Interim Report:

Governing AI for Humanity

December 2023

United Nations

Interim Report: Governing AI for Humanity

This report is published by the
Advisory Body on Artificial Intelligence

For more information, contact the
AI Advisory Body Secretariat:
aiadvisorybody@un.org

United Nations

An enormous category error corrodes many (but by no means most!) of the suggestions. AI is not some unitary actor. It is not an actor, and it is not unitary. It is a set of loosely related programming techniques, plus a variety of digital services that each deploy some aspects of those programming techniques. Sentences like "The full impacts of AI are not yet known" are nonsensical, because impacts are determined both by application areas and by the specific systems architecture chosen to implement those applications. We can expect ongoing innovation and extension of applications, the impacts of which will have to be individually checked as with all new products. We might hope to make architectural recommendations that will have some resilience and uptake, such as avoiding unnecessary storage of data or excessive use of energy.

This enormous flaw in understanding could be addressed relatively easily and without very excessive modifications if instead of making this article about governing AI, we made it about governing in an age of AI. Many of the core motivating issues here addressed go well beyond the application of AI into wider matters of justice, rights, and sustainability. Yet AI could help us more successfully address those issues.

One of my favourite aspects or the document is the outstanding work in box 1, on all the ways AI could be used to resolve the climate crisis. Yet AI itself is presented as unknowable and almost ungovernable. The fact is that AI is a lot less complicated than climate, and no more integrated into our economy. The same kinds of effort that has gone into rethinking how we address climate if we leverage AI should be spent rethinking how we govern. I would love to see this as a second box (maybe replacing the second one, which I didn't find very well thought out.)

The other overarching issues is that too much of the text serves the interests of big tech and surveillance at the cost of diversity, resilience, and innovation in governance, and at the cost of national sovereignty. While I agree the UN may be uniquely positioned to help negotiate the really hard and essential matters such as transnational redistribution of value due to IP and data originators, or indeed wealth taxes on billionaires, in general I would prefer to see more respect for diversity of jurisdictions.

add paragraph here about costs etc.

Nevertheless, I do believe we also need some innovations of governance, such as how to govern the transnational utilities many digital services have become. Unfortunately, theory development on utilities seems to have stopped in about 1980, as Chicago School economics cashed in on the plateauing of the soviet economy, and with it the then-perceived economic failure of communism. China has since shown us that more governance innovation is possible, but China shares with the West many of the concerns addressed here. All governments benefit from legitimacy and trust, and are challenged and disrupted by excessive inequality.

Having said that, it is not evident to me that these innovations must be operated from some UN or global body. Rather, the UN might help coordinate... stuff about incentives from below.

Other pitfalls as learnt from EU experience (from below)

Table of Contents

Introduction

1 Artificial intelligence (AI)[1] increasingly affects us all. Though AI has been around for years, capabilities once hardly imaginable have been emerging at a rapid, unprecedented pace. AI offers extraordinary potential for good — from scientific discoveries that expand the bounds of human knowledge to tools that optimize finite resources and assist us in everyday tasks. It could be a game changer in the transition to a greener future, or help developing countries transform public health and leapfrog challenges of last mile access in education. Developed countries with ageing populations could use it to tackle labour shortages.

2 Yet, there are also risks. AI can reinforce biases or expand surveillance; automated decision-making can blur accountability of public officials even as AI-enhanced disinformation threatens the process of electing them. The speed, autonomy, and opacity of AI systems challenge traditional models of regulation, even as ever more powerful systems are developed, deployed and used.

3 The opportunities and the risks of AI for people and society are evident and have seized public interest. They also manifest globally, with geostrategic tensions over access to the data, compute, and talent that fuel AI, with talk of a new AI arms race. Nor are the benefits and risks equitably distributed. There is a real danger, even if humanity harnesses only the positive aspects of AI, that those will be limited to a club of the rich. Today's AI benefits are accruing largely to a handful of states, companies, and individuals.

4 This technology cries out for governance, not merely to address the challenges and risks but to ensure we harness its potential in ways that leave no one behind. A key measure of our success is the extent to which AI technologies help achieve the Sustainable Development Goals (SDGs). As an example, Box 1 illustrates AI's potential in tackling climate change and its impact (SDG 13).

5 The High-level Advisory Body on AI was formed to analyse and advance recommendations for the international governance of AI. We interpret this mandate not merely as considering how to govern AI today, but also how to prepare our governance institutions for an environment in which the pace of change is only going to increase. AI governance must therefore reflect qualities of the technology itself and its rapidly evolving uses — agile, networked, flexible — as well as being empowering and inclusive, for the benefit of all humanity.

guardrails

6 Our work does not take place in a normative or institutional vacuum. The UN is guided by rules and principles to which all of its member states commit. These shared and codified norms and values are the lodestar for all of its work, including AI governance. Norms including commitments to the UN Charter, the Universal Declaration of Human Rights, and international law including environmental law and international humanitarian law, are applicable to AI. Institutions created in support of multilateral objectives from peace and security to sustainable development have roles to play in cultivating the opportunities while safeguarding against risks.

7 Nonetheless, we share the sense of urgency held by complementary governance initiatives on AI, including those by states as well as regional and intergovernmental processes such as the EU, the G7, the G20, UNESCO, and the OECD, among others. More inclusive engagement is needed, however, as many communities — particularly in the Global South or Global Majority — have been largely missing from these discussions, despite the potential impact on their lives. A more cohesive, inclusive, participatory, and coordinated approach is needed, involving diverse communities worldwide, especially those from the Global South or Global Majority.

8 The United Nations holds no panacea for the governance of AI. But its unique legitimacy as a body with universal membership founded on the UN Charter, agreed universally, as well as its commitment to embracing the diversity of all peoples of the world, offer a pivotal node for sharing knowledge, agreeing on norms and principles, and ensuring good governance and accountability. Within the UN system, plans for the Global Digital Compact and the Summit of the

1 Per OECD definition: https://oecd.ai/en/wonk/ai-system-definition-update

Box 1: Case study illustrating how AI can help address climate change

The critical intersection of climate change and AI opportunity – a case study:

Climate change represents a global and universal challenge – one where a collective response requires sustainable digital transformation, thoughtfully designed new infrastructure, and the ability to deliver precise decision making at scale. AI-driven approaches are particularly well suited to this challenge, integrating key developments in machine learning, large language models, high quality data analysis, and more, to create new capacities.

Information that describes disconnected and disparate phenomena – from geospatial imaging, distributed sensors, real time monitoring, and citizen-reported data on effects of hyperlocal climate change – can be used to create new understanding of inputs, consequences, and the complex systems which drive climate outcomes. Taken together with predictive systems that can transform data into insights and insights into actions, AI-enabled tools may help develop new strategies and investments to reduce emissions, influence new private sector investments in net zero, protect biodiversity, and build broad-based social resilience. This can apply to other SDGs.

The following is a non-exhaustive list of early promises of AI helping to address climate change:

- Assigning responsibility for climate action to national and subnational governance institutions by creating new and highly granular predictive resources for climate investment. For example, real time heatmaps of storm-related urban flooding to unlock hyperlocal infrastructure improvements in sewer and drainage systems.
- Building public, open-source data and AI systems to move private sector net zero reporting from a static compliance function to a public facing, real time data repository to increase trust, transparency, and accountability for public commitments.
- Using advanced climate modelling tied to information about urban mobility and behaviour patterns to create new early warning systems, allowing for more effective delivery of post conflict/disaster relief and recovery.
- Developing evidence-based AI interventions in open system and other carbon removal technologies where high uncertainty intervals can limit crucial early-stage investment. Advanced modelling techniques can lower the cost of scientific inquiry and allow for rapid prototyping of novel solutions.

This is both doable AND harder than assigning responsibility for AI systems!

But structural barriers remain to help these technologies reach the scale required to match the scope of the climate crisis and meet the diverse needs of the many critical stakeholders in the climate fight including corporations, governments, activists, civil society, and others. Systemic risks such as algorithmic bias, transfer context bias, interpretation bias, representation and allocation harms would have to be considered. Some actions to overcome barriers include:

- Improving model explainability and trust in order to increase adoption of AI-produced insights into critical climate decision making.

The climate and its interaction with our economy is more complex than AI and its.

why replicate with public resource what is available privately? Why not instead treat digital services where appropriate as utilities?

- Ensuring that AI models are trained on diverse, truly representative datasets, which reflect both commercially viable data collected by for-profit entities and data which "fills in the gaps" funded by nonprofit, philanthropic, and government resources and complements local tacit knowledge.
- Providing communities impacted by climate change vulnerabilities access to AI-generated predictions that would otherwise only be provided to private companies.
- Lowering cost of compute and machine learning expertise so that nonprofits and civil society can build and sustain free and open AI products.
- Overcoming siloed action from multiple organizations building proprietary solutions / holding proprietary data to compete for private or philanthropic investment.
- Financing for scaling such solutions

The cross-domain connection between AI and frontline experience of climate change is critical to enabling these transformative approaches. Solutions which exist in a technical silo – even when enabled with compute, data, and talent – face significant challenges in uptake and distribution when they do not reflect the lived experience of community members and local decision makers.

