitutions, electoral processes, and citizen-jury systems.

On June the 28th, 2023, in response to AI's growing challenges and opportunities, TCA initiated a movement to create three new global intergovernmental organisations and their participatory constituent processes leading up to their creation. We aimed to develop a framework to govern AI and digital communications for the global public good. This initiative was presented at one of the United Nations events. It incorporated the Trustless Computing Certification Body as one of the prospective agencies of a new IGO. It was showcased at the United Nations public event organised by the Community of Democracies, which includes 32 member states.

Governance of our association is vested in designated bodies, each composed of three members with equal voting rights, ensuring decisions align with our foundational principles.

Mission and Vision 2030

Summary

Trustless Computing Association's (TCA's) initiative envisions the establishment of a new inter-governmental organisation dedicated to leveraging the potential of Information Technology (IT) and Narrow Artificial Intelligence (AI). Our mission aims to prevent catastrophic outcomes and dystopias by promoting global dialogue and effective coordination while empowering individual and collective control over technologies to enhance privacy, security, health, and abundance.

Acknowledging the next decade as potentially the most decisive in human history, we recognise unprecedented opportunities and threats. Opportunities stem from rapid advancements in IT and AI, which promise enhanced well-being and security. Conversely, threats include climate change, nuclear proliferation, and the concentration of power and wealth, forming a complex, multidimensional global crisis that requires urgent and strategic global responses.

Our organisation advocates for a transformative approach to global cooperation, mirroring successful federation processes from national contexts like the United States, Germany, and Switzerland, but on a global scale. This involves forming a new type of transnational federation initiated by diverse nations. These nations would commit to an open, representative process, setting a foundation for equitable global governance.

We cannot emphasise enough the potential dangers that may arise due to the insufficient regulation of current digital platforms, which are controlled by a select few nations and wealthy individuals. The risks of technology misuse for manipulation and surveillance and ineffective regulatory responses are significant and concerning. Therefore, we champion a parallel, more accountable digital communications infrastructure governed by a transparent, participatory intergovernmental body.

The ultimate goal is to reframe IT and AI as tools for collective empowerment, enhancing global coordination and democracy, thereby realising a vision of abundant health, peace, and well-being for All. This requires a concerted effort from well-intentioned global actors to shift control from the few to the many, ensuring technology serves humanity's broadest interests rather than narrow, elite ones. TCA takes leadership in such efforts driving the Annapolis convention of AI.

-

Concept of Trustlessness

Introduction to Trustlessness

The term "trustless" might initially sound negative, suggesting something or someone not worthy of trust. However, in the context of our Trustless Computing Association, it carries a much more nuanced and crucial significance.

Modern Interpretation of Trustlessness

In today's world, "trustless" reflects a proactive approach to safety and reliability, particularly in technology and governance. It implies not just a lack of trust, but a systematic approach to verify and validate trustworthiness across all aspects of a system or institution. This methodology is derived from understanding that while most people act in good faith, the potential for significant harm exists if just a few act maliciously or under pressure.

The concept of trustlessness is about ensuring that no single person or group can cause disproportionate damage. It's about designing systems and mechanisms to assess and mitigate the risks posed by any entity within the system, thus safeguarding against the weakest link. This is vital in everything from maintaining the integrity of democratic processes to ensuring the safety of commercial aviation.

Historical Context of Trustlessness

Historically, the term "trustless" has evolved. Centuries ago, societal norms demanded unquestioning trust in authority figures like monarchs or religious texts, often at severe personal risk for dissenters. Over time, as democratic and scientific methods developed, these norms shifted. The skepticism once punished is now the foundation of modern critical thinking and democratic integrity.

In its primary, modern usage, "trustless" describes a stance that eschews blind trust, advocating instead for rigorous verification and oversight. This approach underpins the reliability of systems from electoral politics to nuclear safety.

Trustless vs. Trusted Computing

In the realm of information technology, "trustless computing" stands in contrast to "Trusted Computing" as advocated by major IT firms. While Trusted Computing relies on inherently trusting certain designated components deemed secure by their manufacturers, Trustless Computing demands transparency and testability across all components. It rejects the notion of pre-approved trustworthiness, focusing instead on comprehensive, impartial verification to prevent vulnerabilities.

Philosophical Underpinnings

Adopting a trustless approach does not imply a cynical view of human nature. Rather, it recognizes that even well-intentioned individuals can fail under certain pressures. By designing systems that do not rely on inherent trust, we better equip them to handle unexpected challenges, thereby enhancing overall safety and integrity.

In conclusion, the Trustless Computing Association is committed to advancing a model where trust is earned through rigorous scrutiny and constant vigilance, ensuring technologies and institutions remain robust and trustworthy in the face of both everyday and extraordinary challenges.

Funding

We are immensely proud to announce that our organisation is self-funded by our main partners Rufo Guerreschi and Alexandre Hovarth, underscoring our own deep commitment to the Initiative. The dedication and volunteer efforts of our world-renowned experts have enabled us to achieve critical milestones and advance our projects significantly. We extend our heartfelt gratitude for the generous financial support from the following entities: EIT Digital, ECSEL-JU (two EU governmental agencies). Additionally, we are thankful for the invaluable non-financial assistance from three accelerator programs: Hardware.co (Berlin, 2016), Fintech Fusion (Geneva, 2019), MACH37 (McLean, 2021). This collective support not only fueled our progress but also reinforced the strength and dedication behind our mission.

As we forge ahead, the engagement and financial support of new partners are crucial to the continuation and expansion of our efforts. We invite new entities to join us in this transformative journey. Your support will play a pivotal role in enabling us to reach new heights and achieve broader impact. Partnering with us offers a unique opportunity to contribute directly to visionary initiatives led by top-tier experts. We are eager to welcome new supporters who are passionate about making a tangible difference in the world.

For further details on how you can become involved and support us financially, please contact:

partnerships@trustlesscomputing.com

Transparency

Transparency is a fundamental value, embedded at the core of our organisation. We believe that openness in our operations and communications fosters trust and accountability, both essential for building and maintaining strong relationships with our partners and supporters. By clearly sharing our goals, processes, and outcomes, we ensure that all stakeholders are fully informed and can see the direct impact of their contributions. This commitment to transparency not only upholds our integrity but also enhances our collaborative efforts, enabling more effective and sustainable advancements in our initiatives.

The Trustless Computing Association is a non-profit association within the meaning of Articles 60 et seq. of the Swiss Civil Code (“CC”). It was created on May 21st, 2021 in Geneva.

