Multi-Stakeholder Consultation on AI Governance for the Global Digital Compact

PUBLISHED
April, 2023

Contents

Introduction

The Global Digital Compact (GDC) is the stakeholder entry point for shaping the technology agenda in the Summit of the Future. Cognizant of this meeting's importance, the Future of Life Institute (FLI) organized a group of 41 individuals representing 25 non-profits, academia, firms, and multilateral organizations for a virtual consultation to offer recommendations on the subject of artificial intelligence (AI) governance.[1]

Our collective aspiration is that the United Nations (UN) Office of the Secretary-General's Envoy on Technology takes our input for the purpose of improving the likelihood that AI contributes to the 2030 agenda on sustainable development. Furthermore, we believe that the UN has a unique and critical role in furthering this technology's governance to spread its benefits and mitigate its potential harms. When the UN serves as a universal forum to catalyze constructive collective action, the whole world wins.

This document synthesizes the consultation's results into two sections. The first section describes how the input was gathered from consultation participants. In short, we asked stakeholders to share their insights over three rounds of feedback during March of 2023. The second section delivers an analysis of the responses given and prioritized by participants. In addition to the GDC's two prompts on principles and key commitments, pledges, and actions, participants answered four extra prompts expressing their views on the submission process and the UN.

If you have any questions or comments on this consultation, please contact Carlos Ignacio Gutierrez, UN Representative and AI Policy Researcher at FLI, at carlos@futureoflife.org.

1 Participant organizations in this process include: Ada Lovelace Institute, AI Governance and Safety Canada, Brookings, C Minds, Canadian Institute for Advanced Research, Carnegie Endowment for International Peace, Center for Long-Term Cybersecurity, Centre for Long-Term Resilience, Centre for the Governance of AI, Centro Nacional de Inteligencia Artificial, Clarkson University, Concordia, Corporación Andina de Fomento, DeepMind, DiploFoundation, Future of Life Institute, GobLab - Universidad Adolfo Ibáñez, International Centre of Expertise in Montréal on Artificial Intelligence, Legal Priorities Project, Organisation for Economic Co-operation and Development, Partnership on Artificial Intelligence, Responsible Artificial Intelligence Institute, Simon Institute for Long-term Governance, The Future Society, and the University of Richmond.

Process

FLI is the organization responsible for designing and coordinating this consultation on AI governance. Its objective was to recruit a group of organizations with varied interests, expertise, and perspectives on AI governance. To this end, 25 entities accepted our call to participate in this process (see Annex I for a complete list of representatives). It is relevant to note that individuals were allowed to provide their feedback anonymously in all rounds of input. Our intent with this methodology was to create a space that fostered an honest exchange of ideas.

The consultation process was divided into three rounds:

Round one consisted of a survey with six prompts (Annex II contains a copy of the survey). The first two prompts reflected the language used in the GDC's original submission form and included contextual information to aid our participants in their responses. The latter four prompts explored the willingness of our group to engage in the GDC process and their opinions on the UN's role in the AI governance space. We received a total of ***224 submission ideas*** during this round.

Round two represented an opportunity to share participant input with the entire group. The FLI team consolidated the feedback from the first two questions into groups via a pile-sorting methodology and utilized an electronic platform to enable commenting and up and down voting on ideas (see Annex III for a screenshot of this system). This round was fundamental in prioritizing the principles and key commitments, pledges, and actions submitted in round one. A total of ***458 votes and 56 comments*** were submitted in this round.

Round three was centered on a final group call. Participants attended a one-hour virtual meeting where FLI shared the results of the consultation. This was an opportunity to express input on any part of the consultation's process and its outputs.

Participant Input

Participants in the virtual consultation were asked for feedback on six prompts related to the GDC. This section contains highlights of these responses, divided into four parts. The first two reflect on the GDC's submission request that focuses on principles and commitments, pledges, and actions. The subsequent two explore participant awareness and likelihood to engage in the GDC process. FLI's role in drafting these sub-sections was to package ideas concisely, yet remain true to participant submissions and preferences.

Principles

The first round of the virtual consultation received 93 principle proposals. This sub-section condenses those submissions into five principles using the 253 votes and 33 reactions gathered in round two of the process (see Annex IV for the list of prioritized principles that influenced each of the proposals below). An overarching theme in this section centers on the duties of AI systems designers, developers, and deployers to individuals, groups, and society. Specifically, it is understood that the creation and commercialization of AI technologies that respect the rights of stakeholders are a critical step towards achieving the UN's 2030 agenda on sustainable development.

Mitigating Risks throughout the AI Lifecycle

The incorporation of mechanisms to identify and mitigate AI system risks throughout their lifecycle is essential. By 'lifecycle,' our consultation participants refer to the progression of a system from its conception in the design phase, through its development, to its deployment, and eventual use by consumers. In all of these phases, entities should ensure that real or potential direct and indirect harms are proactively avoided (this could be done through, for example, a risk management framework).

> AI systems should proactively incorporate mechanisms to systematically mitigate their real or perceived direct and indirect risks throughout their entire lifecycle (design, development, and deployment). This analysis should be performed at the individual, group, and societal level.

Transparency and Explainability

Our virtual consultation participants specified two levels of information asymmetry that merit action. On one level, an explanation of how an AI system processes information and generates outputs should be available to relevant stakeholders. As the use cases of AI grow, individuals, communities, and policymakers should understand what data is utilized and how, in situations where automated decisions can generate consequential harms.

The second level of information asymmetry centers on the entity responsible for an AI system. Here, participants stressed the need for an expansion of system transparency and explainability that details how a system is used, its capabilities, limitations, and what opportunities exist to challenge its outputs and decisions. Improving the public's understanding of a

system and its characteristics benefits society and advances the protection of human rights.

> The outputs produced by AI systems, in the form of actions or decisions, need to be understandable and available for the inspection and independent oversight of interested parties.
>
> Equally important, entities in charge of AI systems must establish open lines of communication where the ability to consent, opt-out, challenge errors, and learn about a system's capabilities, limitations, and real or perceived impact is possible for individuals and communities.

Accountability of Stakeholders

The advantages of AI systems do not exist in a risk vacuum. Incidents where harm is experienced because of this technology are documented in all corners of the globe. This reality underscores the need to align the incentives between entities responsible for a system and social welfare by establishing an effective accountability framework. Such a framework must adapt to the changing nature, capabilities, and potential harms caused by this technology.