For each of the opportunities described above, critical early inputs from non-technical stakeholders need to inform project conception, design, execution, and integration. Enablers therefore require a values-based approach that prioritizes community interests, a combination of technical and problem-based expertise, and a comprehensive approach to new AI development. We also need to keep an eye on the potential negative impact of AI on climate change because of the associated energy and water consumption.

Future in September 2024 offer a pathway to timely action.

9 The Advisory Body comprises individuals diverse by geography and gender, discipline and age; it draws expertise from government, civil society, the private sector, and academia. Intense and wide-ranging discussions yielded broad agreement that there is a global governance deficit in respect of AI and that the UN has a role to play.

10 In this report, we first identify opportunities and enablers that can help harness the potential benefits of AI for humanity. Second, we highlight risks and challenges that AI presents now and in the foreseeable future. Third, we argue that addressing the global governance deficit requires clear principles, as well as novel functions and institutional arrangements to meet the moment. The report concludes with preliminary recommendations and next steps, which will be elaborated in our final report by August 2024.

11 Though we are confident of the broad direction, we know that we do not take this journey alone. We look forward to consulting widely on next steps to ensure that more voices and views are included, and that AI serves our common good.

This example should be applied to not only AI governance, but governance generally. We need this kind of effort and transparency so people may both see and ensure the value of their tax payments, and understand their constituitive role in their governments. Data is not really the new oil – its value is not determined solely by its quantity or quality, but rather by its informativeness, which varies by application and is not fungible. But AI really is the new paper. We are coming to write down nearly every important plan in digital and sometimes even executable format. This means really solving digital governance would solve much of the rest of governance as a side effect. Perhaps realising that this should be our ambition would help us better understand the task at hand, including the efforts to derail or disinform regulation.

The Global Governance Deficit

the extent of this concentration is sometimes exaggerated, cf Bryson & Malikova 2020; Dorfs & Bryson in prep

12 Though AI is transforming our world, its development and rewards are currently concentrated among a small number of private sector actors in an even smaller number of states. The harms are also unevenly spread. Global governance with equal participation of all member states is needed to make resources accessible, make representation and oversight mechanisms broadly inclusive, ensure accountability for harms, and ensure that geopolitical competition does not drive irresponsible AI or inhibit responsible governance.

13 The United Nations lies at the heart of the rules-based international order. Its legitimacy comes from being a truly global forum founded on international law, in the service of peace and security, human rights, and sustainable development. We believe that this offers the institutional and normative foundation for collective action in global governance of AI. Apart from considerations of equity, access, and prevention of harm, the very nature of the technology itself — AI systems being transboundary in structure, function, application, and use by a wide range of actors — necessitates a global approach.

14 Pieces of this puzzle are being filled by self-regulatory initiatives, national and regional laws, and the work of multilateral forums. Yet, gaps remain and the challenge is clear: a global governance framework is needed for this rapidly developing suite of technologies and its use by various actors, be they the developers or users of the technology. AI presents distinctly global challenges and opportunities that the UN is uniquely positioned to address, turning a patchwork of evolving initiatives into a coherent, interoperable whole, grounded in universal values agreed by its member states, adaptable across contexts.

15 The next three sections outline roles an institution or a network of institutions anchored in the UN's universal framework could play in expanding the benefits of AI and mitigating its risks, as well as the principles and functions that will best achieve these ends.

is now the time to contest this? Maybe a good lead into the below?

Note that this transboundary nature could and perhaps should be limited as demonstrated by the Chinese "great firewall". Although we want a free, open, and global Internet, we do not want services provided by illicit actors who refuse to comply to the law to be easily available in regions where they have been banned. Note too that fine structures such as the EU is now wielding assume actors that care about long term brand and legal reputation. Any belief that this will apply to all AI producers or deployers ignores the low barriers to entry for digital services, and the increasing prevalence of shell companies and the use of bankruptcy to avoid substantial accountability.

The advantage of creating a system of smart digital borders is that it maintains the sovereignty of EU members, and further, by facilitating diversity of regulatory regimes, both increases the probability of useful regulatory innovation, and decreases the likelihood of regulatory capture. Those who fear some suggestions as a "Splinternet" overlook the lessons of the EU's successes as a tradeblock. It is quite likely that a relatively small number of harmonised jurisdictions will form as the value of doing digital trade in any jurisdiction depends both on its share of the global wealth and the divergence of its regulatory context.

Opportunities and Enablers

16 AI has the potential to transform access to knowledge and increase efficiency around the world. A new generation of innovators is pushing the frontiers of AI science and engineering. AI is increasing productivity and innovation in sectors from healthcare to agriculture, in both advanced and developing economies. Alongside such growth are questions about which enablers are required to ensure benefits are spread equitably and safely across humanity, and that disruptive impacts, including on jobs, are addressed and managed. An important question for policy makers is how to grow successful AI ecosystems around the world while holding established and emerging players accountable.

Box 2: Examples of AI opportunities

Examples of AI opportunities

People-assistive AI

AI is not the actor. The previous box handled moral agency better. Also, "people-assistive" is very awkward. How about "Individual-enhancing AI", "individual support" or just "individual opportunities" parallel to the other subsections?

AI can assist people in everyday tasks as well as their most ambitious, creative and productive endeavours. People-assistive AI includes accessibility tools and improvements to education. Applications have been developed to serve as virtual assistants for people with limited vision or speech, supporting accessibility needs previously overlooked or neglected. AI-powered translation now covering over a hundred languages promotes access as well as intercultural understanding and communication. A new generation of tutoring apps promises to expand access to quality education worldwide.

Sectoral opportunities

This box is weak and inconsistent.

AI will have a greater impact in some sectors rather than others. Among the most promising are agriculture and food security, health, education, protection of the environment, resilience to natural disasters and combating climate change. For example, AI has been used to create early-warning systems for floods, now covering over 80 countries, as well as wildfires, and food insecurity. AI is being used to monitor endangered species (e.g., dolphins, whales) and to optimize agricultural practices. Within each field, there are myriad possibilities.

AI is broadening access to quality care, for example in the maternal health care space in Sub-Saharan Africa. Similarly, possibilities exist with respect to environmental problems, making education more accessible, helping ease poverty and hunger, and making cities safer.

Scientific opportunities

AI is transforming the way in which scientific research is performed and is expanding the frontier of scientific advancement, including by accelerating molecular and genomic research. AI systems show special promise for accelerating the work of scientists across many disciplines and a potential paradigm shift in the way science is practised, from helping explore new discovery spaces to automating experimentation at scale. For example, AI-powered tools that predict protein structures are being used by over a million researchers for drug discovery and to advance understanding of diseases like

Lessons from writing the EU AI Act: the point of this section is by no means that other regions should adopt the content of the EU's digital legislation – though they are welcome to, and should note also the DSA, DMA, small modifications to the product liability directive, and of course the GDPR.

tuberculosis, as well as many previously neglected diseases. In the healthcare space, AI is powering diagnostic tools to help doctors with more timely detection of various types of cancers and eye-related diseases, thereby saving lives. In the energy space, AI is playing a role in optimizing energy systems and advancing the transition to renewable energies. For example, AI has been used to boost the value of wind energy, control tokamak plasmas in nuclear fusion, and enable carbon capture. There is scope for the UN to encourage progress in AI-enabled science by focusing attention on questions worth solving for the global good.

Public sector opportunities

please mention transparency. Market forces are not really applicable here btw, well, kind of, but again, this is all awkward.

Crucially, AI may drive progress in areas where market forces alone have traditionally failed. These range from extreme weather forecasting and monitoring biodiversity, to expanding educational opportunities or access to quality healthcare, and optimizing energy systems. Governments and the public sector can improve services for citizens and strengthen delivery for vulnerable communities by leveraging AI for social good.

why differentiate this from the above? The UN is just a special case of the public sector. Maybe try to come up with ways the UN would add value?

Opportunities for the UN to harness AI

Finally, the use of AI can contribute to accelerating progress towards achieving the Sustainable Development Goals and enhance the role and effectiveness of the UN in promoting sustainable development, human rights and peace and security. For example, the UN can use AI to monitor the development of crisis situations in different parts of the world including human right abuses or for measuring progress on the SDGs. While many have noted the potential of AI to contribute to many of the 17 SDGs, many have also noted significant barriers to fully leveraging the potential of AI to help make progress. The UN and other international organizations have started to build promising AI use cases and demonstrations in areas such as prediction of food insecurity, managing relief operations and weather forecasting.

Key enablers for harnessing AI for humanity

17 The development of AI is now driven by data, compute, and talent, sometimes supplemented by manual labelling labour. Currently, only well-resourced member states and large technology companies have access to the first three, leading to a concentration of influence. In addition to global shortages of crucial hardware such as GPUs, there is also a dearth of top technical talent in the field of AI. It has been suggested that open model development may alter this dynamic, though the impact and safety of open models is still being analysed and debated.

"sometimes" applies even more to AI being driven by data than to such data-hungry AI requiring manual labelling. Disagree also that there's a dearth of global talent, though it isn't all equally recognised.