- *Founders' Meeting*, of May 21st 2021 (pdf)
- Current Statute, revised on May 9th 2024 (pdf)
- *Minutes of the General Assembly and Board Meeting of the Trustless Computing Association of May 23rd, 2023* (Available on qualified request). Includes:
  - Annual Activity Report (May 2021- May 2022)
  - Annual Activity Report (May 2023- May 2023)
  - Financial Report (May 2021- May 2022)
  - Financial Report (May 2022- May 2023)
  - Programme of Action (May 2023- May 2024)
  - Budget (May 2023- May 2024)
- A *General Assembly and Board Meeting of the Trustless Computing Association* will be held on May 9th, 2024 to approve Annual Activity and Financial Report of the past year and the Programme and Budget for the next.
- *Swiss Registration*: For Geneva-based associations, it is not required to register with the Chamber of Commerce, unless they perform commercial activities. Yet, we plan to register and request for "tax-exempt status" as soon as we receive our first donations.

Our Journey

Summary

In 2015, we embarked on a journey to enhance TCCB and Seevik Net, initiating a series of scholarly research and publications. This year also marked the launch of the inaugural Free and Safe in Cyberspace conference in Brussels, setting the stage for a series of impactful discussions. By 2019, our endeavours led to the creation of Trustless.AI, a startup spin-off that successfully drew private investments to develop an initial architecture, ecosystems, proofs-of-concept, and systems aligned with TCCB standards, concluding its operations in September 2023.

The year 2021 was another landmark as we introduced the Trustless Computing Certification Body and the Seevik Net Initiative during the 8th edition of the Free and Safe in Cyberspace event in Geneva. Our progress continued through 2023, by which time we had hosted eleven editions of the Free and Safe in Cyberspace across cities like Geneva, Zurich, Brussels, New York, and Berlin, attracting over 120 distinguished participants from around the globe. Moreover, we expanded our network of world-class advisors and research partners, significantly advancing the Trustless Computing Paradigms and engaging with over 15 countries to further the initiative.

Why Trustless

At the Trustless Computing Association (TCA), we are driven by a mission that matters to all of humanity. Our work is about achieving a collective impact and addressing critical AI challenges. Together, we can achieve remarkable things and leave a lasting legacy. By choosing to work with us, you are taking part in making history.

Summary

In contemporary societies, the efficacy of advanced technologies like commercial aircraft and the stability of established democratic institutions rely on a twofold understanding. First, it is assumed that the majority of people naturally act in the collective interest, establishing a baseline of trust. Second, there is an acknowledgment of the need to proactively address the possibility of harmful actions by individuals or small groups, driven by either internal motives or external pressures.

Consequently, it is imperative to maintain a consistent level of scrutiny towards both individuals and institutions, coupled with the implementation of systems to assess their reliability and curtail potential damages. This vigilance is critical in mitigating risks they may pose, as the integrity of any system is as strong as its most vulnerable component. This cautious approach should be applied universally, from a lone passenger intending to commit terrorism aboard an aircraft to a senior official in the Federal Aviation Administration engaged in corruption. Similarly, it encompasses elected officials who may seek to extend their power unduly, and groups that threaten to undermine democratic norms through acts of sedition.

Leadership

Introduction

We are privileged to be guided by a team of esteemed experts whose leadership and profound knowledge drive our organisation forward. Our leaders bring decades of experience across various fields, giving us invaluable insights and strategic guidance. We are deeply fortunate to tap into their expertise, which propels our mission and ensures we remain on the right track.

Board of Directors

Rufo Guerreschi - President and Founder of the Trustless Computing Association
Member of the Board of Directors

Rufo Guerreschi is an accomplished activist, researcher, and entrepreneur dedicated to advancing liberty and democracy globally through innovative digital technologies. He founded the Trustless Computing Association and its spin-in, TRUSTLESS.AI. Rufo has been the visionary behind the Free and Safe in Cyberspace conference series since 2015, which advocates for a groundbreaking IT security paradigm—the Trustless Computing Paradigms—to harmonise personal freedoms with public safety.

Previously, he was the founder and CEO of Participatory Technologies Srl, a company that provided open-source e-democracy solutions to transnational political organisations across three continents. As the Global Vice President at 4thPass, he pioneered the first Java mobile app store system, securing over 10 million euros in deals with major clients like Telefonica. As the CEO of Open Media Park, Rufo significantly increased the project's valuation from 3 million to 21 million euros, focusing on developing a cutting-edge cybersecurity and new media technology park in Rome.

Roberto Savio - Member of the Board of Directors

Roberto Savio is an esteemed Italo-Argentinian journalist, communication expert, and a respected political commentator. He is also a passionate activist for social and climate justice, advocating for global governance. He co-founded Inter Press Service in 1964 and led its transformation into the premier news agency of the Global South over the following decades. Roberto served as the Deputy Director of the World Policy Forum Scientific Council, an initiative founded by Mikhail Gorbachev.

Since its inception in 2001, he has been an active member of the International Committee of the World Social Forum. Roberto oversees international relations at the Belgrade European Centre for Peace and Development. Previously, he was the Chairman of the Board of the Alliance for a New Humanity, further demonstrating his commitment to fostering global collaboration and peace.

Davide Cova - Member of the Board of Directors

Davide Cova is the founder and director of the Dorjeling Center in Piediluco, Italy, he is an experienced teacher of secular Buddhist meditation, psychology, and philosophy. Holding a degree in Political Philosophy, he has further specialised in Conflict Resolution and gained valuable experience as a political analyst at the United Nations in New York and Copenhagen.

Davide was a dedicated disciple of the Dalai Lama and the Venerable Lama Geshe Ciampa Gyatso for a decade. He completed a rigorous seven-year full-time Masters Program in Buddhist Studies and spent three years as a Buddhist monk. His spiritual exploration included pilgrimages to prominent Buddhist sites in India and significant Christian sites in the Middle East, followed by a 16-month solitary meditative retreat.

Steering Committee

Akash Wasil - Member of the Steering Committee

Akash Wasil is an AI Policy Researcher at Control.AI, specialising in the critical field of AI governance. Previously, he served as an AI Governance Researcher at the Center for AI Safety in Berkeley, California, and as an Analyst at the Stanford Existential Risks Initiative, where he contributed to understanding and mitigating risks associated with advanced AI technologies.

Akash graduated Phi Beta Kappa in Psychology from Harvard University, where he distinguished himself academically and laid a solid foundation for his contributions to the intersection of technology and policy.

Marco Landi - Member of the Steering Committee

Marco Landi serves as President of EuropIA Institut; he leads a non-profit organisation dedicated to advocating for safe and sovereign AI. He is also the President of Questit, a pioneering AI startup. Previously, he was President and Chief Operating Officer at Apple Computers in Cupertino, overseeing Global Operations, Marketing, and Sales, significantly contributing to the company's success.

