March 2023

Submission to the Global Digital Compact

Contact
carlos@futureoflife.org

Dr. Amandeep Singh Gill
Secretary-General's Envoy on Technology
United Nations
New York, NY

Dear Dr. Amandeep Singh Gill,

The Future of Life Institute (FLI) is an independent non-profit organization that works on fostering the benefits of technology and mitigating its risks. FLI created one of the earliest sets of AI governance tools - the Asilomar Artificial Intelligence (AI) principles - and maintains a large network of global AI researchers. In addition, we participated in the United Nations (UN) Secretary General's Digital Cooperation Roadmap as the civil society champion for AI.

FLI welcomes the opportunity to provide feedback on the Global Digital Compact (GDC) organized by the Office of the Secretary-General's Envoy on Technology. In preparation for the Summit of the Future, we wholeheartedly agree that digital technologies should help to achieve the UN's 2030 agenda on sustainable development. Moreover, all stakeholders charged with designing, developing, deploying, and governing these technologies must take responsibility as stewards of a future where life is protected.

Our organization cares deeply about the impact of AI on society. We believe this technology holds great promise to solve important global challenges, but it can also catalyze significant barriers to sustainable development. Because of this, our feedback for the GDC process focuses exclusively on AI governance. In this document, we propose four core principles accompanied by key commitments, pledges, and actions. It is our hope that these ideas will complement the UN's efforts to support the GDC's prosocial agenda through the most effective means available. If you have any questions or comments about the information in this document, please contact Carlos Ignacio Gutierrez, FLI's UN Representative and AI Policy Researcher, at carlos@futureoflife.org.

Regards,

The Future of Life Institute

Core principles

- Risk identification and mitigation: Stakeholders involved in the design, development, and deployment of AI systems must earnestly and systematically identify and mitigate their negative impact on individuals, communities, and the planet.
- Technological partnership: Cooperation between parties designing, developing, deploying, or governing AI systems should be prioritized over profits, politics, and power asymmetries.
- Shared benefit and prosperity: The deployment of AI-based technologies should advance a flourishing future that empowers as many individuals as possible, improves their social and economic condition, and decreases inequality.
- Loyalty: AI systems should, first and foremost, consider the user or consumer's interests when completing tasks.

Key commitments, pledges, or actions

The four core AI governance principles proposed by FLI for the GDC are the basis for the key commitments, pledges, and actions we believe are critical to achieve the UN's agenda on sustainable development. We first provide a summary of our ideas below followed by seven sections that go in depth into each of them.

- Risk identification and mitigation
  - Stakeholders should prioritize the allocation of resources into AI Safety research, a field specialized in developing techniques for building AI that is beneficial to humans, over the rapid and uncontrolled deployment of AI systems.
  - Entities that participate in any part of the AI lifecycle should put into practice a risk management system that, to the degree possible, effectively identifies and addresses harms emanating from these systems.
- Technological partnership
  - We propose the creation of an advisory body within the UN where multilateral discussions on any governance issue related to AI can take place without the jurisdictional or thematic restrictions present in existing fora.
  - The UN should establish a trustworthy body of AI experts whose objective is to serve as an informational clearinghouse that generates recommendations to address the most important AI problems facing society.
- Shared benefit and prosperity
  - The UN has a role in connecting entities to its sustainability agenda via the publication of high-impact challenges open to any AI product or service providers.
  - We ask the UN to assess methods for distributing the economic benefits of AI to the developing world with the objective of solving some of the world's most pressing problems.
- Loyalty
  - We recommend that the notion of "loyalty" be recognized as an integral element in national and multilateral discussions that examine the characteristics of trustworthy AI systems.

Risk identification and mitigation

AI systems have the potential to generate direct and indirect harms that derail the achievement of the UN's sustainable development goals. Increasingly, the geographic, demographic, or socio-economic borders that protect individuals from this technology's effects are falling away. This should prompt the GDC to invite stakeholders involved in the design, development, and deployment of AI systems to identify and mitigate their negative impact.