> To protect society from harms, all deployed AI systems must assign a party that is responsible for its output. In parallel, government authorities should generate the necessary incentives to minimize gaps in the accountability of these systems.

Fairness

One of the advantages of AI is its scalability, allowing virtually limitless numbers of individuals to reap its benefits. The other side of the coin to this reality is that individuals are confronted with systems that are designed in places with different cultures, idiosyncrasies, biases, and objectives. This can lead to outputs that produce harms such as discrimination against groups because of their demographic characteristics, the reinforcement of power asymmetries, or the minimization of non-dominant ideas or perspectives. All of these harms should be avoided.

> Designers, developers, and deployers have an obligation to account for the differential treatment of individuals and groups by their AI systems. This ranges from assessing discrimination due to demographic characteristics, contextualizing the varying effects of outputs, analyzing the distribution of harms and benefits on different groups, and protecting intellectual diversity (e.g., non-dominant ideas and perspectives).

Shared Prosperity

AI systems are capable of generating efficiencies that not only improve an individual's quality of life, but fundamentally advantage an entire community. To reach the sustainable develop-

ment goals, these benefits should not be circumscribed by geography or socio-economic factors. Consultation participants believe that all members of society or their representatives should have access to tools that solve their most challenging problems, to the furthest possible extent.

> Access to AI systems should be widely distributed to guarantee that diverse populations can benefit from them to solve their most challenging problems. In addition, the achievement of the sustainability agenda is a shared responsibility of all parties and should be incorporated into the objectives of AI system designers, developers, and deployers.

Key commitments, pledges, and actions

The virtual consultation received 63 proposals for commitments, pledges, and actions to improve AI governance. During round two of the consultation, 205 votes and 23 comments on these proposals enabled the FLI team to prioritize these ideas (see Annex V for the list of prioritized commitments, pledges, and actions that influenced the proposals below).

To differentiate between proposals, we identified the entities with the notional responsibility for implementing each idea. Hence, this section is divided into five sub-sections, including: all stakeholders, the United Nations, regulating bodies, independent third parties, and designers, developers, and deployers of AI systems.

All Stakeholders

Regardless of their remit, all entities can be empowered to improve the national, regional, or global governance of AI. This sub-section highlights proposals applicable to any party wanting to help ensure that this technology contributes to achieving the sustainable development goals and minimizing risks to society.

RISK MANAGEMENT

All systems have the capability to generate direct and indirect harms. To mitigate these consequences, any party can and should serve a leading or supporting role in these efforts. For instance, designers, developers, and deployers can self-govern by adopting best practices in the form of a risk management framework or a human rights impact assessments to identify and classify risks. Independent third parties in the form of civil society, standards setting organizations, or multilateral institutions can operate as auditors to validate the implementation of risk management systems. Organizations that produce research can prioritize the funding and public distribution of AI safety outputs to identify and mitigate the harms emanating from these systems. Meanwhile, governments should feel empowered to proactively mandate actions deemed necessary to protect the well-being of users within their jurisdiction.

COORDINATION

We live in a world where thousands of principles, standards, best practices, strategies, and frameworks exist to manage the benefits and consequences of AI. While independent action

in this field can set an important precedent in complementing the UN's agenda on sustainable development, coordination is critical in guaranteeing the scalability of these actions. Therefore, all parties should devote resources to facilitate the reaching of a consensus that establishes which ideas are most effective and ensures their interoperability, so as to minimize implementation barriers.

In addition, all entities are responsible for contributing to the sustainable development goals. We ask that institutions in the private, public, and non-profit worlds assess how they can, independently or in concert, contribute to these objectives.

United Nations

As the sole multilateral forum with universal representation, the UN has a unique role in the AI governance space. Consultation participants highlighted two of the organization's comparative advantages: improving dialogue and generating public goods.

IMPROVING DIALOGUE

Existing fora dedicated to the discussion of AI governance are characterized by two types of limitations: geographic, where participating states are from a subset of countries or regions, or thematic, confined to specific issues or siloes with respect to AI (e.g., human rights, transportation, and the environment, among others). The UN is the only entity where all countries are represented and any issue affecting states can be discussed. With this in mind, consultation participants urge the UN to fulfill its role as chief catalyzer of collaboration, cooperation, and relationship-building between states in the field of AI governance.

GENERATION OF PUBLIC GOODS

The UN has unique experience in generating public goods that benefit society. With respect to AI governance, there are several opportunities for concrete action:

- Create a body similar to the Intergovernmental Panel on Climate Change dedicated to analyzing the most important AI challenges. Such an initiative would provide an authoritative body able to generate a consensus on critical issues with global repercussions.[2]
- Complement existing monitoring and reporting efforts with information on AI. For instance, the Common Country Assessments could include data on national technological developments, build an AI accident database, and provide a basis for multi-stakeholder dialogue on this technology. Prospective initiatives, such as the Futures Lab, could systematically measure and monitor the possible and real impact of AI systems and make this data public to improve stakeholder decision-making.
- For autonomous weapon systems, there is a need for an agreement on several types of procedures. These include best practices on incident reporting, de-escalation, auditing of AI-enabled military systems, non-proliferation, and strategic stability, as well as the consideration of a ban on non-human control over nuclear weapon launches.

2 There are proposals in the literature that advocate for the creation of an International Atomic Energy Agency for AI. In theory, such a body could have enforcement authority over systems that pose high-risks for society.

Regulating Bodies

All levels of government have tools at their disposal to compel or persuade parties within their jurisdiction to manage AI systems in a beneficial manner. The sub-sections below provide suggestions for how regulators can improve from within, increase their cooperation with other states, and protect the rights of their constituents through the implementation of safeguarding mechanisms.

BUILDING CAPACITY AND IMPROVING COORDINATION

As AI continues to push the boundaries of regulation, policymakers must prioritize how they prepare to engage with challenging issues in a knowledgeable and consistent manner. In this sense, all governments would benefit from actively developing and expanding the internal resources at their disposal for adequate oversight of high-risk AI systems.