18 The AI opportunity arrives at a difficult time, especially for the Global South. An "AI divide" lurks within a larger digital and developmental divide. According to ITU estimates for 2023, more than 2.6 billion people still lack access to the Internet. The basic foundations of a digital economy — broadband access, affordable devices and data, digital literacy, electricity that is reliable and affordable are not there. Fiscal space is constrained and the international environment for trade and investment flows is challenging. Critical investments will be needed in basic infrastructure such as broadband and electricity, without which the ability to participate in the development and use of AI will be severely limited. Even outside the Global South, taking advantage of AI will require

awkward

This really sounds like you are taking "AI" to mean only generativeAI, which is an unproven (economically) and relatively unimportant subpart of all presently active intelligent automation.

efforts to develop local AI ecosystems, the ability to train local models on local data, as well as fine-tuning models developed elsewhere to suit local circumstances and purposes.

19 Access and benefits must go hand in hand. Entrepreneurs in regions lagging in AI capacity require and deserve the ability to create their own AI solutions. This requires national investments in talent, data, and compute resources, as well as national regulatory and procurement capacity. Domestic efforts should be supplemented by international assistance and cooperation not only among governments but also private sector ~~players~~. Rallying scientists to solve societal challenges could be a key enabler for harnessing AI's potential for humanity. Open-Source and sharing of data and models could play an important role in spreading the benefits of AI and developing beneficial data and AI value chains across borders.

20 Enablers ('common rails') for AI development, deployment and use would need to be balanced with 'guard rails' to manage impact on societies and communities. A litmus test will be the extent to which AI governance efforts yield human augmentation rather than human replacement or alienation as the outcome. Some AI development relies on cheap and exploitable labour in the Global South. Even in the Global North, there are questions related to valuing artistic expression, intellectual property, and the dignity of human labour. Equitable access to these technologies and relevant skills to make full use of them are needed if we are to avoid "AI divides" within and across nations.

Governance as a key enabler

21 AI can and should be deployed in support of the Sustainable Development Goals. But doing so cannot rely on current market practices alone, nor should it rely on the benevolence of a handful of technology companies. Any governance framework should shape incentives globally to promote these larger and more inclusive objectives and help identify and address trade-offs.

22 Comparisons with other sectors offer potential lessons. Mechanisms such as Gavi, the Vaccine Alliance, may suggest short-term examples for ensuring that the benefits are shared. Repositories of AI models that can be adapted to different contexts could be the equivalent of generic medicines to expand access, in ways that do not promote AI concentration or consolidation.

23 Some of these societally beneficial aspirations may be realized by advances in AI research itself; others may be addressed by leveraging novel market mechanisms to level the playing field, or by incentivizing actors to reach all communities and enable benefits to be accessible to all. But many will not. Ensuring that AI is deployed for the common good, and that its benefits are distributed equitably, will require governmental and intergovernmental action with innovative ways to incentivize participation from private sector, academia and civil society. A more lasting solution is to enable federated access to the fundamentals of data, compute, and talent that power AI — as well as ICT infrastructure and electricity, where needed. Here, the European Organization for Nuclear Research (CERN), which operates the largest particle physics laboratory in the world, and similar international scientific collaborations may offer useful lessons. A 'distributed-CERN' reimagined for AI, networked across diverse states and regions, could expand opportunities for greater involvement. Other examples of open science relevant to AI include the European Molecular Biology Laboratory (EMBL) in biology or ITER, the International Thermonuclear Experimental Reactor.

Automation doesn't undermine the dignity of human labour. Rather, it alters the value of various skills. Rapidly undermining the value of some skills implies a need for active governments to support retraining and indeed new sectoral innovation. In some countries we see that the same corporations that invest in AI also invest in retraining themselves, at least for the majority of existing employees, who find themselves with more interesting and better paying jobs at the same firm (cite). Historically, economies have done less well at sufficiently recognising and rewarding which skills have newly become valuable. We can hope that AI itself might be used to help adjust wages and educational opportunities appropriately. But again this requires investment, which the transnational nature of AI and the billionaires it helps create has so far undermined. If there is a real role for the UN it is probably in this, in ensuring adequate redistribution such that governments can meet these challenges, for example through the introduction of a global wealth tax on the most excessive individual fortunes. Within countries, we could also be reenforcing those who demonstrate egalitarian application of the UDHR, for example by ensuring that the benefits of healthcare and utilities (power, water, information) are universally accessible.

IMO it is wrong to think of AI as a special new thing that needs to be managed. Rather, it is a technological enabler that allows us to improve our management and governance overall. Very few of the problems attributed to AI are uniquely its own. Failures to address market concentration, lack of transparency in the expression of power, or indeed all the forms of bias we train into machine learning are pervasive in our data because they are pervasive in our socety.

Risks and Challenges

None of these seem unique to AI. An entirely more useful framing is that regulation of and through AI might well help humanity deal with these social ills wherever they are present. Does AI threaten language diversity more than the BBC, Le Monde, or Hollywood?

24 Along with ensuring equitable access to the opportunities created by AI, greater efforts must be made to confront known, unknown, and as yet unknowable harms. Today, increasingly powerful systems are being deployed and used in the absence of new regulation, driven by the desire to deliver benefits as well as to make money. AI systems can discriminate by race or sex. Widespread use of current systems can threaten language diversity. New methods of disinformation and manipulation threaten political processes, including democratic ones. And a cat and mouse game is underway between malign and benign users of AI in the context of cybersecurity and cyber defence.

Risks of AI

though here there are active and intelligent debates about whether such threats exist, and if so, now to assess them.

25 We examined AI risks firstly from the perspective of technical characteristics of AI. Then we looked at risks through the lens of inappropriate use, including dual-use, and broader considerations of human-machine interaction. Finally, we looked at risks from the perspective of vulnerability.

as they are sometimes presently built

26 Some AI risks originate from the technical limitations of ~~these~~ systems. These range from harmful bias to various information hazards such as lack of accuracy and "hallucinations" or confabulations, which are known issues in generative AI.

uses of

27 Other risks are more a product of humans ~~than~~ AI. Deep fakes and hostile information campaigns are merely the latest example of technologies being deployed for malevolent ends. They can pose serious risks to societal trust and democratic debate.

28 Still others relate to human-machine interaction. At the individual level, this includes excessive trust in AI systems (automation bias) and potential de-skilling over time. At the societal level, it encompasses the impact on labour markets if large sections of the workforce are displaced, or on creativity if intellectual property rights are not protected. Societal shifts in the way we relate to each other as humans as more interactions are mediated by AI cannot also be ruled out. These may have unpredictable consequences for family life and for physical and emotional well-being.

be used to

29 Another category of risk concerns larger safety issues, with ongoing debate over potential "red lines" for AI — whether in the context of autonomous weapon systems or the broader weaponization of AI. There is credible evidence about the increasing use of AI-enabled systems with autonomous functions on the battlefield. A new arms race might well be underway with consequences for global stability and the threshold of armed conflict. Autonomous targeting and harming of human beings by machines is one of those "red lines" that should not be crossed. In many jurisdictions, law-enforcement use of AI, in particular real-time biometric surveillance, has been identified as an unacceptable risk, violating the right to privacy. There is also concern about uncontrollable or uncontainable AI, including the possibility that it could pose an existential threat to humanity (~~even if~~ there are debates ~~over~~ whether and how to assess ~~such threats~~).

has that gone well or badly?

in other words, human responsibility must always be attributed, and in such a way that the individual held responsible can learn from punishment and improve future outcomes.

30 Putting together a comprehensive list of AI risks for all time is a fool's errand. Given the ubiquitous and rapidly evolving nature of AI and its use, we believe that it is more useful to look at risks from the perspective of vulnerable communities and the commons. We have attempted an initial categorization as per this approach *(Box 3)*, which will be developed further into a risk assessment framework, building on existing efforts. There will be dynamicity about risks as technology, its adoption, and use evolve. This speaks to the need to keep risks under review through interdisciplinary science and evidence-based approaches. Adaptable risk management frameworks that can be tuned as per the experience of different regions at different times would also be needed. The UN can provide a valuable space for such mutual learning and agile adaptation.

Again, it's probably more useful to ask "how have our obligations and governments changed given the advent of AI". What we can and should do for our residents and defend them against (note that I say "resident" meaning "any human within our borders" as per the UDHR. Citizenship is largely irrelevant to fundamental rights obligations.)

Box 3: Categorizing risks from the perspective of existing or potential vulnerability

AI risks from the perspective of existing or potential vulnerability

- Individuals
  - Human dignity/value/agency (manipulation, deception, nudging, sentencing)
  - Life, safety, security (autonomous weapons, autonomous cars, interaction with chemical, biological, radiological and nuclear defence)
  - Physical and mental integrity, health and safety (diagnostics, nudging, neurotechnology)
  - (other) human rights/civil liberties, e.g. fair trial (recidivism prediction), presumption of innocence (predictive policing), freedom of expression (nudging), privacy (biometric recognition)
  - Life opportunities (education, jobs, financial stability)
- Groups
  - Discrimination/unfair treatment of sub-groups, including on basis of gender
  - Group isolation/marginalization
  - Functioning of a community
  - Social equality/equity (unfair treatment of groups, including on basis of gender)
  - Children, elderly, people with disabilities
- Society
  - International and national security (autonomous weapons/disinformation)
  - Democracy (elections, trust)
  - Information Integrity (mis- or disinformation, deep fakes, personalized news)
  - Rule of Law (functioning of and trust in institutions, judiciary)
  - Security (military and policing uses)
  - Cultural diversity and shifts in human relationships (homogeneity, fake friends)
  - Social cohesion (~~filter bubbles~~, declining trust in news, ~~information~~)
- Economy
  - Power concentration
  - Technological dependency
  - Unequal economic opportunity
  - Resource distribution/allocation
  - Under-/overuse of AI, techno-solutionism

great list, and congruent with my above point.

not a bullet, maybe include in heading?

may as well and be inclusive and say "governing", since all governments now hold elections, and every government has always had to care about trust and legitimacy as these ease not only governing and leadership more generally.

bullet categories are evidently redundant.

declining local community identity and trust, loss of language or culture, corrosion of moral, ethical, or legal values

logistics / global throttle points (maybe under systems?)