Our suggestions for identifying and mitigating the risks of AI systems are divided into research commitments and actions concerning deployment. First, we strongly believe that the existing gap between the resources allocated to AI safety research and the commercialization of AI systems needs to be addressed. Second, stakeholders involved in any aspect of the AI lifecycle must prioritize the implementation of an effective risk management system to prevent harms.

Prioritizing AI safety research

AI Safety is the field of research into "techniques for building AI that is beneficial for humans."[1] The challenges in harnessing AI systems in a manner that minimizes harms to individuals are continuously evolving.[2] Yet resources to support this field pale in comparison to those dedicated to the rapid deployment of AI-based products and services. In other words, society is prioritizing short-term benefits over long-term resilience and stability. Thus, our recommendation is to request that stakeholders re-evaluate the importance of AI Safety funding in their research portfolios.

All of society benefits from devoting resources to a field whose main goal is to minimize the harms and improve the benefits of AI technology. Therefore, stakeholders must work together to identify the most pressing issues in AI safety and channel sufficient funding to address them. Importantly, when funding is allocated to this field, it must be distributed in a manner that builds research capacity around the world. This not only improves the diversity of viewpoints able to generate solutions, but also represents an opportunity to address those issues otherwise neglected in the geographies where leading work in this field currently occurs.

Implementation of an effective risk management system

Organizations with a role in the AI lifecycle are responsible for protecting society against its potential harms. This requires a concerted effort, in which reactive and proactive measures are generated with the explicit purpose of ensuring the safety of individuals. With this in mind, we recommend that all organizations put into practice a risk management system that, to the greatest possible degree, effectively identifies and addresses harms emanating from AI.

There are several vectors of action that a risk management system needs to address. For instance, much like biosafety laboratory levels, research entities or laboratories dedicated to creating AI systems should secure their infrastructure to avoid data or system leakages.[3] Similarly, whistleblower programs are critical in highlighting potential managerial or system hazards. Lastly, a systematic evaluation of how an AI system will negatively affect individuals from different jurisdictions and cultures is fundamental.

Success with this recommendation can take many shapes. On the one hand, governments could take a hands-on approach using regulation or 'hard law'. The strength of this alternative is that entities are forced to comply with the law if they want to distribute their system in a jurisdiction. An example is the European Union's (EU) proposed AI Act, which has the potential to become the most comprehensive

1 Mislav Juric, Agneza Sandic & Mario Brcic, *AI safety: state of the field through quantitative lens*, in 2020 43RD INTERNATIONAL CONVENTION ON INFORMATION, COMMUNICATION AND ELECTRONIC TECHNOLOGY (MIPRO) 1254 (2020); Future of Life Institute, *FLI AI Safety Research Landscape - Extended - v0.43*, https://futureoflife.org/valuealignmentmap/ (last visited Mar 2, 2023).

2 Dario Amodei et al., *Concrete problems in AI safety*, ARXIV PREPRINT ARXIV:1606.06565 (2016).

3 Centers for Disease Control and Prevention, *Infographic: Biosafety Lab Levels*, (2023), https://www.cdc.gov/orr/infographics/biosafety.htm (last visited Mar 2, 2023).

such effort to date.[4] In the Act, use cases of AI are categorized by their perceived risk ranging from low to unacceptable, each with different requirements to ensure their safety. In particular, systems that pose a high risk are required to generate a risk management system to attain EU approval.[5]

Another example of hard law, supported by FLI, is the proposal for a legally-binding instrument on autonomous weapons systems, with a two-tier approach of prohibition and regulation. This is currently being discussed at the UN Convention on Certain Conventional Weapons (CCW) and is supported by over 80 countries. The development of legally-binding rules on this issue, rather than voluntary guidelines, would ensure that the strongest form of global governance is utilized to prohibit a class of weapon that threatens peace around the world. International law, in this respect, would provide a clear and homogeneous framework by which governments can abide when using weapons that have autonomous functions and that therefore require meaningful human control, predictability, explainability, and traceability.