This can be performed through several initiatives. One is the hiring or training of staff who can complement how policymakers at the executive, judicial, and legislative branches of government understand the consequences of AI. Another is the creation of external bodies composed of representative users impacted by AI to allow citizens to raise their concerns with policymakers on emerging issues.

A second type of engagement focuses on external and internal relationships. External relationships are those between different countries where a significant amount of learning can be accomplished through the exchange of best practices and strategies. Consultation participants encourage all countries to form bilateral or multilateral bonds that inform how they can best govern AI. On the other hand, internal relationships involve cooperation on the management of AI between distinct jurisdictions within a country. Concretely, participants believe it is important for national, regional, provincial/state, and local policymakers to create a shared understanding of their synergies in the management of AI.

SAFEGUARDING MECHANISMS

Consultation participants were in favor of mechanisms that safeguard consumer interaction with AI. In particular, they emphasized the protection from systems capable of making consequential decision-making. They believe that this issue can be addressed through requirements that allow individuals to challenge mistakes made by systems in use by the public or private sector. The ambition of this idea is to improve the transparency of AI and empower individuals to stand up for their rights.

Other ideas proposed by participants center on how governments could assess system risks. One is to compel impact assessments based on the level of risk, using frameworks such as those created by Canada in its Algorithmic Impact Assessment Tool.[3] Another is to develop a process where independent parties are given a mandate to test systems and audit the provenance of their training data based on pre-established thresholds of risk.

Independent Third Parties

Entities disconnected from the design, development, or deployment of AI systems (e.g., uni-

3 Treasury Board of Canada Secretariat, *Algorithmic Impact Assessment Tool*, (2021), https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html (last visited Apr 10, 2023).

versities, non-profits, and standard-setting organizations, among others) have a vital role in analyzing their impact on society and informing the public of their findings. Consultation participants believed that these institutions can fundamentally shape the responsible management of AI systems.

IDENTIFYING AND MONITORING RESPONSIBLE AI PRACTICES

Organizations without a stake in the commercialization of an AI system are unbound from the pressure to contribute towards its success. This allows the potential for objective analysis of a system's risks and benefits. One way society can take advantage of this independence is by having these entities dedicate their resources to identifying and broadcasting best practices and standards in managing AI. Increasing the awareness of prosocial ideas to mitigate AI's harms facilitates their incorporation into public and private practices throughout the world.

A second idea circulated by participants is to have these entities operate as auditors. In other words, the entities can monitor the management practices of designers, developers, and deployers and publicly highlight both positive and negative outputs. In this regard, several participants noted that although independence should be highly valued, the same goes for credibility. Not all organizations have equal standards for rigorous analysis. Therefore, support for an entity as an auditor should be routed through to those with a track-record of objective work that is aligned with the UN's agenda for sustainable development.

Designers, Developers, and Deployers of AI Systems

Entities that design, develop, or deploy AI systems have a contractual responsibility to their shareholders. At the same time, they have an implicit responsibility to protect the welfare of their users and society. This section emphasizes self-governance initiatives aimed at improving system safety. These initiatives should represent a baseline of actions implemented by all organizations responsible for the creation and commercialization of AI.

DESIGN AND DEVELOPMENT

Significant diversity exists in how entities approach the research and development process for state-of-the-art AI systems. Although there is no one "correct" process to minimize risks, it is imperative that entities devote substantial attention to follow the most up-to-date AI safety research, devote their own resources to identifying successful practices, and implement them as soon as possible. One concrete idea offered by consultation participants is the cultivation of an ecosystem of inspection. This means that entities should be open to the external scrutiny of their cutting edge AI systems and are empowered to subject them to public or independent pre-deployment inspection without the risk of prosecution.

DEPLOYMENT

The decision to make an AI system publicly available generates a duty to account for its impact on users. Norms that should characterize a responsible release strategy include:

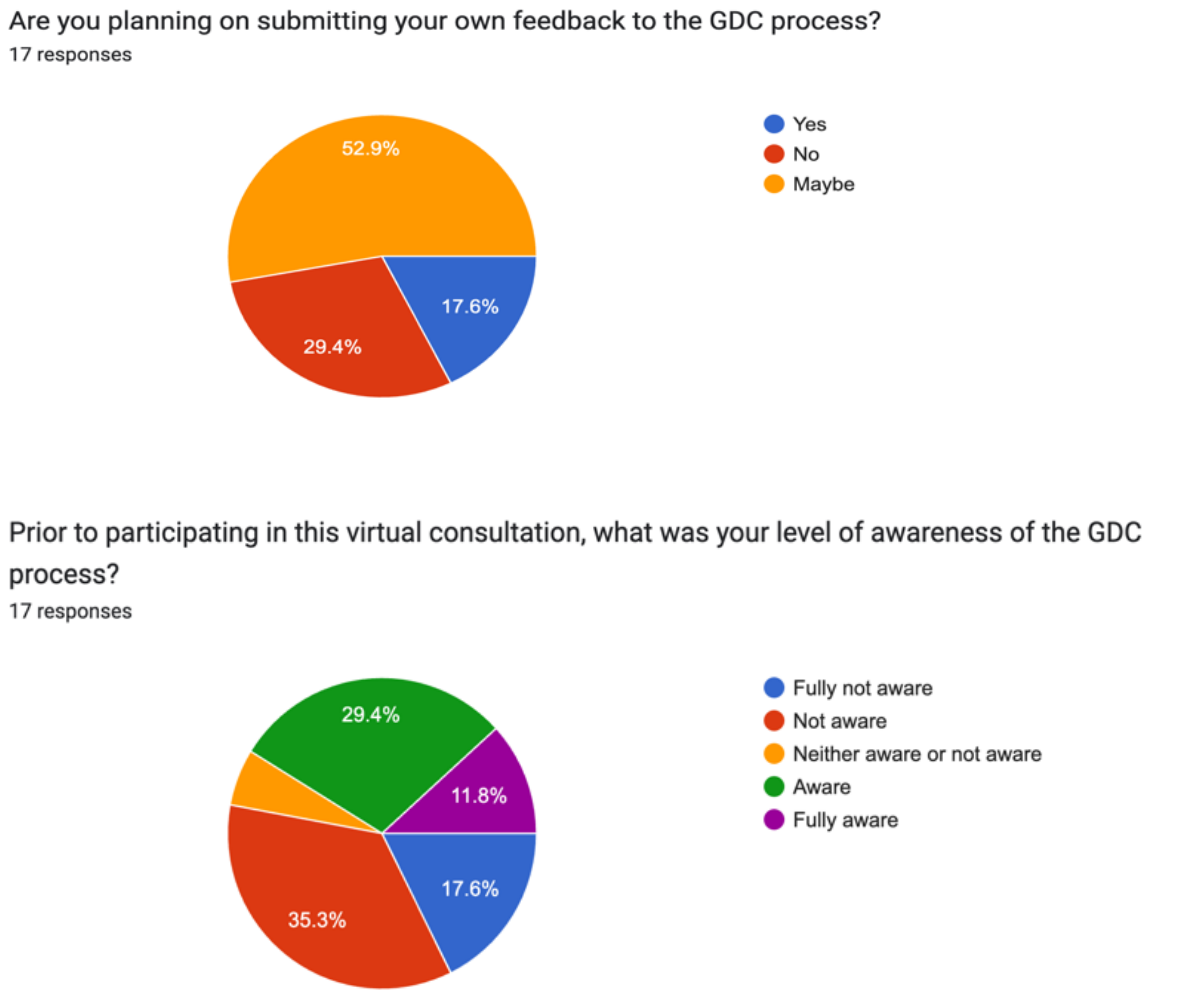
- Transparently documenting the system, which includes publishing its characteristics, capabilities, limitations, how it was trained, its intended purposes, as well as its risks and impacts.