In addition to his corporate roles, Marco founded the World AI Cannes Festival, one of Europe's leading AI conferences and expos. This event highlights cutting-edge developments in AI technology, attracting industry leaders and innovators from around the globe.

Jan Philipp Albrecht - Member of the Steering Committee

Jan Philippe Albrecht is a European and German politician recognised for his significant contributions to digital civil rights. As a member of the EU Parliament, he notably served as the Vice-Chair of the European Parliament Committee on Civil Liberties, Justice, and Home Affairs, emphasising his commitment to civil liberties. From 2018 to 2022, he took on the Minister for Energy, Agriculture, the Environment, Nature, and Digitalization role in the German State of Schleswig-Holstein, enhancing his impact on environmental and digital policy.

Additionally, Jan Philippe is a member of NOYB – European Center for Digital Rights, an organisation spearheaded by Max Schrems that focuses on protecting digital privacy. He currently leads the Heinrich Böll Foundation as its President, driving initiatives promoting the environment, civil rights, and multilateral democracy across 30 countries with a dedicated staff of 180.

Advisory Board

Paul Nemitz - Member of the Advisory Board

Paul Nemitz currently serves as the Principal Adviser on Justice Policy at the EU Commission, where his expertise guides significant policy developments. Additionally, he is a member of the German Data Ethics Commission and the Global Council on Extended Intelligence, reflecting his deep commitment to ethical standards in data use and artificial intelligence. He also holds a position as a Visiting Professor of Law at the College of Europe, where he imparts his extensive knowledge to the next generation of legal professionals.

Previously, Paul was the Director of Fundamental Rights and Union Citizenship and later a Principal Advisor at the Directorate-General for Justice of the European Commission. In these roles, he led pivotal work on the GDPR and the EU–US Privacy Shield, focusing on data protection and privacy within law enforcement and national security frameworks.

Ansgar Koene - Member of the Advisory Board

Ansgar Koene serves as the Global AI Ethics and Regulatory Leader at Ernst & Young, one of the world's top five consultancy firms. His leadership extends to his role as Chair of the IEEE Working Group P7003 Standard for Algorithm Bias Considerations, where he focuses on developing ethical guidelines for AI applications. Additionally, he is a Trustee at the 5Rights Foundation, advocating for children's digital rights.

Angsar is also a member of the AI Ethics Board at Hayden AI, contributing his expertise to guide the development of responsible AI technologies.

Muhhamadou Kah - Member of the Advisory Board

Amb. Prof. Muhammadou M.O. Kah currently serves as the Ambassador Extraordinary and Plenipotentiary of The Gambia to Switzerland and the Permanent Representative of The Gambia to the United Nations organisations at Geneva (UNOG), the World Trade Organization (WTO), and other international organisations in Switzerland. His distinguished career also includes his role as Vice Chancellor/Rector of The University of The Gambia, where he made history as the institution's third and first Gambian-born leader. He has further enhanced academic and technological landscapes internationally as the Provost/Vice President for Academic Affairs at the American University of Nigeria, Yola, and the Vice-Rector for Technology and Innovation and Founding Dean of the School of IT and Engineering at ADA University in Baku, Azerbaijan.

In addition to academic and diplomatic responsibilities, Amb. Prof. Kah has held numerous significant positions in the financial and educational sectors. He was the founding chairman of Zenith Bank Gambia Limited and served on the Governing Board of The African University of Science and Technology in Abuja, Nigeria. He was a founding Board member of The International Digital Health and AI Research Collaborative (I-DAIR) in Geneva, Switzerland, and a Governing Council Member of the DCAF-Geneva Center For Security Sector Governance Foundation. His leadership extends to his past role as Chairman of the Africa Group of Ambassadors in Geneva from April 2021 to September 2021, Vice President (Africa) for the UN Human Rights Council for 2022 and 2023, Vice Chairman for UNCTAD's Commission on Science and Technology for Development, and as a member of the Advisory Board of the UNCTAD Trade and Development Bureau from June 2021 to July 2023. He also served as one of two Ambassadors designated as Friends of the Chair of the World Intellectual Property Organization (WIPO) General Assembly. He continues his involvement as one of the Vice Chairs of the General Assembly of WIPO.

Mohamed Farahat - Member of the Advisory Board

Mohamed Farahat serves on the United Nations High-Level Advisory Board for Artificial Intelligence; he contributes to shaping global AI strategies and ethical guidelines. Mohamed is also a member of the Multistakeholder Advisory Group of the African Internet Governance Forum, engaging in discussions that influence the continent's internet policy framework.

As a legal and political researcher, he has authored 20 published articles addressing various topics, including international refugee law, statelessness, politics, democracy, parliamentary studies, digital rights, and internet governance. His scholarly work informs and impacts ongoing debates and developments in these critical fields.

Gordon Laforge - Member of the Advisory Board

Gordon Laforge is a Senior Policy Analyst at the New America Foundation, specialising in global AI governance and geopolitics. His expertise in this field is recognised internationally, and he contributes to influential policy developments. Previously, he was a Senior Researcher at Princeton University, where he conducted extensive studies on technology and international relations.

A Fulbright Fellow, his academic and professional accomplishments reflect a deep commitment to understanding and shaping the strategic implications of artificial intelligence on a global scale.

Camila Lopez Badra - Member of the Advisory Board

Camila Lopez Badra currently serves as the Chair of the Outreach Committee of the World Federalist Movement and as the International Director of the United Nations Parliamentary Assembly Model. In addition, she is the Executive Director of Democracia Global – Movimiento por la Unión Sudamericana y el Parlamento Mundial, driving initiatives that advocate for greater regional and global political integration.

Beyond her leadership roles, Camila coordinates the campaign to establish a Latin American and Caribbean Criminal Court against Transnational Organized Crime (COPLA). She also serves as a Parliamentary Advisor in the Honorable Chamber of Deputies of the Argentine Republic. Apart from her extensive political and advocacy work, she is a passionate tango teacher, blending her professional pursuits with cultural engagement.

Tjerk Timan - Member of the Advisory Board

Tjerk Timan is a PhD scientist with a specialisation in Trustworthy AI, digital privacy, and security. His expertise positions him at the forefront of technology policy and ethical AI practices. He serves as the Principal Consultant at Technopolis Group, instrumental in shaping high-impact innovation strategies and technology policies.