Alternatively, self-governance or 'soft law' measures are a flexible alternative to direct government action. These are defined as "instruments or arrangements that create substantive expectations that are not directly enforceable" by government.[6] There are hundreds of these measures in existence that focus on an increasingly wide variety of AI issues.[7] FLI's own Asilomar AI Principles represented one of the first efforts to codify a set of desired practices in the generation of highly-capable systems.[8] Recently, the National Institute of Standards and Technology of the United States Government published their AI Risk Management Framework, intended to be implemented by any stakeholder.[9]

Technological partnership

We ask stakeholders to prioritize AI-related cooperation to fulfill the UN's sustainable development goals. Our belief is that all parties would generate significant value from minimizing technological 'races to the bottom' - prompted by economic, military, or political asymmetries. As a result, both our recommendations are aimed at consensus building efforts that leverage the UN's strength as a global convener.

Below, our suggestions center on the creation of bodies meant to catalyze discussion and partnership between AI stakeholders. The first body centers on inviting all representatives of member states to engage in AI governance discussions where any nation can participate without restrictions. The second body is an expert group that generates consensus-based recommendations to find trustworthy approaches to solve AI challenges. FLI considers that the creation of these two groups would enhance the UN's ability to reach its sustainable development goals.

Multi-Stakeholder Advisory Body

As a civil society champion in the UN's Roadmap for Digital Cooperation, we are committed to supporting efforts to create a forum that welcomes all countries in pursuit of dialogue and cooperation regarding AI. In this respect, one of our recommendations is to create a "multi-stakeholder advisory body on global AI cooperation."[10] Its objective is to decrease the barriers of entry for all parties

4 European Union, *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*, (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (last visited Oct 13, 2021).

5 Future of Life Institute, *Lessons from the NIST AI RMF for the EU AI Act*, (2022), https://futureoflife.org/wp-content/uploads/2022/08/Lessons_from_NIST_AI_RMF-v2.pdf.

6 Gary E Marchant & Brad Allenby, *Soft law: New tools for governing emerging technologies*, 73 BULLETIN OF THE ATOMIC SCIENTISTS 108 (2017).

7 CARLOS IGNACIO GUTIERREZ & GARY E. MARCHANT, *A Global Perspective of Soft Law Programs for the Governance of Artificial Intelligence*, (2021), https://papers.ssrn.com/abstract=3855171 (last visited Sep 30, 2021).

8 Future of Life Institute, *AI Principles*, FUTURE OF LIFE INSTITUTE (2017), https://futureoflife.org/ai-principles/ (last visited Oct 1, 2021).

9 NIST, *AI Risk Management Framework*, NIST (2021), https://www.nist.gov/itl/ai-risk-management-framework (last visited Dec 25, 2022).

10 United Nations, *PROVIDING GLOBAL STEERAGE ON ARTIFICIAL INTELLIGENCE*, (2021), https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/general/Artificial_Intelligence_Summary_PDF.pdf.

interested in multilateral discussions regarding the benefits and harms of AI.

This advisory body can accomplish a variety of goals. First, it would build governance capacity for the development and use of AI. Concretely, this technology's spread has made it imperative for national governments to improve their understanding of its capabilities. Yet not all states have the resources to effectively oversee the continuous stream of novel issues that could affect them. A new forum would facilitate the sharing of experiences and best practices for improving AI governance.

Second, this body can combat the lack of representation and inclusiveness in global AI discussions. To date, many existing fora dedicated to the discussion of technical and governance matters lack the input and perspective of countries not at the forefront of AI design, development, and deployment. These underrepresented parties are nonetheless unequivocally affected by the technology's proliferation. Moreover, there is a need for a neutral convening point where multilateral discussions on any governance issue related to AI can take place without the jurisdictional or thematic restrictions present in existing fora. It is our belief that all parties must have a seat available to participate in talks about how AI is affecting them and their relationships with other members states.