- Guaranteeing that upstream and downstream entities acknowledge their level of responsibility for a system's harms.
- Analyzing and publicly declaring which uses would violate human rights.
- Actively monitoring high-risk and/or incorrect uses of a system and make attempts to disallow them when they occur.

Awareness + Plans to Submit Feedback to the GDC Process

As part of this consultation, FLI measured the engagement of our cross-section of AI governance institutions with the GDC process. In this regard, we received 34 responses and report two findings (see Graph 1 and 2). First, the majority of institutions (52.9%) were not aware of the GDC process. Second, an equal number of participants did not plan on submitting an independent response to the GDC.

Graph 1 and 2 – Responses to Awareness and Plans to Submit Feedback to GDC Process



Round 3 of the consultation provided an opportunity for several individuals to express their reasoning for the responses above. They stated the following:

- **Bandwidth limitations**: Many organizations do not have individuals with bandwidth ded-

icated to identifying and responding to multilateral efforts such as these.

- **Information asymmetry + uncertainty**: A potential reason why the GDC is not identified as a priority to invest bandwidth in is information asymmetry. Institutions lack data about how this process will shape multilateral AI governance, the likely impact of their submission on the eventual Summit of the Future, and the underlying forces that are managing the GDC process. In addition, the UN system has a large catalog of AI governance initiatives, and consultation participants were unsure about which ones they should invest their resources in.

Considering the above, a recommendation to improve awareness and participation in future initiatives related to AI governance is to **designate thematic and/or regional institutional ambassadors**. Doing so would incentivize the formation of coalitions and decentralize the UN's efforts to spread information regarding the importance of partaking in multilateral efforts. In return, formal collaboration with the UN generates clear positive externalities to entities selected to perform a concrete role. In particular, it legitimizes their interest to contribute towards shaping international technology governance and fortifies their standing with respect to peer organizations.

Role of the UN with Respect to AI Governance

The consultation included two questions, from which we received 34 responses, intended to inform UN stakeholders about the perception of their organization amongst AI governance organizations outside of the GDC process. We began by asking participants what they believed was the UN's role in this field. The overwhelming agreement is that it should maximize its comparative advantage, which is to function as the only truly global multilateral convenor. This convening power allows unheard voices in the developing world to present their perspectives on the benefits and challenges of AI; this in turn facilitates the management of economic and military disagreements about this technology's use.

A secondary objective for the UN should be to produce AI-centered public goods that benefit society. This can be in the form of analyses that spread awareness of this technology's benefits and risks. A concrete avenue of action proposed by a consultation participant is for the UN to explore the links between AI and the universal declaration of human rights. The latter is, after all, a technology-neutral instrument that has withstood the test of time and should be regarded as the main driving force in protecting individuals from risks. In addition, the UN is well-placed to identify mechanisms for countries and entities to follow this document's precepts via the alignment of incentives (e.g., increasing the reputational costs of not following these mechanisms and reducing the ability to externalize harms).

The last question posed to participants was what role should not be pursued by the UN in the field of AI governance. Participants expressed less of a desire for initiatives where few incentives exist for enforcement or implementation. Examples include the development of new principles or frameworks that do not consider the importance of aligning the interests of stakeholders. Instead of aiding AI governance, these efforts could serve as a means for entities to 'ethics-wash' their efforts. Finally, participants were weary that, despite the UN's

ability to foster discussion, the conflicting interests between nation states may make it difficult or impossible to reach a consensus on AI regulatory guidance that effectively prioritizes democracy and human rights.

Annex I – List of Participants

Organization	Name
Ada Lovelace Institute	Connor Dunlop
AI Governance and Safety Canada	Wyatt Tessari L'Allié
	Brianna Brownell
Brookings	Chris Meserole
C Minds	Claudia May Del Pozo
Canadian Institute for Advanced Research	Brent Barron
	Rachel Parker
Carnegie Endowment for International Peace	Matthew O'Shaughnessy
	Hadrien Pouget
	Aubra Anthony
Center for Long-Term Cybersecurity	Jessica Newman
Centre for Long-Term Resilience	Jess Whittlestone
Centre for the Governance of AI	Eoghan Stafford
Centro Nacional de Inteligencia Artificial	Rodrigo Duran
	Maria Jose Escobar
	Pablo Barcelo
	Claudia Lopez
	Gabriela Arriagada
	Rolando Martinez
	Marcelo Mendoza
Clarkson University	Jeanna Matthews
Concordia	Brian Tse
Corporación Andina de Fomento (Consultant)	Armando Guio
DeepMind	Sebastien A. Krier
	Alexandra Belias
DiploFoundation	Katharina Höne
Future of Life Institute	Carlos Ignacio Gutierrez
GobLab - Universidad Adolfo Ibáñez	Romina Victoria Garrido Iglesias
	Maria Paz Hermosilla Cornejo
International Centre of Expertise in Montréal on Artificial Intelligence	Mathieu Marcotte
	Lama Saouma
Legal Priorities Project	Matthijs Maas
Organisation for Economic Co-operation and Development	Luis Aranda
	Lucia Russo
Partnership on Artificial Intelligence	Rebecca Finlay
	Stephanie Ifayemi
Responsible Artificial Intelligence Institute	Ashley Casovan
	Var Shankar
Simon Institute for Long-term Governance	Konrad Seifert
The Future Society	Yolanda Lannquist
University of Richmond	Anne Toomey McKenna

Annex II – Round 1: Survey

Global Digital Compact - UN Virtual Consultation on Artificial Intelligence

The team at FLI is incredibly grateful for your participation in our virtual consultation. Please note the following:

- This round of feedback will be open from March 1st to the 10th.
- Your answers to this form are anonymous.