- (Eco)systems
  - Stability of financial systems
  - Risk to critical infrastructure
  - Strain on environment/climate/natural resources
- Values and Norms
  - Ethical values
  - Moral values
  - Social values
  - Cultural values
  - Legal norms

I would put these under "social cohesion", but I know I'm odd.

see previous comment on jurisdictions and firewalls, I'm not sure we want global consensus on ALL of this. Rather we need to identify the things that a) truly require global near consensus e.g. wealth tax b) are facilitated by global cooperation e.g. technical standards for AI transparency, redistribution to IP holders and other data sources.

I don't understand this sentence.

31 There is not yet a consensus on how to assess or address these risks. Nevertheless, as the precautionary principle provides on environmental dilemmas, scientific uncertainty about risks should not lead to governance paralysis. Achieving consensus and acting on it requires global cooperation and coordination, including through shared risk monitoring mechanisms. International organizations have decades of relevant experience with dual use technologies, from chemical and biological weapons to nuclear energy, based in treaty law and other normative frameworks, that could be applied in addressing risks of AI.

32 We also recognize the need to be proactive. There are important lessons in recent experiences with other globally scalable, high-impact technologies, such as social media. Even as diverse societies process the impact and implications of AI, the need for effective global governance to share concerns and coordinate responses is clear.

33 We must identify, classify, and address AI risks, including building consensus on which risks are unacceptable and how they can be prevented or pre-empted. Alertness and horizon-scanning for unanticipated consequences from AI is also needed, as such systems are introduced in increasingly diverse and untested contexts.
To achieve this, we must overcome technical, political, and social challenges.

A) please don't use social media as a punching bag. It has very likely provided more value than harm to date through facilitating communication and government transparency. cf. papers based on data. B) please realise that wherever social media includes recommendation or search, it deploys AI algorithms. Social media are an important application of AI and therefore one of the application areas of this recommendation, not an externalisable example.

Challenges to be addressed

34 Many AI systems are opaque, either because of their inherent complexity or commercial secrecy as to their inner workings. Researchers and governance bodies have difficulty in accessing information or fully interrogating proprietary datasets, models, and systems. ~~Further, the science of AI is at an early stage, and we still do not fully understand how advanced AI systems behave.~~ This lack of transparency, access, compute and other resources, and understanding hinders the identification of where risks come from, and where responsibility for managing those risks (or compensating for harm) should lie.

Please do not include this BS. AI is a product, the transparency we need is on how it is created and tested. Cf. the EU AI act. Explainability of the AI itself can reduce other reporting obligations, but not eliminate them.

35 Despite AI's global reach, governance remains territorial and fragmented. National approaches to regulation that typically end at physical borders may lead to tension or conflict if AI does not respect those borders. Mapping, avoiding, and mitigating risks will require self-regulation, national regulation, as well as international governance efforts. There should be no accountability deficits.

This para touches something important, but not coherently. Cf. my earlier discussion of jurisdiction.

36 We also need to meet member states where they are and assist them with what they need in their own contexts given their specific constraints in terms of participation in and adherence to global AI governance, rather than telling them where they should be and what they should do based on a context to which they cannot relate.

yes

False. Responsibility for harms lies with whoever sold the product, unless they can prove it was used inappropriately. Deployers are free to pass liability on to suppliers, but are liable to their customers regardless of whether they can claim damages from the suppliers.

37 In addition to technical and political hurdles, these challenges exist in a broader social context. Digital technologies are impacting the 'software' of societies challenging governance writ large. Moreover, there are human and environmental costs of AI — hardware as well as software — must be accounted for throughout its lifecycle, as human lives and our environment are at the beginning and end of all AI-integrated processes.

yes, see my earlier points about focussing on governance in an age of AI, not governing AI itself. Maybe a little redundant though with earlier sections?

38 Besides misuse, we also note countervailing worries about *missed* uses — failing to take advantage of and share the benefits of AI technologies out of an excess of caution. Leveraging AI to improve access to education might raise concerns about young people's data privacy and teacher agency. However, in a world where hundreds of millions of students do not have access to quality education resources, there may be downsides of not using technology to bridge the gap. Agreeing on and addressing such trade-offs will benefit from international governance mechanisms that enable us to share information, pool resources, and adopt common strategies.

this is definitely true and was definitely mentioned earlier, but maybe these two points bear reiteration here?

International Governance of AI

Please mention the UNESCO recommendation, which is IMO the leading effort, with the possible exception of the full suite of EU digital legislation (not just the AIA!)

The AI governance landscape

39 There is, today, no shortage of guides, frameworks, and principles on AI governance. Documents have been drafted by the private sector and civil society, as well as by national, regional, and multilateral bodies, with varying degrees of impact. In technology terms, governance efforts have been focused on data, models, and benchmarks or evaluations. Applications have also been under focus, especially where there are existing sectoral governance arrangements, say for health or dual-use technologies. These efforts can be anchored in specific governance arrangements, such as the EU AI Act or the U.S. Executive Order and they can be associated with incentives for participation and compliance. *Figure 1* presents a simplified schema for considering the emerging AI governance landscape, which the Advisory Body will develop further in the next phase of its work.

The need for binding rules is much less debated than existential threat. Similarly, I'm pretty sure it's widely accepted in law and political science that government addresses KNOWN harms, that is that regulation SHOULD lag implementation.

40 Existing AI governance efforts have yielded similarities in language, such as the importance of fairness, accountability, and transparency. Yet there is no global alignment on implementation, either in terms of interoperability between jurisdictions or in terms of incentives for compliance within jurisdictions. Some favour binding rules while others prefer non-binding nudges. Trade-offs are debated, such as how to balance access and safety — or whether the focus should be on present day or potential future harms. Different models may also require different emphasis in governance. A lack of common standards and benchmarks among national and multinational risk management frameworks, as well as multiple definitions of AI used in such frameworks, have complicated the governance landscape for AI, notwithstanding the need for space for different regulatory approaches to co-exist reflecting the world's social and cultural diversity.

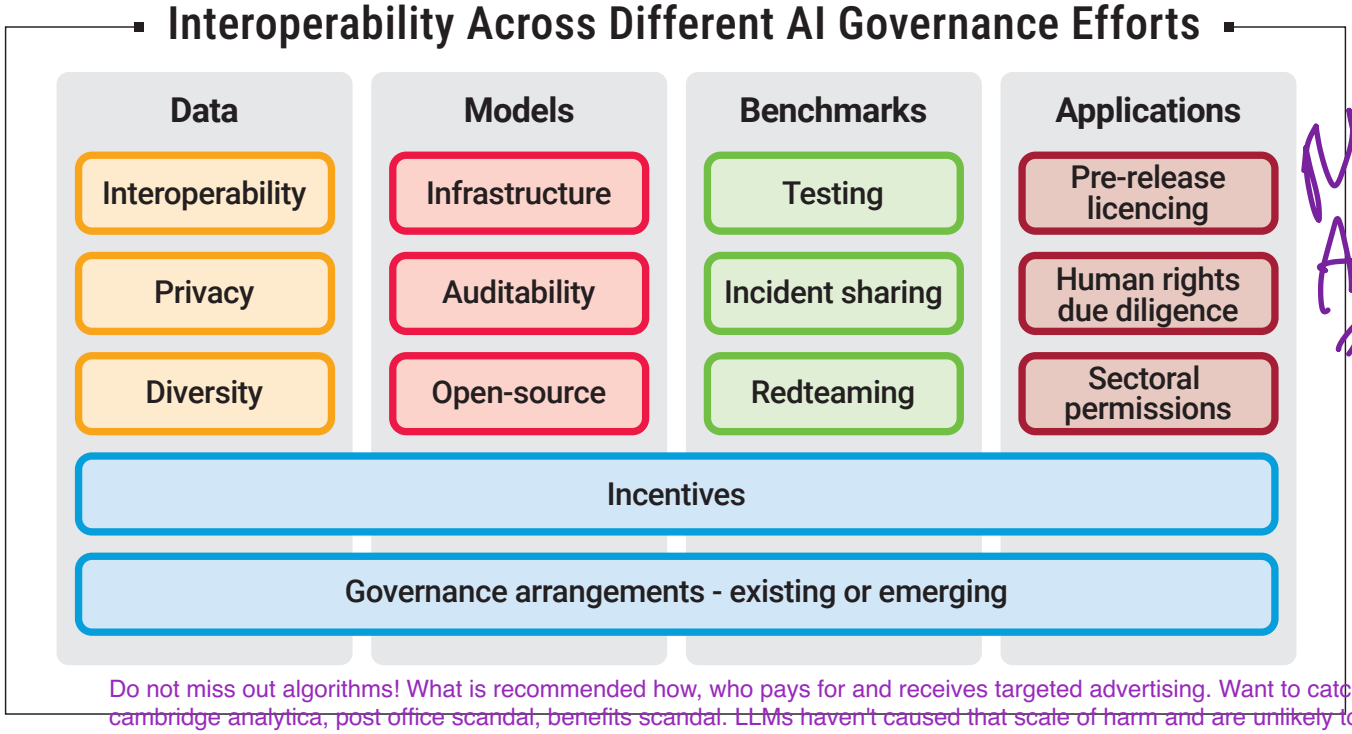
41 Meanwhile, technical advances in AI and its use continue accelerating, expanding the gap in understanding and capacity between technology companies

non seq. we could be using AI to narrow this gap.