Previously, Tjerk was a Senior Scientist in Strategy & Policy for AI at TNO Group. In this role, he focused on developing and integrating ethical guidelines and policy frameworks to govern the deployment of artificial intelligence technologies.

Chase Cunningham - Member of the Advisory Board

Chase Cunningham serves as the Chief Strategy Officer at Ericom, where he shapes strategic initiatives and drives technological advancements within the company. Before joining Ericom, he was Chief Cryptologic Technician at the National Security Agency (NSA), leading technical teams across multiple operations in collaboration with the FBI and CIA. His leadership facilitated significant advancements in national security projects.

Chase's extensive background includes roles such as Vice President and Principal Analyst at Forrester Research, where he developed the Zero Trust eXtended framework. Additionally, he has served as the Director of Threat Intelligence at Armor and Director of Cyber Analytics at Decisive Analytics, enhancing cybersecurity measures through innovative analytical strategies.

Boris Taratine - Member of the Advisory Board

Boris Taratine is a leading global cybersecurity executive with extensive expertise in the financial sector. His experience includes roles as Chief Cybersecurity Architect at Lloyds Banking Group and Principal Security Architect at VISA, where he played a pivotal role in enhancing security architectures within these major financial institutions. He is pursuing a PhD in Physics at the University of St. Petersburg, Russia, demonstrating his deep commitment to advancing scientific and technical knowledge.

In addition to his professional and academic pursuits, Boris holds dozens of patents in physics and cyber defence. As a citizen of the UK, Canada, and Russia, he actively promotes freedom of speech and seeks to mitigate the impacts of global propaganda and the fight for informational superiority.

Mika Lauhde - Member of the Advisory Board

Mika Lauhde leads the IT R&D department at the International Committee of the Red Cross, focusing on developing innovative technological solutions to enhance humanitarian efforts worldwide. Before this role, he was the Global Vice-President of Cyber Security and Privacy at Huawei Technologies, where he spearheaded initiatives to strengthen global data protection and privacy protocols.

His extensive background in security also includes serving as the Head of Business Security and Continuity at Nokia for 13 years, where he was integral in establishing robust security frameworks. Mika has been an influential figure in cybersecurity policy, having been a member of the ICT Security Advisory Board of the Republic of Finland, the European Network and Information Security Agency (ENISA), and the EU Commission's Network and Information Security (NIS) platform. Additionally, he contributed to the EU Commission security advisory board under Commissar Redding and was a Trust In Digital Life founding board member.

Flynn Devine - Member of the Advisory Board

Flynn Devine is a dedicated researcher focused on AI governance. His notable contributions include leading the 'The Recursive Public' project, funded by OpenAI's Democratic Inputs to AI grant. This initiative explores the interplay between artificial intelligence and democratic processes. He is also involved with the Collective Intelligence Project, enhancing collaborative approaches to complex problem-solving in AI.

Previously, Flynn was part of the Public Policy program at The Alan Turing Institute, where he applied data science and artificial intelligence to address key policy challenges. Additionally, he co-initiated the Global Assembly for COP26, facilitating international discussions on climate action and sustainable policy.

Aicha Jeridi - Member of the Advisory Board

Aicha Jeridi is the Vice President of the North African School and Forum of Internet Governance. In this role, she significantly influences the development and implementation of Internet governance policies in the region, fostering dialogue and collaboration among key stakeholders.

Additionally, Aicha is a member of the African Union Multi-Stakeholder Advisory Group on Internet Governance. This involvement allows her to contribute to broader discussions and initiatives to improve digital infrastructure and governance across the African continent.

James S. Fishkin - Member of the Advisory Board

James S. Fishkin is a Tenured Professor of Communication and Political Science at Stanford University, where he also serves as the Director of the Center for Deliberative Democracy. His academic and leadership roles at Stanford have positioned him at the forefront of research and education in political communication and democratic theory.

His notable contribution to the field, Deliberative Polling, was introduced in 1988 and has since gained global recognition as an effective method for public consultation. Professor Fishkin’s scholarly works, including "When the People Speak" (2009), "Deliberation Day" (2004, co-authored with Bruce Ackerman), and "Democracy and Deliberation" (1991), explore the complexities and practicalities of fostering deliberation in democratic processes.

Richard Falk - Member of the Advisory Board

Richard Falk is a renowned Professor Emeritus of International Law at Princeton University. He is recognised globally as a leading academic in the field of global democratic governance. His tenure has established him as a pivotal figure in discussing and advancing international legal frameworks and democratic principles.

From 2008 to 2014, he served as the UN Human Rights Council Rapporteur on the situation of human rights in the Palestinian territories, highlighting his commitment to human rights irrespective of his Jewish descent. Professor Falk is also the Chairman of the Board of the Nuclear Age Peace Foundation. A prolific scholar, he is the author or co-author of over 20 books and the editor or co-editor of another 20 volumes focused on promoting global democratic governance reform and institution building.

Daniel Archibugi - Member of the Advisory Board

Daniel Archibugi is a world-renowned expert in global governance and global constituent processes, specialising in the development of accountable global institutions. His extensive academic career includes teaching positions at prestigious universities such as Sussex, Naples, Cambridge, Rome Sapienza, Rome LUISS, the London School of Economics and Political Science, and Harvard.

Daniel holds the Professor of Innovation, Governance, and Public Policy position at the University of London, Birkbeck College, School of Business, Economics, and Informatics. He advises prominent international organisations, including the European Union, Council of Europe, OECD, and several UN agencies. Additionally, he has contributed significantly to research as the Research Director at the Italian National Research Council (CNR) in Rome. He is the author of numerous books on global governance and UN reform.

Alexandre Horvath - Member of the Advisory Board

Alexandre Horvath is a seasoned IT security executive with over 15 years of C-level experience, specialising in developing and implementing robust security strategies. He serves as the Chief Information Security Officer at Cryptyx AG, a leading Swiss IT security consultancy.

Previously, Alexandre was the Chief Information Security Officer at DHL Suisse and the Cyber Risk Engineering Global Practice Leader at Allianz Suisse. His extensive career also includes roles as Managing Consultant IT Security at Detecon International, Credit Suisse, and PwC. He holds a master's degree in IT engineering from the ZHAW School of Engineering and over 15 international IT security degrees and certifications.

Sunil Abraham - Member of the Advisory Board

Sunil Abraham is the Founder and former Executive Director of the Center for Information and Society, India’s foremost digital civil rights NGO. His leadership has been instrumental in advancing digital rights and advocacy efforts in the region.

Previously, Sunil served as the director of the International Open Source Network, a project of the United Nations Development Programme’s Asia-Pacific Development Information Programme, providing valuable support to 42 countries in the Asia-Pacific region. He also held the position of Public Policy Director - Data Economy and Emerging Tech at Meta India, where he contributed to shaping policies and strategies in the digital domain.