If you have any comments or questions, please contact:
Carlos Ignacio Gutierrez at carlos@futureoflife.org

*Required

1. The first prompt to the Global Digital Compact (GDC) submission is: "Core principles that all governments, * companies, civil society organisations and other stakeholders should adhere to."

   Principles are broad statements that serve as high-level norms for any type of organization. They are unlikely to include actionable items, instead they are an organization's aspirational objectives.

   We are seeking your written contribution on which principles should be included in the GDC that are directly related to artificial intelligence (AI) governance and the UN 2030 agenda. In terms of format, we recommend selecting ideas you deem relevant and supplementing them with a concise description.

   As guidance, there are hundreds of principle sets in AI governance. This project by Arizona State University compiled 158 sets (the original database can be found here). Another project by Harvard found several dozens. For your convenience, FLI has compiled the feedback provided to the GDC on AI governance at this link.

2. The second prompt of the GDC is: "Key commitments, pledges, or actions that in your view should be * taken by different stakeholders – governments, private sector, civil society, etc."

   Responders may decide to link their suggestions to the principles above or make them independent. In addition, we urge you to think about the viability of your feedback. If proposing direct government action, also known as hard law, we ask that you state concrete actions and support them with examples. If you opt for suggesting self-governance methods, known as soft law, it is critical to address or mention the incentives needed for stakeholders to adopt such initiatives.

3. Prior to participating in this virtual consultation, what was your level of awareness of the GDC process? *

   Mark only one oval.

   - ◯ Fully not aware
   - ◯ Not aware
   - ◯ Neither aware or not aware
   - ◯ Aware
   - ◯ Fully aware

4. Regardless of the GDC process, what do you believe should be the role of the UN with respect to AI * governance?

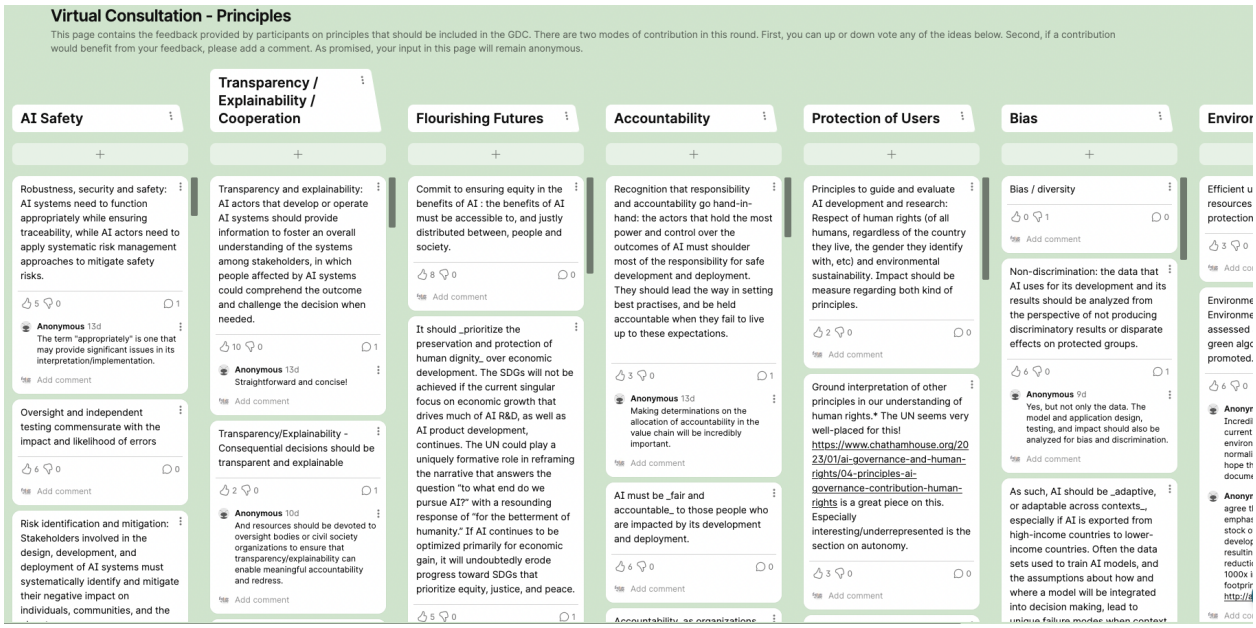
5. Do you believe that there is an area of work, related to AI governance, that the UN should stay away from? * If yes, what would it be?

6. Are you planning on submitting your own feedback to the GDC process? *

   Mark only one oval.