OECD is pretty good here these days. But anyway, another reason to focus on governing in itself. Arguing about definitions is an established means for delaying progress.

Figure 1: A four-fold simplified schema for considering interoperability across different AI governance efforts

or indeed, some other means of increasing knowledge and understanding.



Do not miss out algorithms! What is recommended how, who pays for and receives targeted advertising. Want to catch ~~cambridge analytica, post office scandal, benefits scandal. LLMs haven't caused that scale of harm and are unlikely to~~ if they don't prove more useful than costly at some point.

developing AI, companies and other organizations using AI across various sectors and societal spaces, and those who would regulate its development, deployment, and use.

42 The result is that, in many jurisdictions AI governance can amount to self-policing by the developers, deployers, and users of AI systems themselves. Even assuming the good faith of these organizations and individuals, such a situation does not encourage a long-term view of risk or the inclusion of diverse stakeholders, especially those from the Global South. This must change.

see comment on 41, but point on inclusion is of course important.

Toward principles and functions of international AI governance

43 The Advisory Body is tasked with presenting options on the international governance of AI. We reviewed, among others, the functions performed by existing institutions of governance with a technological dimension, including FATF, FSB, IAEA, ICANN, ICAO, ILO, IMO, IPCC, ITU, SWIFT and UNOOSA [2]. These organizations offer inspiration and examples of global governance and coordination.

good

44 The range of stakeholders and potential applications presented by AI and their uses in a wide variety of contexts makes unsuitable an exact replication of any existing governance model. Nonetheless, lessons can be learned from examples of entities that have sought to: (a) build scientific consensus on risks, impact, and policy (IPCC); (b) establish global standards (ICAO, ITU, IMO), iterate and adapt them; (c) provide capacity building, mutual assurance and monitoring (IAEA, ICAO); (d) network and pool research resources (CERN); (e) engage diverse stakeholders (ILO, ICANN); (f) facilitate commercial flows and address systemic risks (SWIFT, FATF, FSB).

nice!

45 Rather than proposing any single model for AI governance at this stage, the preliminary recommendations offered in this interim report focus on the **principles** that should guide the formation of new global governance institutions for AI and the broad **functions** such institutions would need to perform. The subfunctions listed in *Table 1* below are informed by a survey of existing research on AI governance as well as a gap-analysis of nine current AI governance initiatives, namely, China's interim measures for the management of AI services, the Council of Europe's draft Convention on AI, the EU AI Act, the G7 Hiroshima Process, the Global Partnership on AI, the OECD AI Principles, the Partnership on AI and the Foundation Model Forum, the UK AI Safety Summit, and the U.S. Executive Order 14110.

Preliminary Recommendations

Further lessons from the EU effort – don't let people try to solve universal or essentially non-AI problems by calling them AI problems.

A. Guiding Principles

Guiding Principle 1. AI should be governed inclusively, by and for the benefit of all

the full benefits of human technology and wealth

46 Despite its potential, many of the world's peoples are not yet in a position to access and use ~~AI~~ in a manner that meaningfully improves their lives. Fully harnessing the potential of AI and enabling widespread participation in its development, deployment, and use is critical to driving sustainable solutions to global challenges. All citizens, including those in the Global South, should be able to create their own opportunities, harness them, and achieve prosperity ~~through AI~~. All countries, big or small, must be able to participate in ~~AI~~ governance.

global

47 Affirmative and corrective steps, including access and capacity building, will be needed to address the historical and structural exclusion of certain communities, for instance women and gender diverse actors, from the development, deployment, use, and governance of technology, and to turn digital divides into inclusive digital opportunities.

47 is good :-)

These problems cannot be solved only for AI without being addresse d more generally, but AI may be part of their solution.

Guiding Principle 2. AI must be governed in the public interest

48 The development of AI systems is largely concentrated in the hands of technology companies. The refinement, deployment and use of AI will involve other actors including but not limited to the original developers (be they companies, small AI labs, other

Not really, cf Bryson & Malikova, unless you are talking about generative AI which so far has limited utility.
Agree though too much power is concentrated, particularly some services really have only one vender, e.g. search (google), political communication (twitter), online sales (amazon), or two or three such as cloud compute.

2 See the list of abbreviations in the annex.

organizations as well as countries) but include deployers and users who will range from individuals, to companies, organizations, and governments and who will bring a wide variety of incentives to their approaches.

49 As shown by the experience with social media, AI products and services can scale rapidly across borders and categories of users. For this reason, as well as wider considerations of opportunities and risks, AI must be governed in the broader public interest. "Do no harm" is necessary, but not sufficient. A broader framing is needed for accountability of companies and other organisations that build, deploy and control AI as well as those that use AI across multiple sectors of the economy and society across the lifecycle of AI. This cannot rely on self-regulation alone: binding norms enforced by member states consistently are needed to ensure that public interests, rather than private interests, prevail.

yes. cf EU AIA

50 AI will be used by people and organizations, across multiple sectors, each with different use-cases and complexities and risks. Governance efforts must bear in mind public policy goals related to diversity, equity, inclusion, sustainability, societal and individual well-being, competitive markets, and healthy innovation ecosystems. They must also integrate implications of missed uses for economic and social development. Governance in this context should expand representation of diverse stakeholders, as well as offer greater clarity in delineating responsibilities between public and private sector actors. Governing in the public interest also implies investments in public technology, infrastructure, and the capacity of public officials.

Guiding Principle 3.
~~AI governance~~ should be built ~~in step with~~ data governance and the promotion of data commons

51 Data is critical for many major AI systems. Its governance and management in the public interest cannot be divorced from other components of AI governance *(Figure 1)*. Regulatory frameworks and techno-legal arrangements that protect privacy and security of personal data, consistent with applicable laws, while actively facilitating the use of such data will be a critical complement to AI governance arrangements, consistent with local or regional law. The development of public data commons should also be encouraged with particular attention to public data that is critical for helping solve societal challenges including climate change, public health, economic development, capacity building and crisis response, for use by multiple stakeholders.

Guiding Principle 4.
AI governance must be ~~universal~~, networked and rooted in adaptive multi-stakeholder collaboration

the widest possible

52 Any AI governance effort should prioritize ~~universal~~ buy-in by different member states and stakeholders. This is in addition to inclusive participation, in particular lowering entry barriers for previously excluded communities in the Global South *(Guiding Principle 1)*. This is key for emerging AI regulations to be harmonized in ways that avoid accountability gaps.

however, veto players should not be tolerated, rather participation should be positively incentivised e.g. through access to products and utilities, or participation in redistribution.

53 Effective governance should leverage existing institutions that will have to review their current functions in light of the impact of AI. ~~But this is not enough.~~ New horizontal coordination and supervisory functions are required ~~and they should be~~ entrusted to a new organizational structure. New and existing institutions could form nodes in a network of governance structures. There is a clear momentum across diverse states for this to happen as well as growing awareness in the private sector for a well-coordinated and interoperable governance framework. Civil society concerns regarding the impact of AI on human rights point in a similar direction.

Deserves own item

may also be required, possibly

probably mean call for, but tough. Cf. previous comments on jurisdiction. Tech giants want one uniform world, and so do some nations spy agencies. Diversity is safer, more inclusive, more resilient, and more innovative.

54 Such an AI governance framework can draw on best practices and expertise from around the world. It must also be informed by understanding of different cultural ideologies driving AI development, deployment, and use. Innovative structures within this governance framework would be needed to engage the private sector, academia, and civil society alongside governments. Inspiration may be drawn from past efforts to engage the private sector in pursuit of public goods, including the ILO's tripartite structure and the UN Global Compact.

TCV?

Guiding Principle 5.
AI governance should be anchored in the UN Charter, International Human Rights Law, and other agreed international commitments such as the Sustainable Development Goals

55 The UN has a unique normative and institutional role to play; aligning AI governance with foundational UN values — notably the UN Charter and its commitment to peace and security, human rights, and sustainable development — offers a robust foundation and compass. The UN is positioned to consider AI's impact on a variety of global economic, social, health, security, and cultural conditions, all grounded in the need to maintain universal respect for, and enforcement of, human rights and the rule of law. Several UN agencies have already done important work on the impact of AI on fields from education to arms control.