Elizabetta Tranta - Member of the Advisory Board

Elizabetta Tranta holds distinguished credentials as a Former Minister of Defense of the Republic of Italy and a Former Security and International Relations Professor at Link Campus University of Rome. With a career marked by exemplary service and scholarly contributions, she brings a wealth of expertise to national security and international relations.

In addition to her tenure in ministerial leadership, Elizabetta Tranta brings over 20 years of extensive experience as a senior intelligence and security analyst. Her illustrious career encompasses significant roles within the Italian Ministry of Foreign Affairs, where she played a pivotal role in shaping strategic initiatives. Furthermore, Elizabetta Tranta has collaborated with several esteemed think tanks, leveraging her expertise to foster international cooperation and address complex security challenges on a global scale.

Toufi Saliba - Member of the Advisory Board

Toufi Saliba is the President of the Decentralized AI Alliance and the esteemed Global Chair for International Protocols for AI standards at IEEE, the world's largest technical professional organisation. With a remarkable track record in the convergence of decentralised digital technologies, artificial intelligence, and democratic governance, Toufi is recognised as a leading authority in the field.

In addition to these pivotal roles, he serves as the CEO of Toda.Network, a decentralised network protocol poised to revolutionise digital infrastructure. Furthermore, Toufi leverages his expertise as CEO of PrivacyShell, providing cybersecurity advisory services to large organisations and nurturing cybersecurity startups towards success.

Paolo Lezzi - Member of the Advisory Board

Paolo Lezzi is a renowned expert in intelligence-grade cybersecurity, distinguished by his role as the Founder and CEO of In The Cyber. This esteemed company stands at the forefront of providing state-grade lawful spyware systems, solidifying Paolo’s position as a pioneer in the industry. Notably, in 2019, In The Cyber acquired the Italian Hacking Team, a globally recognised leader in spyware technology.

In addition to his entrepreneurial endeavours, he serves as the Chairman of Conferenza Nationale sulla Cyberwarfare, Italy's premier conference dedicated to cyber warfare. Furthermore, Paolo holds the esteemed position of Executive Vice President at the European Center for Advanced Cybersecurity (EUCACS), contributing to the advancement of cybersecurity initiatives on a continental scale.

Violet Abthani - Member of the Advisory Board

Violet Abthani is a seasoned Silicon Valley-based senior executive, esteemed founder, and dedicated activist for peace and democracy through the transformative power of information technology and artificial intelligence. Serving as the CEO of Platonic Holdings and as a Co-founder of Enya Labs, she is at the forefront of driving innovation and positive change in the tech industry.

Before these roles, Violet held the prestigious position of Chief Operating Officer and Co-founder at Boba Network. Boba Network emerged as a leading level-2 platform, specialising in secure computing over the Ethereum blockchain. Violet's visionary leadership played a pivotal role in establishing Boba Network as a trailblazer in decentralised computing solutions.

Marta Jastrzębska - Member of the Advisory Board

Marta Jastrzębska is an asset management professional with a deep understanding of traditional and non-traditional investment strategies. She began her career in 2007 at Charlemagne Capital, an emerging markets specialist later acquired by Fiera Capital, a global firm managing C$160 billion in assets.

Marta's extensive client relations experience spans global markets, including Europe, Asia, MENA, and North America, covering a diverse investor base. Additionally, Marta became a member of Fiera’s Global Sustainable investing team, serving as Fiera Europe’s representative and advocated for Women’s empowerment in the workplace. She holds an MSc in Finance from the London School of Economics, specialising in portfolio management and an MA in Law from Wrocław University, specialising in Constitutional Law and Human Rights.

Scientific Advisory Board

Roberto Gallo - Member of the Scientific Board

Roberto Gallo holds the CEO and Chief Scientist positions at Kryptus, a leading Brazilian strategic defence IT company. Additionally, he serves as the President of the Brazilian Defense Industry Association, showcasing their influential leadership in the defence sector.

As a renowned IT security expert in Latin America, Roberto has significantly contributed to the field. Noteworthy accomplishments include designing the hardware security architecture of the Brazilian voting machines (T-DRE, Urna Eletrônica), overseeing the development of the ASI-HSM, and pioneering the creation of the SCuP, the first Secure Microprocessor in the southern hemisphere.

Koen Maris - Member of the Scientific Board

Koen Maris serves as the Cybersecurity Lead of PwC Luxembourg, overseeing the entirety of PwC's cybersecurity practice within the Luxembourg region. In addition, he spearheads the organisation's annual EU-wide cybersecurity startup competition, showcasing his commitment to fostering innovation and collaboration within the cybersecurity ecosystem.

In his prior role as Cybersecurity CTO for the ATOS Group in the Benelux and the Nordics, Koen was pivotal in delivering military-grade cybersecurity solutions to top-tier organisations. Notably, his contributions extended to serving esteemed clients such as the European Defence Agency, demonstrating his expertise in catering to the unique security needs of high-profile entities.

Gerhard Knecht - Member of the Scientific Board

From 2007 to 2019, Gerhard Knecht held the esteemed positions of Chief Information Security Officer and Global Head of Information Security Services at UNISYS. During his tenure, UNISYS solidified its position as a renowned global IT consultancy, boasting 20,000 employees and generating $3 billion in revenue annually. Gerhard played a pivotal role in driving UNISYS' strategic positioning as a leading provider of IT security solutions and services, anchored around their original Zero Trust approach.

As a visionary leader, he championed UNISYS' adoption of the Zero Trust philosophy, which advocates the principle of “never trust, always verify.” Through his strategic guidance, UNISYS successfully aligned its offerings with this innovative approach, enhancing its reputation as a trusted partner in IT security solutions.

Reinhold Wochner - Member of the Scientific Board

Reinhold Wochner is the Group Chief Information Security Officer (CISO) at Delivery Hero, a prominent global leader in the food delivery industry. In this capacity, he leads the organisation’s cybersecurity initiatives and ensures the protection of sensitive data and systems across Delivery Hero's operations worldwide.

Prior to his role at Delivery Hero, Reinhold held distinguished positions as the Group Chief Information Security Officer of Raiffeisen International Bank and as the Group Chief Security Officer of Erste Bank. These roles placed him in charge of cybersecurity and security operations within two of the largest Austrian and Eastern European banking groups, underscoring his expertise in safeguarding critical assets in highly regulated environments.