   - ◯ Yes
   - ◯ No
   - ◯ Maybe

Annex III – Round 2: Voting and Commenting System



Annex IV – Prioritized Principles

Mitigating Risks throughout the AI Lifecycle

- Safety: AI systems should function reliably, and they should not pose a risk of harm to living beings or the environment in which they are deployed, as defined by the European High-Level Expert Group on AI. The principle of safety must again be incorporated at every stage in the AI lifecycle.
  - During development, AI systems should be subjected to rigorous testing and evaluation that can assess their risk of failure or misuse.
  - After deployment, AI operators should continue to test and monitor the performance of the system and be prepared to take action if the system fails in unexpected ways.
- Risk identification and mitigation: Stakeholders involved in the design, development, and deployment of AI systems must systematically identify and mitigate their negative impact on individuals, communities, and the planet.
- Human-centered values and fairness: The values of human rights, democracy, and rule of law should be incorporated throughout the AI system's lifecycle, while allowing human intervention through safeguard mechanisms.
- Robustness, security and safety: AI systems need to function appropriately while ensuring traceability, while AI actors need to apply systematic risk management approaches to mitigate safety risks.
- Considering society-level outcomes: In general, principles have been aimed at individual systems and users, rather than society-level outcomes. While they are of course related, including principles aimed at society-level outcomes is especially important for governments (which the UN is well-placed to advise!). These could include:
  - Guarding against misuse and abuse of systems, especially as they become more prolific
  - Protecting the information environment of people (deepfakes, spam, etc, related to autonomy point above)
  - Concentration of power, wealth, etc.
  - Considering complex interactions between AI systems, which may not always be obvious but can result in negative outcomes even if each system seems individually sound

Transparency and Explainability

- Oversight and independent testing commensurate with the impact and likelihood of errors
- Transparency: AI systems should be designed and deployed in a way that allows people to understand when they are being used and how they were developed and trained. Actors should commit to transparency in AI systems to foster increased understanding, allow individuals to make more informed decisions, and to make oversight of AI possible.

Furthermore, the use of AI in predictions, recommendations, or decisions should be disclosed to ensure public trust and visibility. Transparency is included in a wide variety of AI principles, including those articulated by the OECD.

- Explainability: AI systems should be designed to provide the evidence, support, or reasoning that produced a given outcome in a way that is interpretable by humans. As noted by NIST in the US, explainability can build public trust in the deployment of AI systems and provide redress to those impacted by their decisions. AI developers and operators should ensure that those affected by an AI system are able to understand the outcome based in information that, in line with NIST's principles of explainability, is meaningful and accurate.
- Build in transparency across the AI lifecycle: AI systems are inherently opaque and difficult to understand. All stakeholders developing and deploying such systems must adhere to actions supporting explainable and interpretable AI – from the research and development of new AI technologies and the design of AI products through to the deployment of AI systems and their interactions with affected persons.
- Transparency and explainability: AI actors that develop or operate AI systems should provide information to foster an overall understanding of the systems among stakeholders, in which people affected by AI systems could comprehend the outcome and challenge the decision when needed.
- Meaningful two-way communication between developers and impacted users. (4.1) Developers and deployers of automated systems should obtain meaningful consent from users, and provide a meaningful option to opt-out. (4.2) Developers should accurately, clearly, and proactively communicate the capabilities and limitations of their products. (4.3) Developers and deployers should consult with impacted users and communities at all stages of development. (4.4) Structures should be developed and supported to enable information flow to and from impacted communities that are geographically, linguistically, and socioeconomically diverse. (4.5) Consideration should be given to the rights and well-being of impacted groups at all stages of development.
- Right to redress or review. Meaningful access to humans empowered to investigate errors and harms. Those human teams must be incentivized to find errors and identify harms. System maintainers must be incentivized to fix those errors and mitigate harms.

Accountability of Stakeholders

- AI governance should strengthen democratic accountability and individuals' fundamental human rights, rather than concentrate political, economic, or other forms of power in the hands of elites.
- Accountability: AI developers and operators must ensure that these systems function properly throughout their life cycle. AI must be designed and deployed in accordance with applicable regulatory frameworks and systems should be in place at each stage in the AI lifecycle to assign responsibility and monitor compliance by the relevant human parties. Accountability for human actors features prominently in a number frameworks, including

UNESCO's Recommendation on the Ethics of AI.

- Accountability: norms and legal structures should be defined that disincentivize (and provide meaningful recourse for) potential harms. (3.1) Policy and norms should be developed that counter the ways in which AI systems make traditional recourse more difficult. For example, the EU's proposed AI Liability Directive offers policy solutions that account for the opacity of AI systems. (3.2) Provide norms and requirements for the sharing of information by the developers and deployers of AI systems that encourage transparency and a right to opt out. (3.3) Development of, and support for, institutions that represent under-resourced individuals and communities in obtaining recourse from harms caused by AI systems. (3.4) Processes, norms, and regulation that reduce the externalization of potential harms, such as potential future threats to privacy that stem from the collection and use of personal data.
- AI must be _fair and accountable_ to those people who are impacted by its development and deployment.

Fairness

- Non-discrimination: the data that AI uses for its development and its results should be analyzed from the perspective of not producing discriminatory results or disparate effects on protected groups.
- As such, AI should be _adaptive, or adaptable across contexts_, especially if AI is exported from high-income countries to lower-income countries. Often the data sets used to train AI models, and the assumptions about how and where a model will be integrated into decision making, lead to unique failure modes when context isn't appropriately factored into AI's design and deployment. Prioritizing adaptability or contextual relevance of models would better accommodate limitations arising from the global supply chain that often underpins AI at scale.
- AI must be _trustworthy and non-discriminatory_. AI governance should foster trust, but for that to be achieved, the technology must be trustworthy. Trust is intimately connected to power, and as such, AI's use must not exacerbate or reinforce existing power asymmetries and inequities. There must also be systems of accountability in place to consistently test and reaffirm trustworthiness and non-discrimination.
- Fairness and inclusivity: AI systems should be developed in a manner consistent with fundamental principles of equality and fairness. AI tools possess a strong potential for bias, discrimination, and reification of existing societal inequities. They should be assessed on measures of fairness, and both standards and the technologies themselves should be developed through inclusive processes that seek input from impacted stakeholders that may otherwise be left out. Fairness and nondiscrimination were included as principles in 100% of the frameworks analyzed by the Harvard study cited above.
- justice and equality: developers and regulators of algorithmic systems should consider how potential benefits and harms are distributed across groups defined by traits such as geography, socioeconomic status, race, ethnicity, gender, etc. (2.1) Structures that allow

for the co-design of algorithmic systems with impacted communities should be developed and enhanced. (2.2) Developers should make public the results of this analysis, as well as making available data and other information needed for communities, civil society, and government to conduct independent analysis of its impacts. (2.3) Developers, civil society, governments, and intergovernmental bodies should strive to ensure that access to the benefits of new algorithmic systems are broadly and equitably distributed

- AI should _preserve the pluralistic aspects of humanity_. Care should be taken to preserve non-dominant cultures, non-dominant worldviews, and non-dominant languages in the way AI is constructed and integrated into society. Similarly, AI principles themselves must not conform overmuch to one worldview but be flexible enough to accommodate multiple approaches to understanding and reflecting values/norms.