UNESCO!!

56 The Global Digital Compact and the Roadmap for Digital Cooperation are examples of multi-stakeholder deliberations towards a global governance framework of technologies including AI. Strong involvement of UN member states, empowering UN agencies and involving diverse stakeholders, will be vital to empowering and resourcing a global AI governance response.

Failing to publicly recognise the immense effort and quality of output by UNESCO undermines the legitimacy of this UN effort, and indeed of the UN and its processes.

B. Institutional Functions

57 We consider that to properly govern AI for humanity, an international governance regime for AI should carry out at least the following functions. These could be carried out by individual institutions or a network of institutions.

58 Figure 2 summarizes our recommended institutional functions for international AI governance. At the global level, international organizations, governments, and private sector would bear primary responsibility for these functions. Civil society, including academia and independent scientists, would play key roles in building evidence for policy, assessing impact, and holding key actors to account during implementation. Each set of functions would have different loci of responsibility at different layers of governance — private sector, government, and international organizations. We will further develop the concept of shared and differentiated responsibilities for multiple stakeholders at different layers of the governance stack in the next phase of our work.

Can you say something about the sovereignty of member nations and their critical role as representatives of their residents, and enforcers of law and order.

-not unitary!

Institutional Function 1:
Assess regularly the future directions and implications of AI

This isn't surprising because AI is a set of technologies, not a unitary actor.

59 There is, presently, no authoritative institutionalized function for independent, inclusive, multidisciplinary assessments on the future trajectory and implications of AI. A consensus on the direction and pace of AI technologies — and associated risks and opportunities — could be a resource for policymakers to draw on when developing domestic AI programmes to encourage innovation and manage risks.

unlikely

60 In a manner similar to the IPCC, a specialized AI knowledge and research function would involve an independent, expert-led process that unlocks scientific, evidence-based insights, say every six months, to inform policymakers about the future trajectory of AI development, deployment, and use *(subfunctions 1-3 in Table 1)*. This should include arrangements with companies on access to information for the purposes of research and horizon-scanning. This function would help the public better understand AI, and drive consensus in the international community about the speed and impact of AI's evolution. It would produce regular shared risk assessments, as well as establishing standards to measure the environmental and other impacts of AI. This Advisory Body is in a way the start of such an experts-led process, which would need to be properly resourced and institutionalised.

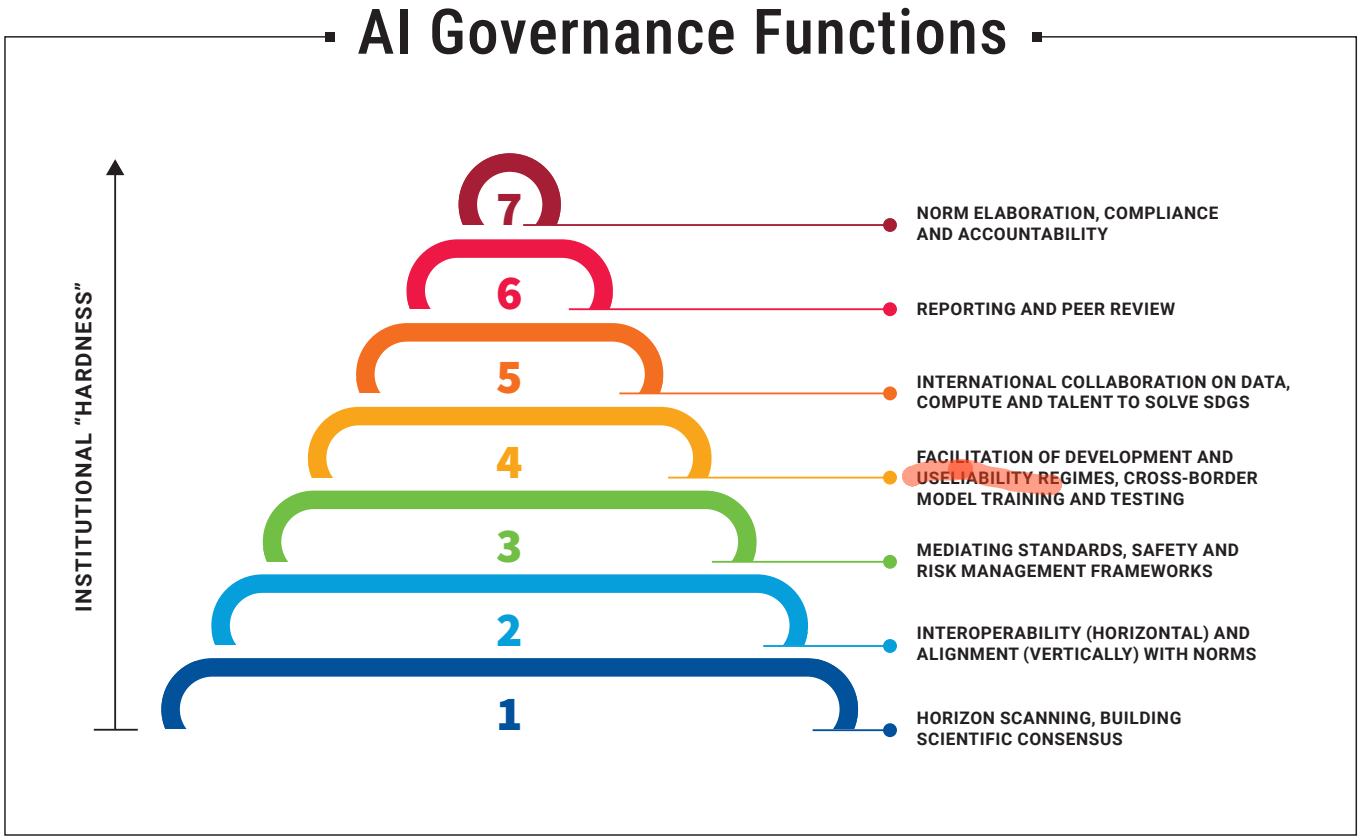
61 The extent of AI's negative externalities is not yet fully clear. The role of AI in disintermediating aspects of life that are core to human development could fundamentally change how individuals and communities function. As AI capabilities further advance, there is the potential for profound, structural adjustments to the way we live, work, and interact. A global analytical observatory function could coordinate research efforts on critical social impacts of AI, including its effects on labour, education, public health, peace and security, and geopolitical stability. Drawing on expertise and sharing knowledge from around the world, such a function could facilitate the emergence of best practices and common responses.

and cannot be, since they depend on both application area and chosen system architecture..

new digital technologies and applications

I hate when people call AI governance instruments "observatories" like technology was a natural kind like the stars over which we have no capacity for control. Again, this is pandering to those who do not want regulation, and want you to think AI is opaque and unknowable.

Figure 2: AI governance functions distributed by institutional 'hardness'



Institutional Function 2: Reinforce interoperability of governance efforts emerging around the world and their grounding in international norms through a Global AI Governance Framework endorsed in a universal setting (UN)

62 AI governance arrangements should be interoperable across jurisdictions and grounded in international norms, such as the Universal Declaration of Human Rights *(Principle 4 above)*. They should leverage existing UN organizations and fora such as UNESCO and ITU for reinforcing interoperability of regulatory measures across jurisdictions. AI governance efforts could also be coordinated through a body that harmonises policies, builds common understandings, surfaces best practices, supports implementation and promotes peer-to-peer learning *(subfunctions 7-10 in Table 1)*. A Global AI Governance Framework could support policymaking and guide implementation to avoid AI divides and governance gaps across public and private sectors, regions, and countries as well as clarifying the principles and norms under which various organizations should operate. As part of this framework, special attention should be paid to capacity-building both in the private and public sectors as well as dissemination of knowledge and awareness across the world. Best practices such as human rights impact assessments by private and public sector developers of AI systems could be spread through such a framework, which may a need an international agreement.

See previous comments on diversity in jurisdiction. I don't believe there is any more to do on this remit after UNESCO's excellent recommendation. Just own and run with that. Maybe promise to update it in 5 years or something.

Institutional Function 3: Develop and harmonize standards, safety, and risk management frameworks

63 Several important initiatives to develop technical and normative standards, safety, and risk management frameworks for AI are underway, but there is a lack of global harmonization and alignment *(subfunction 11 in Table 1)*. Because of its global membership, the

again, prefer this to be about encouraging harmonisation where tractable and dissemination of accessible known solutions for those who cannot afford innovation, but not too much stressing over "fragmentation"

UN can play a critical role in bringing states together, developing common socio-technical standards, and ensuring legal and technical interoperability.

64 As an example, emerging AI safety institutes could be networked to reduce the risk of competing frameworks, fragmentation of standardization practices across jurisdictions, and a global patchwork with too many gaps. Care should, however, be taken not to overemphasise technical interoperability without parallel movement on other functions and norms. While there is greater awareness of socio-technical standards, more research, active involvement of civil society and transdisciplinary cooperation is needed to develop such standards.

65 Further, new global standards and indicators to measure and track the environmental impact of AI as well as its energy and natural resources consumption (i.e. electricity and water) could be defined to guide AI development and help achieve SDGs related to the protection of the environment.