Jovan Golic - Member of the Scientific Board

Jovan Golic is a world-renowned cryptographer who is celebrated for his significant contributions to cybersecurity. During his distinguished career spanning decades, he served as the Senior Technical Leader of the Security Lab at Telecom Italia, where he played a pivotal role in shaping the company's security strategy and initiatives.

In addition to his role at Telecom Italia, Jovan served as the Action Line Leader for Privacy, Security & Trust within the esteemed EU EIT ICT Labs. This position, situated within one of the six action lines of the €3 billion EIT Digital program, emphasised his leadership in driving innovation and fostering trust in ICT solutions within the European Union.

Sanusi Drammeh - Member of the Scientific Board

Sanusi Drammeh is a seasoned Cybersecurity specialist who serves as an anchor for the Gambia Government, bringing over 15 years of invaluable experience to his role. Currently, he holds the esteemed position of Director of Cybersecurity at the Ministry of Communications & Digital Economy in Gambia. He fulfils various responsibilities in this capacity, from providing policy advisory and formulation to executing practical technical functions, demonstrating his multifaceted expertise in the field.

He possesses a diverse skill set and a keen interest in various aspects of ICT and Cyber Policies. His expertise extends to Project Management, Personal Data Protection, Privacy, and the combatting of Cybercrime. Moreover, Sanusi has represented the Government of Gambia in numerous domestic and international forums, including the AU, ECOWAS, UNCTAD, GFCE, COE, and IEC. His educational background includes a Master's of Engineering degree in Cyber Security from the University of Maryland, College Park, USA, as well as a Diploma in the

International Program on Cyber Security Studies from the George C. Marshall European Center for International Studies, The College of Garmisch-Partenkirchen, Germany.

Ecosystem

Introduction

We are immensely proud to collaborate with distinguished partners from the realms of science, academia, and other globally recognised advisory entities. Our partnerships are foundational to our success, enriching our work with expert insights and cutting-edge research. We consider ourselves fortunate to have the support and collaboration of such reputable organisations, which not only enhance our capabilities but also affirm our commitment to excellence.

Research & Development Partners - Governance

European Organisation for Security (Belgium)

The European Organisation for Security, based in Belgium, is a strategic institution supported by its members, dedicated to research, dissemination, and strategic initiatives. It represents a broad coalition of leading European entities in IT security, including major providers, research institutions, universities, clusters, and associations such as Thales, Almaviva, Atos, CEA, Fraunhofer, Engineering, Airbus, Indra, Saab, and STM. The organisation's role involves contributing to the analysis and formulation of recommendations from the standpoint of major IT industry actors. Additionally, it focuses on networking and dissemination activities targeted at similar stakeholders.

Federal Chief Information Officer of Austria (Austria)

The Federal Chief Information Officer (CIO) of Austria, a role held by Reinhard Posch since 2001, operates directly under the auspices of the Austrian Chancellor. The CIO oversees all aspects of Digital Austria and e-government initiatives nationwide. This includes leading the "Digital Austria ICT Board," which establishes the legal and technical frameworks necessary for e-government and coordinates the planning and development of e-government solutions among the federal government, provincial governments, and local authorities. Additionally, the CIO is the Director General of A-SIT, coordinating Austria's contributions to SOGIS and leading the nation's involvement in crucial cybersecurity standardisation and certification activities.

ISCOM - Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione (Italy)

ISCOM, the Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione in Italy, is a General Directorate within the Italian Ministry of Economic Development. It oversees OCSI, the Italian Scheme for ICT Security Certification, which is the official Italian member of both the Common Criteria Recognition Arrangement (CCRA) and the Senior Officials Group for Information Systems Security Mutual Recognition Agreement (SOG-IS MRA). As an authorising member of these agreements, OCSI issues ICT security certifications that are recognised both in Europe and globally. OCSI primarily focuses on certifying ICT products and systems for civil use. Additionally, ISCOM houses Ce.Va, a specialised evaluation center laboratory that assesses ICT products and systems handling classified data.

Data Protection Authority of the State of Schleswig-Holstein (Germany)

The Unabhängiges Landeszentrum für Datenschutz (ULD), or Independent Centre for Privacy Protection, serves as the Data Protection Authority for Schleswig-Holstein, the northernmost federal state of Germany. Headquartered in Kiel with a staff of 40 employees, ULD is led by Marit Hansen, the Privacy Commissioner of Schleswig-Holstein. The authority oversees freedom of information and data protection for private and public sector entities located within Schleswig-Holstein.

Lombardia Informatica (Italy)

Lombardia Informatica, established in December 1981, is a public capital service company initiated by the Regional Government of Lombardia (Regione Lombardia) in Italy. The company employs approximately 630 individuals and generates a turnover of around 200 million euros. The primary mission of Lombardia Informatica is to enhance the productivity of the regional system and improve the quality of life for citizens, as well as to boost the competitiveness of local businesses through innovative Information Technology solutions. As the IT partner for Regione Lombardia, Lombardia Informatica is responsible for designing and implementing ICT systems for the regional government and acts as the sole interface between Regione Lombardia and the marketplace. The LISPA team possesses comprehensive expertise in delivering public services, with extensive experience in managing complex services, including critical privacy and security services in the fields of e-government and e-health. This experience ensures they can effectively manage pilot sites involving citizens and public employees.

The Secure Information Technology Center of Austria A-SIT (Austria)

The Secure Information Technology Center of Austria (A-SIT) is Austria's premier public body for IT standardisation and certification. It collaborates with or represents, Austrian public authorities in numerous international and EU bodies, including the Council of Europe, ENISA Management Board, Common Criteria Management Board, SOG-IS, and the OECD. A-SIT's members comprise the Austrian Federal Ministry of Finance (BMF), the Central Bank of the Republic of Austria (Oesterreichische Nationalbank, OeNB), the Federal Computing Centre of Austria (BRZ), and Graz University of Technology (TU Graz).

Additionally, A-SIT fulfils several formal responsibilities: it is the Competent Authority responsible for certifying online collection systems for the European Citizen Initiative under EU Regulation 211/10, Article 6(4), conducts security assessments of technical components used in e-voting for student union elections, and provides expert opinions for the Data Protection Commission. Following a Cabinet Council decision, Austrian federal ministries are directed to consult A-SIT for research inquiries or issues related to its mission, positioning A-SIT as a national ICT security advisory body, functioning as an association rather than a conventional agency.

Municipality of Barcelona (Spain)

The Municipality of Barcelona, the capital of the autonomous community of Catalonia in Spain, is the country's second-largest city, with a population of 1.6 million within its administrative boundaries. As the capital, Barcelona houses the Catalan government, known as the Generalitat de Catalunya, which includes the executive branch, the parliament, and the Supreme Court of Catalonia. The city has a distinguished history of leadership in both government and e-government practices within the European Union, particularly focused on enhancing citizen autonomy.