Shared Prosperity

- Commit to ensuring equity in the benefits of AI: the benefits of AI must be accessible to, and justly distributed between, people and society.
- It should _prioritize the preservation and protection of human dignity_ over economic development. The SDGs will not be achieved if the current singular focus on economic growth that drives much of AI R&D, as well as AI product development, continues. The UN could play a uniquely formative role in reframing the narrative that answers the question "to what end do we pursue AI?" with a resounding response of "for the betterment of humanity." If AI continues to be optimized primarily for economic gain, it will undoubtedly erode progress toward SDGs that prioritize equity, justice, and peace.
- Shared benefit and prosperity: The deployment of AI-based technologies should advance a flourishing future that empowers, improves the social and economic condition, and decreases inequality for as many individuals as possible.
- Environmental protection. Environmental impacts should be assessed and the development of green algorithms should be promoted.

Annex V – Prioritized Commitments, Pledges, and Actions

All Stakeholders

- Find novel solutions for R&D governance: Academic research has shown that, at present, "far greater emphasis is placed on deployment in downstream applications, neglecting risks arising during research and development. For example, while several leading AI companies have agreed to a set of best practices for safely deploying large language models, no (publicly known) agreement has been reached on safety protocols for research and development. Similarly, while there exists a database of safety incidents encountered in the deployment of automated systems, there is no equivalent platform for reporting safety incidents encountered during AI research and development". We have seen this issue arise in the context of the EU AI Act, and would therefore compel all stakeholders to set guardrails via R&D governance and compelling responsible release strategies by upstream developers.
- Governments, private sector stakeholders should commit to performing Human Rights Impact Assessments for high-risk uses of AI, and to meaningfully engaging with civil society around findings from these assessments.
- Coordinate on international monitoring and horizon scanning: activities for identification of potentially risky models will be important. This could entail measuring the capabilities of state-of-the-art AI systems, mandatory disclosure regimes in which AI developers audit their own systems (or third parties do so) and report findings to policymakers, and/or whistleblower protections to empower individuals with access to relevant information (such as software engineers at leading AI labs) to disclose information pertaining to automated systems that pose large-scale societal risks.
- In service of principles of transparency and accountability, a key commitment that can be undertaken by all AI actors is documentation. Documentation standards are included throughout the NIST AI Risk Management Framework for example, and should include components such as characteristics, capabilities, and limitations of an AI system, how it was trained, its intended purposes, as well as its risks and impacts.
- Prioritizing AI safety research: AI Safety is the field of research into "techniques for building AI that is beneficial for humans." The challenges in harnessing AI systems in a manner that minimizes harms to individuals are continuously evolving. Yet, resources to support this field pale in comparison to the increasing pace of deployment for AI-based products and services. In other words, society is prioritizing short-term benefits over long-term resilience and stability. Hence, all stakeholders should re-evaluate the importance of AI safety funding in their research portfolios.
- In service of principles of human centered values and fairness and others, another key commitment that can be undertaken by all AI actors is to make public declarations to avoid development and uses of AI that would violate human rights, for example cases that would put people's lives at risk or violate the right to privacy, protection from gender or sexual violence, protection of children's rights, freedom of expression, or the right to fair

trial and peaceful assembly.

- Adopt a risk-based approach to AI regulation: Both the EU and OECD have employed a risk-based approach to AI. The EU considers four categories of risk (unacceptable risk, high risk limited risk, and minimal risk) and limits application to only "where strictly ended and in a way that minimizes the burden for economic operators, with a light governance structure." Alternately, the NIST AI Risk Management Framework offers a risk-based approach that attempts to be scalable to organizations and AI across disciplines. Clear identification and classification of risks, convergence on cases where risks are too high to be mitigated, and convergence on the type of risk assessment to be performed will be an important step towards a system of responsible global AI governance.
- Many organisations are developing frameworks to assess risks and impacts in AI, including NIST in the US, the Council of Europe, CEN-CENELEC, IEEE and ISO, among others. It is crucial however to ensure that these frameworks are as interoperable as possible, else we risk making implementation of Trustworthy AI way more complex and costly in practice, and therefore less effective and less enforceable. This is a large risk that will inhibit innovation.
- Proactive participation in the UN's sustainability agenda: As society has thus far witnessed, there are clear advantages in harnessing AI to solve complex challenges. In parallel, achieving the sustainable development goals requires ingenuity in deciphering the combination of levers and actions able to solve the myriad of constraints faced by governments throughout the world. Thus, there is an overwhelming need to marry the efficiencies and useful outputs of AI with the sustainability agenda. Considering the above, we ask that all entities participating in the AI lifecycle proactively participate in the UN's sustainable development agenda.

United Nations

- One could aim for the Integration of AI development monitoring into Common Country Assessment (CCA). This could be an efficient way to monitor developments as well as build an AI accident database, and provide a basis for multi-stakeholder dialogues on the speed of technological change and its implications.
- Multi-Stakeholder Advisory Body: We support the UN's Roadmap for Digital Cooperation recommendation on AI stating the need for the creation of a "multi-stakeholder advisory body on global AI cooperation." Its objective is to decrease the barriers of entry for all parties interested in multilateral discussions regarding the benefits and harms of AI. This advisory body can accomplish a variety of goals with the UN system: build capacity for the development and use of AI, combat the lack of representation and inclusiveness in global AI discussions, and coordinate how countries harness AI in support of the UN's sustainability agenda.
- Disarmament and confidence-building initiatives require future-proofing (e.g. NPT, IAEA, or new fora). One could establish international dialogues on autonomous incident agreements (procedures for de-escalation) and on auditing AI-enabled military systems. Much

of the conversation at the LAWS GGEs focuses narrowly on the risk to civilians of autonomous drones. There should also be scope for the discussion of escalation, non-proliferation, and strategic stability. The GDC could recommend a general ban on non-human control over nuclear weapons platform launches, building on commitments from nuclear weapons states, such as the United Kingdom and the US.