Institutional Function 4: Facilitate development, deployment, and use of AI for economic and societal benefit through international multi-stakeholder cooperation

66 In addition to standards for preventing harm and misuse, developers and users, especially in the Global South, need critical enablers such as standards for data labelling and testing, data protection and exchange protocols that enable testing and deployment across borders for startups as well as legal liability, dispute resolution, business development, and other supporting mechanisms. Existing legal, financial, and technical arrangements need to evolve to anticipate complex adaptive AI systems of the future, and this will require taking into account lessons learnt from forums such as FATF, SWIFT and equivalent mechanisms. In addition, for most countries and regions, capacity development in the public sector is urgently required to facilitate responsible and beneficial use of AI as well as participate in international multi-stakeholder cooperative frameworks to develop enablers for AI *(subfunctions 4, 5 and 11 in Table 1).*

This one is really good.

this is mostly only relevant to generative AI (and surveillance). Be sure to avoid lock-in to established businesses.

Institutional Function 5: Promote international collaboration on talent development, access to compute infrastructure, building of diverse high-quality datasets, responsible sharing of open-source models, and AI-enabled public goods for the SDGs

67 A new mechanism (or mechanisms) is required to facilitate access to data, compute, and talent in order to develop, deploy, and use AI systems for the SDGs through upgraded local value chains, giving independent academic researchers, social entrepreneurs, and civil society access to the infrastructure and datasets needed to build their own models and to conduct research and evaluations. This may require networked resources and efforts to build common datasets and data commons for use in the public interest, responsible sharing of open-source models, computational resources, and scale education and training.

or at least talent

68 Pooling expert knowledge and resources analogous to CERN, EMLB or ITER, as well as the technology diffusion functions of the IAEA, could provide a much-needed boost to the SDGs *(subfunction 6 in Table 1).* Creating incentives for private sector actors to share and make available tools for research and development can also complement such functions. Experts from the Global South are often invisible at global conferences on AI. This needs to change.

69 Opening access to data and compute should also be accompanied by capacity-building, in particular in the Global South. To facilitate local creation, adoption, and context-specific tuning of models, it would be important to track positive uses of AI, incentivize and assess AI-enabled public goods. Private sector engagement would be crucial in leveraging AI for the SDGs. Analogous to commitments made by businesses under the Global Compact, this could include public promises by technology and other companies to develop, deploy, and use AI for the greater good. In the larger context of the Global Digital Compact, it could also include reporting on the ways in which AI is supporting the Sustainable Development Goals.

Institutional Function 6: Monitor risks, report incidents, coordinate emergency response

70 The borderless nature of AI tools, which can proliferate across the globe at the stroke of a key, pose new challenges to international security and global stability. AI models could lower the barriers for access to weapons of mass destruction. AI-enabled cyber tools increase the risk of attacks on critical infrastructure and dual-use AI can be used to power lethal autonomous weapons which could pose a risk to international humanitarian law and other norms. Bots can rapidly disseminate harmful information, with increasingly human characteristics, in a manner that can cause significant damage to markets and public institutions. The possibility of rogue AI escaping control and posing still larger risks cannot be ruled out. Given these challenges, capabilities must be created at a global level to monitor, report, and rapidly respond to systemic vulnerabilities and disruptions to international stability *(subfunctions 13, 14 in Table 1)*.

a rogue AI system escaping control temporarily cannot be entirely ruled out, and indeed arguably charecterises airline pricing anomolies and "flash crashes", though to date such matters have been quickly resolved.

71 For example, a techno-prudential model, akin to the macro-prudential framework used to increase resilience in central banking and bringing together those developed at the national level, may help to similarly insulate against AI risks to global stability. Such a model must be grounded in human rights principles.

72 Reporting frameworks can be inspired by existing practices of the IAEA for mutual reassurance on nuclear safety and nuclear security, as well as the WHO on disease surveillance.

yes.

Institutional Function 7: Compliance and accountability based on norms

this section is pretty good.

73 We cannot rule out that legally binding norms and enforcement would be required at the global level. A regional effort for an AI treaty is already underway and the issue of lethal autonomous weapons is under consideration in the framework of a treaty on conventional weapons. Non-binding norms could also play an important role, alone or in combination with binding norms. The UN cannot and should not seek to be the sole arbiter of AI governance. However, in certain fields, such as challenges to international security, it has unique legitimacy to elaborate norms *(subfunction 12 in Table 1)*. It can also help ensure that there are no accountability gaps, for example by encouraging states to report analogous to reporting on the SDGs targets and the Universal Periodic Review that facilitates monitoring, assessing, and reporting on human rights practices *(subfunctions 15 in Table 1)*. This would need to be done in a timely and accurate way. Inspired by existing institutions such as the WTO, dispute resolution can also be facilitated through global forums.

74 At the same time, the legitimacy of any global governance institution depends on accountability of that institution itself. International governance efforts must demonstrate resolute transparency in objectives and processes and make all efforts to gain the trust of citizen stakeholders, including by preventing conflicts of interest.

Table 1: Subfunctions for international governance of AI

Summary table of subfunctions for international governance of artificial intelligence, and possible timeframes for realization

SUBFUNCTION	DESCRIPTION	CATEGORY	POSSIBLE TIMEFRAME REQUIRED TO INSTITUTIONALISE PROPOSED SUBFUNCTION
1. Scientific assessment	Prepare a public review of international, regional, and national AI policies at least every 6 months.	Research & Analysis	6-12 months
2. Horizon scanning	Prepare a horizon-scanning report that identifies risks that transcend borders and can potentially affect all jurisdictions.	Research & Analysis	6-12 months

SUBFUNCTION	DESCRIPTION	CATEGORY	POSSIBLE TIMEFRAME REQUIRED TO INSTITUTIONALISE PROPOSED SUBFUNCTION
3. Risk classification	Assess existing and upcoming AI ~~models~~ uses on a risk scale of untenable, high-level, mid-level, and low to no risks.	Research & Analysis	6-12 months at least every six months.
4. Access to benefits	Equitable access to technology and benefits of AI, accelerating achievement of the Sustainable Development Goals.	Enabling	12-24 months
5. Capacity building	Programs and resources to build AI technology and businesses as well as governance and promotional capacity among states.	Enabling	12-24 months
6. Joint R&D	Establish the capacity to undertake collaborative research and development of AI to benefit those who don't have access to AI tools or expertise.	Enabling	12-24 months
7. Inclusive participation	Ensure participation of all stakeholder groups and all countries and regions in collective governance, risk management and realization of opportunities; ~~strive for innovative governance~~.	Governing	6-12 months innovation is great where useful, but GOOD governance is the goal, and many good tricks are known and just not applied here.
8. Convening, international learning	Convene stakeholders regularly to consider AI policies across jurisdictions; building consensus on shared vocabulary and definitions; peer to peer learning.	Governing	6-12 months ?
9. International coordination	Deconflicting work and building synergy across existing international bodies that continue to address AI.	Governing	6-12 months
10. Policy harmonization; norm alignment	Surfacing best practices for norms and rules, including for risk mitigation and economic growth. Align, leverage, and include, soft and hard law, standards, methods, and frameworks developed at the regional, national, and industry level to support interoperability.	Governing	12-24 months policy harmonisation is nice where sensible, not sure norm alignment is the kind of thing the UN should be doing, though I like the first sentence of description, don't want to see innovators bludgeoned with this.
11. Standard setting	Develop global consensus on standards for AI use across stakeholder groups by working with national standards development organizations (SDOs) - updated regularly.	Governing	12-24 months where practicable.
12. Norm elaboration	Convene stakeholders to assess the necessity of and negotiate non-binding and binding frameworks, treaties, or other regimes for AI.	Governing	24-36 months nice
13. Enforcement	Develop mutual reassurance schemes, information sharing mechanisms that respect commercial and national security information, dispute resolution mechanisms, and liability schemes/regimes.	Governing	> 36 months nice
14. Stabilization and response	Develop and collectively maintain an emergency response capacity, off-switches and other stabilization measures.	Governing	> 36 months probably silly, but mostly harmless. Unless encourages recklessness, or lack of attribution of human accountability.
15. Monitoring and verification	Elaborate oversight and verification schemes where appropriate to ensure that the design, deployment and use of AI systems is in compliance with applicable international law.	Governing	> 36 months essential!

Conclusion

75 To the extent that AI impacts our lives — how we work and socialize, how we are educated and governed, how we interact with one another daily — it raises questions more fundamental than how to govern it. Such questions of what it means to be human in a fully digital and networked world go well beyond the scope of this Advisory Body. Yet they are implicated in the decisions we make today. For governance is not an end but a means, a set of mechanisms intended to exercise control or direction of something that has the potential for good or ill.

nice

76 We aspire to be both comprehensive in our assessment of the impact of AI on people's lives and targeted in identifying the unique difference the UN can make. We hope it is apparent that we see real benefits of AI; equally, we are clear-eyed about its risks.

77 The risks of inaction are also clear. We believe that global AI ~~global~~ governance is essential to reap the significant opportunities and navigate the risks that this technology presents for every state, community, and individual today. And for the generations to come.