For example, the discussions on autonomous weapons systems within the UN CCW do not include the 67 member states not party to the CCW. Furthermore, the thematic limitations of this forum restrict discussion of key issues such as the non-military use of autonomous weapons, automated escalation, arms race dynamics, use of autonomous weapons by non-state actors, autonomy embedded in chemical, biological, radiological, and nuclear weapons, and overall geopolitical destabilization.

Lastly, this body must coordinate how countries harness AI in support of the UN's agenda on sustainable development. Each of the 17 goals could leverage the benefits of AI to improve the lives of billions of individuals. Having a body that proactively finds the most effective methods and applications to solve problems in education, health, poverty, among others, will complement national government efforts to improve the livelihood of people, the planet, and our biodiversity.

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 attempt to cope with these issues by responding alone, 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. Our reliance on this "market of ideas" and its incentives currently represents our main source of information on the benefits and weaknesses of different approaches to minimize AI-based risks.

Missing from this market is a credible body that can generate consensus-based assessments on fundamental AI challenges affecting society. Much like the Intergovernmental Panel on Climate Change, the UN could advance 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 our most important AI problems. More than 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. If this idea is successful in enhancing multilateral coordination, future efforts could endeavor to design an agency that emulates the mission of the International Atomic Energy Agency and allows the UN to serve in an action-oriented capacity with respect to AI Safety.

Shared benefit and prosperity

AI has the potential to help humanity achieve each of the UN's 17 sustainable development goals by 2030. In our opinion, this requires the active involvement of stakeholders in the AI lifecycle. We believe that, in addition to pursuing incentives like investing resources to increase market share or pushing the frontiers of knowledge, parties concerned should aim to empower individuals across

geographies and socio-economic characteristics through their technology.

Our first recommendation here centers on the products and services generated by stakeholders. In this regard, the UN should have a role in connecting technology companies to its sustainability agenda via the publication of concrete challenges. Secondly, there is a segment of for-profit firms willing to commit their excess profits for the common good. Along these lines, the UN should assess mechanisms such as the "Windfall Clause" to fund the attainment of its sustainable development goals.

Proactive participation in the UN's sustainability agenda

As society has thus far witnessed, there are clear advantages to harnessing AI in solving complex challenges. In parallel, attempts to achieve the sustainable development goals through technology must somehow circumnavigate the myriad of constraints faced by governments throughout the world. Deciphering the complex combination of levers and actions involved will require no small amount of ingenuity. There is thus, an overwhelming need to meld the efficiencies and useful outputs of AI with the agenda on sustainable development.

Considering the above, we ask that all entities with activities in the AI lifecycle proactively participate in the UN's 2030 agenda on sustainable development. There are diverse means to make this happen. Unilaterally, parties with a prosocial mission can choose to "adopt" a clear and specific goal and devote resources to its solution. Bold action in this form requires entities with strong commitments to sustainability.

An alternative that could benefit all parties is to instruct the UN to serve as a permanent coordinator to generate high-impact challenges open to any AI product or service providers. Having an expertly-assembled concrete list of problems can incentivize individuals and organizations to re-direct their efforts and contribute to achieving the sustainable development goals. For the UN, this exercise could diminish the information asymmetries that currently exist for small, medium, and large entities potentially interested in helping the organization, but lacking the time or dedication to understand its needs.

A version of the coordination initiative suggested above is already implemented by the International Telecommunication Union via its "AI for Good" initiative.[11] However, there is a need for two important changes. First, challenges should be created for all the sustainable development goals. Second, it is necessary to appoint civil society, academic, or private sector champions with the mandate of communicating and improving this initiative over time.

Distribution of economic benefits

If current trends continue, 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. In this recommendation, 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." This leads us to the conclusion that the UN should have a role in assessing these initiatives to fund its agenda on sustainable development directly.