- High-Level Panel of AI Experts: Society is facing a deluge of challenges, both short and long-term, emanating from AI. We find ourselves in a state of affairs where parties involved in the AI lifecycle or its governance cope with issues in a decentralized manner. In other words, many of the lessons learned about mitigating AI risks are either selectively shared or remain secret to protect a competitive advantage. Missing is a credible body that generates consensus-based assessments on fundamental AI challenges affecting society. Much like the Intergovernmental Panel on Climate Change, the UN would benefit its sustainability agenda by establishing a trustworthy body of AI experts. Its objective would be to serve as an informational clearinghouse that generates recommendations to address society's most important AI problems. Unlike any other national, private sector, academic, or multilateral institution, the UN has gained the credibility and experience to organize and manage this effort in a prosocial manner.
- The UN Futures Lab should systematically measure and monitor possible and real impacts of AI systems. International bodies can aggregate information to understand AI's global impacts and potential development trajectories. This information can be made publicly available to promote transparency and improve deployment decisions at national or corporate levels.

Regulating Bodies

- Commit adequate resources for regulatory capacity: Government and regulators will require substantial investment to ensure they have the necessary staffing, experience, knowledge and training to properly govern AI. It is not just technical or digital skills that will be required, but an array of expertise, from law, ethics, and philosophy to mathematics and science. Greater coordination between regulators will be necessary to ensure cross-sectoral enforcement.
- Build governance mechanisms for the representation of affected persons: A recent academic report concluded that the main way the EU's AI governance regime can be strengthened is with more "effective citizen engagement". One mechanism for this would be to have a standing panel of representative users – a type of 'citizens assembly' as suggested in the aforementioned report – as a permanent sub-group of the AI Board. Ada's Citizen's Biometrics Council (in the UK) has been highlighted as a leading example of this type of engagement, and we want to work with policymakers to explore if and how such an approach can be developed under the AI Act
- Hard law should require independent testing throughout the lifecycle of an automated system that is commensurate with the level of impact of the system including the severity of possible impacts of life and well-being and the probability of those impacts. Require-

ments for independent testing, data provenance should be higher the more severe the possible impact of system error and the higher the likelihood of system error.

- Compel impact assessments for risky AI systems: Impact assessments (IA) are a well-established method used to assess human rights, equalities, data protection, financial and environmental impacts of a policy or technology ex ante. Algorithmic impact assessments (AIA) are being explored in a number of jurisdictions, notably Canada, and are already being employed by AI developers (e.g. Microsoft's Responsible AI initative). We would call on Governments to give legislative footing for developers and deployers to conduct impact assessments for risky AI systems.
- I would also like to see hard law that focuses on requirements/processes for redress. There needs to be functional processes to connect with humans who are empowered to investigate and mitigate errors and harms from AI systems. It should not only be powerful/wealthy individuals who will have access to human review, investigation and meaningful redress.
- Governments should develop mechanisms of accountability for citizens to interrogate and contest the use of AI, and systems of transparency (and explanation) for citizens to understand how the use of AI impacts them. These should be present in both public and private sector use of AI.

Independent Third Parties

- Some AI organizations might decide to design and implement such policies on their own, to earn the public's trust. However, human rights watchdogs and other civil society groups can increase the incentives for responsible behavior by proposing public standards for what technologies should not be transferred to governments and what uses are beneficial or harmful, verifying AI organizations' adherence to those standards, and publicly naming and shaming those that facilitate repressive surveillance. (While these voluntary actions by the private sector and civil society would be helpful, I do not think they are sufficient by themselves. Democratic legislatures should also create legal regulations on AI organizations to prevent them from facilitating repressive surveillance at home or abroad.)
- Civil society - should monitor governments' and the private sectors' adherence to ethical and human rights principles in their use and regulation of algorithmic systems and publicize major deviations. Civil society will require adequate resources to safeguard the rights and well-being of under-resourced individuals and communities.

Designers, Developers, and Deployers of AI Systems

- Find novel solutions for R&D governance: Academic research has shown that, at present, "far greater emphasis is placed on deployment in downstream applications, neglecting risks arising during research and development. For example, while several leading AI companies have agreed to a set of best practices for safely deploying large language models, no (publicly known) agreement has been reached on safety protocols for research and development. Similarly, while there exists a database of safety incidents

encountered in the deployment of automated systems, there is no equivalent platform for reporting safety incidents encountered during AI research and development". We have seen this issue arise in the context of the EU AI Act, and would therefore compel all stakeholders to set guardrails via R&D governance and compelling responsible release strategies by upstream developers.

- In service of principles of transparency and accountability, a key commitment that can be undertaken by all AI actors is documentation. Documentation standards are included throughout the NIST AI Risk Management Framework for example, and should include components such as characteristics, capabilities, and limitations of an AI system, how it was trained, its intended purposes, as well as its risks and impacts.
- In service of principles of human centered values and fairness and others, another key commitment that can be undertaken by all AI actors is to make public declarations to avoid development and uses of AI that would violate human rights, for example cases that would put people's lives at risk or violate the right to privacy, protection from gender or sexual violence, protection of children's rights, freedom of expression, or the right to fair trial and peaceful assembly.
- Actively cultivate an ecosystem of inspection: Developers and deployers must be open to scrutiny, which means access for researchers and other stakeholders to cutting edge AI models, and the ability to carry out external inspections (e.g. independent audits, red-teaming) without running the risk of prosecution (by offering safe harbour for researchers). These provisions would build an "ecosystem of inspection", which an Ada report argues can help overcome the challenges of finding adequate capacity for technical oversight, while a Stanford report argues that because ""foundation models can be adapted to myriad downstream applications, no organization can fully anticipate all potential risks Consequently, it is imperative that external researchers representing a diversity of institutions, cultures, demographic groups, languages, and disciplines be able to critically examine foundation models from different perspectives"
- Distribution of economic benefits: If current trends stand, high-added value AI systems will continue to generate substantial profits over the coming decades. This will lead to the creation of firms dedicated to capturing market share through novel innovations. We believe that a proportion of firms dedicated to the AI lifecycle will opt to share their bottom line with the rest of the world as a "social dividend." Considering the existence of proposals such as the "windfall clause," we ask that the UN have a role in assessing these initiatives as a means of fundings its sustainability agenda directly.