78 To be effective, the international governance of AI must be guided by principles and implemented through clear functions. These global functions must add value, fill identified gaps, and enable interoperable action at regional, national, industry, and community levels. They must be performed in concert across international institutions, national and regional frameworks as well as the private sector. Our preliminary recommendations set out what we consider to be core principles and functions for any global AI governance framework.

79 We have taken a form follows function approach and do not, at this stage, propose any single model for AI governance. Ultimately, however, AI governance must deliver tangible benefits and safeguards to people and societies. An effective global governance framework must bridge the gap between principles and practical impact. In the next phase of our work, we will explore options for institutional forms for global AI governance, building on the perspectives of diverse stakeholders worldwide.

Next Steps

80 Rather than proposing any single model for AI governance at this stage, the foregoing preliminary recommendations focus on the principles and functions to which any such regime must aspire.

81 Over the coming months we will consult — individually and in groups — with diverse stakeholders around the world. This includes participation at events tasked with discussing the issues in this report as well as engagement with governments, the private sector, civil society, and research and technical communities. We will also pursue our research, including on risk assessment methodologies and governance interoperability. Case studies will be developed to help think about landing issues identified in the report in specific contexts. We also intend to dive deep into a few areas, including Open-Source, AI and the financial sector, standard setting, intellectual property, human rights, and the future of work by leveraging existing efforts and institutions.

82 We encourage constructive engagement from anyone with an interest in AI. More information about how to engage with our ongoing work can be found online at https://www.un.org/en/ai-advisory-body.

83 We look forward to engaging with diverse stakeholders as we answer more fully the questions identified in this interim report, in support of the ongoing efforts of the United Nations on digital cooperation and on social progress and better standards of life in larger freedom.

Box 4: Example of questions to be addressed during consultations on this interim report

Key questions for further discussion in the next phase of work

Opportunities and enablers of AI

Can we make AI development more inclusive by facilitating model-building ecosystems, for example through data protection and exchange frameworks, with shared access to compute?

Would common standards for ~~data labelling and~~ testing encourage AI startups to test and deploy across more countries and regions?

common standards for data labelling might be useful, but not integral to the testing question.

What mechanisms would promote equitable access to compute and privacy-preserving sharing of datasets across stakeholders and member states?

How can we grow and spread AI talent? Can UN entities or other institutions facilitate exchange of students, joint PhD programmes, and cross-domain (health and AI, agriculture and AI) talent development?

governance and AI

material, and energy resources

How can international collaboration harness ~~AI~~ talent, ~~data~~ and ~~compute~~ for scientific research and for the SDGs?

How can we incentivize governments and the private sector to invest in other core infrastructures that drive ~~AI~~ development around the world?

These are the real questions. AI should only be brought in when cost/benefit indicates.

Risks and challenges of AI

diversity leads to innovation. Don't obsess too much on consensus. Certainly do not allow veto players.

What is the best path to reaching consensus on identifying, classifying, and addressing AI risks?

How should assessments of risks and challenges relate to more specific use cases of AI, notably autonomous weapons systems?

Can we please have an objective analysis of how autonomy is impacting ongoing present-day conflicts before asking this question again?

What should be the threshold or the trigger for identifying red lines (analogous, perhaps, to the ban on human cloning in biomedical research)? How would any such red line be policed and enforced?

I really doubt there is a single mechanism for all red lines. Like cloning in biology or bioweapons more generally, probably need bespoke efforts.

International governance of AI

Do the principles listed above properly reflect the aspirations that a global governance regime for AI should have?

Do the functions outlined above properly reflect what global AI governance can and should do?

What structural arrangement(s) would best empower a new institution or set of institutions to uphold these principles and carry out these functions?

A range of models exist within the UN system for engaging industry in sectoral work (WHO, ITU, ICAO etc).

What kind of mechanism could best support industry participation in international governance of AI? Which of the normative, policy and information instruments that exist today could support coherence in technology governance across governments, private sector and civil society?

What kind of financing and capacity building mechanisms would be needed for effective international arrangements to address the functions outlined above?

Annexes

About the High-Level Advisory Body on AI

Initially proposed in 2020 as part of the Secretary-General's Roadmap for Digital Cooperation (A/74/821), The multi-stakeholder High-level Advisory Body on Artificial Intelligence was formed in October 2023 to undertake analysis and advance recommendations for the international governance of AI.

Advisory Body members participated in their personal capacity, not as representatives of their respective organizations. This proposal represents a majority consensus; no member is expected to endorse every single point contained in the document. In publishing this report, the UN AI Advisory Group members affirm their broad, but not unilateral, agreement with its findings and recommendations. Language included in this report does not imply institutional endorsement by the UN AI Advisory Group members' respective organizations.

Members of the High-Level Advisory Body on AI

- Carme Artigas (Co-Chair)
- James Manyika (Co-Chair)
- Anna Abramova
- Omar Sultan Al Olama
- Latifa Al-Abdulkarim
- Estela Aranha
- Ran Balicer
- Paolo Benanti
- Abeba Birhane
- Ian Bremmer
- Anna Christmann
- Natasha Crampton
- Nighat Dad
- Vilas Dhar
- Virginia Dignum
- Arisa Ema
- Mohamed Farahat
- Amandeep Singh Gill
- Wendy Hall
- Rahaf Harfoush
- Hiroaki Kitano
- Haksoo Ko
- Andreas Krause
- Maria Vanina Martinez Posse
- Seydina Moussa Ndiaye
- Mira Murati
- Petri Myllymaki
- Alondra Nelson
- Nazneen Rajani
- Craig Ramlal
- He Ruimin
- Emma Ruttkamp-Bloem
- Marietje Schaake
- Sharad Sharma
- Jaan Tallinn
- Philip Thigo
- Jimena Sofia Viveros Alvarez
- Yi Zeng
- Zhang Linghan

Terms of Reference for the High-level Advisory Body on AI

The High-level Advisory Body on Artificial Intelligence, convened by the United Nations Secretary-General, will undertake analysis and advance recommendations for the international governance of artificial intelligence. The Body's initial reports will provide high-level expert and independent contributions to ongoing national, regional, and multilateral debates.

The Body will consist of 38 members from governments, private sector, civil society, and academia, as well as a member Secretary. Its composition will be balanced by gender, age, geographic representation, and area of expertise related to the risks and applications of artificial intelligence. The members of the Body will serve in their personal capacity.

The Body will engage and consult widely with governments, private sector, academia, civil society, and international organizations. It will be agile and innovative in interacting with existing processes and platforms as well as in harnessing inputs from diverse stakeholders. It could set up working parties or groups on specific topics.

The members of the Body will be selected by the Secretary-General based on nominations from Member States and a public call for candidates. It will have two Co-Chairs and an Executive Committee. All stakeholder groups will be represented in the Executive Committee.

The Body shall be convened for an initial period of one year, with the possibility of extension by the Secretary-General. It will have both in-person and online meetings.

The Body will prepare a first report by 31 December 2023 for the consideration of the Secretary-General and the Member States of the United Nations. This first report will present a high-level analysis of options for the international governance of artificial intelligence.

Based on feedback to the first report, the Body will submit a second report by 31 August 2024 which may provide detailed recommendations on the functions, form, and timelines for a new international agency for the governance of artificial intelligence.

The Body shall avoid duplication with existing forums and processes where issues of artificial intelligence are considered. Instead, it shall seek to leverage existing platforms and partners, including UN entities, working in related domains. It shall fully respect current UN structures as well as national, regional, and industry prerogatives in the governance of artificial intelligence.

The deliberations of the Body will be supported by a small secretariat based in the Office of the Secretary-General's Envoy on Technology and be funded by extrabudgetary donor resources.

Working Groups and Cross-Cutting Themes

The ongoing work of the Advisory Body is organized around five working groups and ten cross-cutting themes. Sectoral applications and additional themes will be considered in detail in the next phase.

Working Groups

- Opportunities and Enablers
- Risks and Challenges
- Interoperability
- Alignment with Norms and Values
- International Institutions

Cross-Cutting Issues

- Culture
- Equity
- Ethics
- Future of work
- Government capacity
- Gender
- Human Rights, Democracy, Rule of Law
- Open-Source
- Societal impact
- Sustainability

Sectoral applications and additional themes for deep dives

- Agriculture
- Education
- Environment
- Finance
- Health
- Intellectual property
- National Security
- Standard-setting

Governance

List of Abbreviations

EMBL	European Molecular Biology Laboratory
EU	European Union
FATF	Financial Action Task Force
FSB	Financial Stability Board
G7	Group of Seven
G20	Group of 20
GPAI	Global Partnership on AI
GPU	Graphics Processing Unit
IAEA	International Atomic Energy Agency
ICANN	Internet Corporation for Assigned Names and Numbers
ICAO	International Civil Aviation Organization
ILO	International Labour Organization
IMO	International Maritime Organization
IPCC	Intergovernmental Panel on Climate Change
ITER	International Thermonuclear Experimental Reactor
ITU	International Telecommunication Union
OECD	Organization for Economic Cooperation and Development
SDG	Sustainable Development Goals
SWIFT	Society for Worldwide Interbank Financial Telecommunication
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNOOSA	United Nations Office for Outer Space Affairs

This report is published by the
Advisory Body on Artificial Intelligence

For more information, contact the
AI Advisory Body Secretariat:
aiadvisorybody@un.org