Research & Development Partners - Science

COSIC - KU Leuven (Belgium)

The COSIC research group, under the Department of Electrical Engineering-ESAT at KU Leuven, is directed by Professor Bart Preneel. This group is recognised globally for its expertise in digital security and is committed to pioneering innovative security solutions. Their research impacts a wide array of application domains, including electronic payments, communications, identity verification through ID cards, electronic voting, the protection of electronic documents, smart home appliances, automotive telematics, and trusted systems.

Role: The primary responsibilities include designing the cryptographic infrastructure for the specified architecture and leading the analysis and recommendation of enhanced assurance assessment methods and governance strategies.

Applus+ Laboratories (LGAI Technological Center S.A) (Spain)

Applus+, a leader in the testing, inspection, certification, and technological services sector, performs these activities across over 25 industries, including ICT technologies. As the first Spanish multinational in the certification industry and ranked ninth globally, Applus+ boasts a presence on five continents and is organised into four divisions: Applus+ Laboratories, Applus+ IDIADA, Applus+ Auto, and Applus+ Energy & Industry. Applus+ Laboratories will be engaged in this project. The company generates an annual turnover of approximately 40 million euros and employs around 400 highly skilled professionals across various knowledge domains. Applus+ is actively involved in research and development, contributing to numerous national and international projects such as Adv-SCA, Adv-FI, MalpApp, Mobile, Ctless-Tool, NFC-DCC (MICINN-INNPACTO), SINTONIA (CENIT), SMARTCARD, LI-MASH (MICINN), eCID and TRATAMIENTO 2.0 (MITyC-Plan Avanza), TS-CIMONHET (CATRENE), IMPACT-EMR (ENIAC), NET-EMC (Eurostars), COSY3D (Euripides), and COOPERS (FP6).

Role: The key expert in standard setting and certification processes.

TUBITAK BILGEM Cyber Security Institute (SGE) (Turkey)

The Cyber Security Institute (SGE), initially founded in 1997 as a network security department, was formally established as a standalone institute in 2012. It is one of the six institutes within the Informatics and Information Security Research Center (TUBITAK BILGEM) of Turkey’s Scientific and Technological Research Council (TUBITAK). TUBITAK is the premier institution for managing, funding, and conducting research in Turkey and also provides advisory services to the Turkish government on science, technology, and research issues. The Cyber Security Institute focuses on systems security projects predominantly for the public sector. The Institute emphasises applied research, ensuring that the results are immediately actionable, and maintains robust ties with both public and private sector stakeholders in Turkey.

Role: As the leading national technical authority on high-assurance information technology for public sector organisations, responsibilities include strategy development, analysis, specification, and implementation contributions.

Genode Labs GmbH (Germany)

The organisation in question is a German SME that specialises in the development of highly secure operating systems (OS). As the principal architect of the Genode OS Framework, this company champions an open-source OS technology that effectively manages highly dynamic workloads while ensuring optimal security, robustness, and scalability. Notably, it supports compatibility with seL4, the world's only formally verified OS/kernel. The integration of seL4 with the Genode framework and the SCuP SoC creates an unparalleled platform in terms of security, scalability, and verifiability.

Diverging from traditional high-assurance systems, Genode is founded on a fully open and transparent development approach. Established in 2008, Genode Labs has maintained its independence, owned exclusively by its founding members. The diverse community engaged with Genode spans individuals, small to medium enterprises, government bodies, and research departments within multinational corporations.

Role: The company's contribution to the project encompasses providing the foundational operating system and essential software components.

DFKI German Research Centre for Artificial Intelligence (Germany)

The Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) is presently the largest research centre globally specialising in Artificial Intelligence (AI) and its applications, as measured by both staff size and external funding volume. DFKI's consortium of shareholders includes prominent corporations such as Deutsche Post, Deutsche Telekom, Google, Microsoft, SAP, BMW, Intel, and Daimler.

Role: The primary responsibilities include conducting thorough analyses, defining requirements, and making strategic recommendations concerning critical cyber-physical systems. The role also involves assessing the medium—and long-term impacts on AI control, safety, and security. Additionally, this position will oversee the laboratory validation of the CivicCPS, which aims to enhance the security of mobile autonomous systems operating in human-centric environments.

SCYTL Secure Electronic Voting S.A. (Spain)

SCYTL Secure Electronic Voting S.A., based in Spain, is a global leader in electronic voting and high-assurance remote deliberations technologies. The company is dedicated to offering solutions for electoral modernisation that feature the highest security standards in the industry. Scytl's cryptographic measures are supported by over 40 international patents, which play a crucial role in safeguarding the privacy and integrity of sensitive electoral data.

Role: Responsible for prototyping software application layers, with additional responsibilities to be determined.

EMAG Institute of Innovative Technologies (Poland)

The Institute of Innovative Technologies EMAG is a leading research and development entity across various market sectors, including information security management systems, business continuity systems, risk management systems, natural hazards monitoring, and automation and measurement systems. Employing a robust team of 138 professionals, the Institute features 42 scientists and 74 engineers and technicians. EMAG Institute has secured nearly 500 patents, over 80 protection rights, and 6 trademarks, and has received numerous awards at various competitions and fairs. The research team at the Institute possesses extensive knowledge and experience in developing risk assessment methodologies and tools for diverse application domains, such as critical infrastructures and transport utilities.

Role: Responsible for analysing and optimising certification processes, leveraging new IT-enabled methodologies.

Delft University of Technology (The Netherlands)

The Parallel and Distributed Systems Group (PDS) at Delft University of Technology, the oldest, largest, and most comprehensive technical university in the Netherlands, enjoys a prestigious reputation both nationally and internationally. The university brings together over 19,000 students and 2,500 scientists, including 400 professors. The PDS group at TU Delft has a distinguished 15-year history in the design, implementation, deployment, and analysis of peer-to-peer (P2P) systems. Notably, it developed the BitTorrent-based P2P client, Tribler, which features enhanced functionalities such as support for video on demand, live streaming, channels, information dissemination protocols, and a reputation system.

Role: Responsible for designing the P2P and mixed network layers of the target architecture. This position will also lead the analysis and formulation of related recommendations.