There are several efforts dedicated to exploring alternatives for sharing the economic benefits of AI. One we would like to highlight is called the "Windfall Clause."[12] This initiative seeks to pre-commit AI firms to sharing the profits of advanced AI development for the common good in scenarios where these profits are extreme and unexpectedly large (qualifying them as "windfall" profits). The creation of a Windfall Trust would be a distribution mechanism for those profits, detailing how windfall resources are distributed for the "common good." For the UN, such an option could be assessed as

11 United Nations, *AI for Good*, AI FOR GOOD, https://aiforgood.itu.int/ (last visited Mar 10, 2023).

12 Cullen O'Keefe et al., *The windfall clause: Distributing the benefits of AI for the common good*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 327 (2020); Cullen O'Keefe et al., THE WINDFALL CLAUSE.

an alternative that distributes the benefits of technology to the developing world with the objective of solving some of the world's most pressing problems.

Loyalty

Our final suggestion for a key commitment, pledge, or action entails the inclusion of a principle that protects the interests of AI system end users. We recommend that the notion of "loyalty" be recognized as an integral element in national and multilateral discussions that examine the characteristics of trustworthy AI systems.

An AI agent can be considered loyal to another entity, such as a user, insofar as the agent successfully serves or adopts that entity's goals and interests.[13] In the context of trustworthy AI, incorporating this idea into systems crystallizes the clash of incentives present in their design, development, and deployment.

Unlike humans, AI systems are not intrinsically self-interested. An AI system that is loyal to an end user will have goals (or an objective function) that satisfy that individual's objectives. In contrast, a system might face a conflict of interest if it attempts to be loyal to both an individual and another party (e.g., the company that created it). In addition, a system may be disloyal by representing itself as loyal to an individual while prioritizing the interests of other parties instead. Therefore, "loyalty" can affect a system's "perceived trustworthiness" because if a user believes (correctly or not) that a system is serving an interest other than theirs, they are less likely to trust it.

However, loyalty need not be binary. Transparently demarcating degrees of loyalty is important even if only to avoid scenarios where blind allegiance to the objectives of one entity leads to non-physical harms, including the loss of user trust in the system. But "disloyalty" can also lead to the destruction of property or bring material harm to individuals, making the demarcation no less than crucial. Lastly, an AI system could be loyal to several parties or incorporate social maxims (e.g., against breaking international human rights law).

The application of loyalty in systems is therefore useful in a wide range of cases. For AI systems that complement professions with a binding legal commitment to end users, known as fiduciary responsibility, the incorporation of loyalty is imperative. In the medical field, this would mean that AI technologies assisting doctors would provide suggestions aligned with the best interest of patients, instead of hospitals or insurance companies. Analogies of this type can also be made in the fields of law or financial advice.

Non-fiduciary AI systems also benefit from the integration of loyalty. Consumer goods provide a spectrum of services where transparency into the incentives behind their decision-making would benefit end users. Take for instance AI-based virtual assistants. The provision of answers that are biased towards a third-party who pays the developer of the assistant for this advantage will skew the data received by users. This can influence users to favor entrenched and/or manipulative interests. In extreme cases, this can catalyze physical and non-physical harms.

Overall, the consideration of loyalty as a principle for trustworthy AI could decrease the information asymmetries confronted by users when interacting with these systems. With this recommendation, FLI believes that making AI's incentives transparent will improve society's interaction with increasingly complex products and services aimed at serving humanity.

13 ANTHONY AGUIRRE ET AL., *AI Loyalty by Design: A framework for governance of AI*, (2021), https://papers.ssrn.com/abstract=3930338 (last visited Sep 28, 2021); ANTHONY AGUIRRE ET AL., *AI loyalty: A New Paradigm for Aligning Stakeholder Interests*, (2020), https://papers.ssrn.com/abstract=3560653 (last visited Sep 22, 2021).