Kryptus (Brazil)

Kryptus, a Brazilian enterprise, possesses unique global expertise in secure hardware design and system integration. The company has crafted pivotal technologies including Brazil's 400,000 voting machines, fighter-to-fighter communication systems, and the Hardware Security Module (HSM) for the core Root CA of Brazil's principal PKI. It also pioneered the development of the first secure general-purpose CPU microprocessor in the Southern Hemisphere, the SCuP, which operates between 100-300Mhz and features open and verifiable designs alongside Free/Libre and Open Source Software (FLOSS) microcode. This innovation forms the cornerstone of the CivicIT hardware architecture.

Role: This role is responsible for the design and prototyping of critical hardware, including the CPU and SoC of the target architecture. Additionally, it entails contributing to the analysis and formulation of recommendations for enhancing assurance assessment methods.

TECNALIA Research & Innovation (Spain)

TECNALIA Research & Innovation, based in Spain, is a premier private, independent, nonprofit applied research centre recognised internationally for its excellence. As Spain's leading private and independent research and technology organisation and one of the largest in Europe, TECNALIA employs 1,319 staff members, including 198 PhD holders, and reported an income of €94 million in 2014. Its ICT unit holds considerable expertise in the assurance and certification of ICT across various domains. Within the framework of Horizon 2020, TECNALIA has engaged in 87 projects, leading 17 of them through December 2015. It is also a member of EARTO and EUROTECH, connecting major European research centres.

Role: The position involves contributing to the collection and analysis of assurance guidelines and certification schemes, especially concerning critical infrastructures and autonomous cyber-physical systems such as drones and vehicles. This includes collecting and examining assurance guidelines, standards, and certification schemes, as well as analysing requirements and formulating recommendations for assurance models and related certification schemes.

EIT Digital

The Privacy, Security, and Trust Action Line represents one of eight Thematic Action Lines operated by EIT Digital. EIT Digital is one of the five Knowledge and Innovation Communities established by the European Institute of Innovation and Technology (EIT).

ECSEL JU

ECSEL JU (Electronic Components and Systems for European Leadership Joint Undertaking) focuses on Key Enabling Technologies that are integral across all industrial sectors and influence virtually every facet of life. These technologies form the foundational infrastructure for the Internet and are essential for the functionality of portable phones, tablets, and other digital devices.

Association for European NanoElectronics Activities

The Association for European NanoElectronics Activities offers unmatched networking opportunities, policy influence, and facilitated access to funding within the domain of micro- and nano-electronics enabled components and systems.

University of Luxembourg – Interdisciplinary Centre for Security, Reliability and Trust - SnT (Luxembourg)

The University of Luxembourg's Interdisciplinary Centre for Security, Reliability, and Trust (SnT) engages in internationally competitive research in the fields of information and communication technology (ICT), producing work of high socio-economic impact. Beyond its commitment to long-term, high-risk research endeavours, SnT actively participates in demand-driven collaborative projects with industry and public sector partners. To effectively address the strategic challenges faced by these sectors in ICT, SnT has established a Partnership Program which now includes 32 members.

Since its inception in 2009, SnT has experienced significant growth, attracting top scientists and initiating more than 50 projects under EU and ESA auspices. The centre has also developed a technology transfer office (TTO) to manage, protect, and license intellectual property, and has launched four spin-offs. These efforts contribute to a vibrant, dynamic, and interdisciplinary research environment encompassing 260 individuals. These initiatives not only enhance Luxembourg's competitive edge but also extend its impact beyond national borders.

Inria Rennes – Bretagne Atlantique – Décentralisé Team (France)

Inria Rennes – Bretagne Atlantique – Décentralisé Team, based in France, is part of a French public research organisation founded in 1967, exclusively devoted to computational sciences. In 2015, a new high-security laboratory was established in Rennes, focusing on developing secure systems. The laboratory specialises particularly in peer-to-peer (P2P) and privacy-enhancing network protocols. The team is led by Christian Grothoff.

Goethe University – Deutsche Telekom Chair for Mobile Business and Multilateral Security (Germany)

Goethe University's Deutsche Telekom Chair for Mobile Business and Multilateral Security, based in Germany, is at the forefront of research concerning privacy and security within innovative mobile networks, including their related social and economic implications. The chair leads several significant EU research and development projects focused on privacy enhancements, such as ABC4trust, TresPass, and PrivacyOS. The chair is held by Prof. Kai Rannenberg, who is also a member of the NIS Platform for Individual Rights, underscoring his expertise and commitment to advancing individual rights in the digital realm.

American Mini Foundry (USA)

American Mini Foundry (USA) is a dormant yet leading startup specialising in ultra high-assurance IC foundry oversight. The company holds unparalleled world-class expertise in hardware design and fabrication assurance processes. Key management team members involved include President Scodden and Gerry Etzold, who served as the former Technical Director of the NSA Trusted Foundry Program from 2008 to 2009.

23. Other Publications, Articles and Posts

- *Can a global version of the 1786 Annapolis Convention lead to the governance we need for AI?. (*A March 2024, 1200-word blog post)
- *How a public-private consortium could lead to democratic global AI governance*. (A March 2024, 900-word opinion piece the president of the Trustless Computing Association published last March 13th on *The Yuan,* a prestigious Chinese digital and AI policy Journal. It frames our Initiative vis-a-vis global AI supply chains, OpenAI's "$7 trillion AI plan", and the pursuit of an effective, democratic and safe global governance of AI).
- A 33-page Harnessing AI Risk Proposal v.3 PDF, (published January 2024, on *ResearchGate*). It details the Initiative, its rationale, the design of the constituent processes, and the preliminary designs of the IGO and its agencies. It sets an initial framework for the Initiative's co-design with advisors, partners and Summit participants. (*Harnessing AI Risk Proposal v.2* (Oct 2023, 6500-word paper, published on ResearchGate pdf) and *Harnessing AI Risk Proposal v.1* (June 2023, published as Linkedin and blog post)
- A 14-page Grant Proposal and Roadmap 2024-2027 of the Initiative PDF. (March 10th, 2024). Includes 2-page summary.
- A February 2024, 8-page Case for family offices to support and invest in a Global Public Benefit AI Lab and an International AI Safety Agency.
- A December 2023, 700-word blog post, The AI Act and Beyond: EU's Ambitions and Obstacles in the AI Race. It frames our Initiative vis-a-vis EU AI Act and EU AI capacity-building initiatives.
- An October 2023, 3000-word blog post, Towards an Open Transnational Constituent Assembly for AI and Digital Communications.
- For further details about the foreseen ***IT Security Agency*****,** in addition to *Harnessing AI Risk Proposal* (v.3) above, see our Trustless Computing Certification Body and Seevik

Net Initiative (1-pager + 45-pager pdf) and details of our traction so far with over 13 nation-states (1-pager + 32-pager pdf).