

The Global Digital Compact and Foundation Models

Submission by the Simon Institute for Longterm Governance

Summary

In the last 6 months, the Simon Institute for Longterm Governance held three workshops with diverse audiences on future-proofing the UN. Recommending a focus on AI governance to the Global Digital Compact was identified as important. After conducting an analysis of other submissions and consulting all key stakeholder groups, a focus on the governance of foundation models appeared important. These powerful AI systems cannot currently be reliably aligned with human values, unless increased investment in interpretability research makes these black-box algorithms more accessible.

Principles

- 1. The GDC should emphasize the convergence of problems and solutions in AI regulation, and thus advocate for the development of reliable interpretability tools.
- 2. The GDC should balance expertise with democratic input, engaging with private sector and academia while ensuring citizens have a say in designing the future.
- 3. The GDC should advocate for accountability of foundation models, as current regulation is insufficient to establish liability for harms caused.
- 4. The GDC should advocate risk-informed development for advanced AI at international, national, and local levels.

Recommendations

- 1. UN AI capacity-building program: Offer scholarships for experts from low- and middle-income countries to improve AI expertise at leading AI centres, and hire experts for forecasting AI capabilities, opportunities, and harms.
- 2. International consensus on AI auditing: Advocate for international coordination of AI auditing organizations through a multilateral forum, similar to the International Atomic Energy Agency's Regulatory Cooperation Forum.
- 3. Prioritize international cooperation on sharing AI incidents: Promote public sharing of incidents, adopt interoperable standards for reporting AI incidents, and support whistleblowing rights for those who work on advanced AI systems.
- 4. Interoperable standards for safe AI: Use standard-setting bodies to promote international collaboration on safety standards that respect member states' diverging values and regulatory frameworks while remaining interoperable. Private-sector innovation can also provide input.

Submission

Process of preparation

The Simon Institute for Longterm Governance (SI) hosted three workshops with academic experts, national delegates, members of the private sector, international civil servants, and civil society on 'future-proofing the UN' to support Our Common Agenda. During these workshops, a focus on AI governance for the Global Digital Compact was identified as an important step towards the safe, equitable adoption of advanced technologies. Subsequently, SI conducted an analysis of other organisations' inputs to the GDC submitted as of January 31st 2023. Based on its workshops and the existing DC submissions, SI drafted a maximally complementary submission.

Regulation of artificial intelligence

Al regulation is a vast topic, and the GDC has received many strong submissions. In order to constructively contribute to the conversation, the Simon Institute, in consultation with our partners, **has decided to focus our submission on the nature of foundation models** in advanced Al systems, such as ChatGPT.

These AI systems are not designed so much as progressively 'grown' via human feedback. They gain capabilities when developers scale the computing power invested, but we cannot predict which capabilities. These systems' nature makes them 'black boxes'; even the engineers who designed them do not know how foundation models come to their conclusions. The inability to robustly align these systems with human values is known as the 'alignment problem'.

Principles

- The GDC should reflect the above by emphasising a convergence of problems and solutions in AI regulation. Current foundation models present risks of bias, enhanced surveillance, misinformation, privacy violations, and increased inequality. If AI developers continue to scale up such models, experts worry that AI could undermine human control over critical infrastructure and manipulate the human psyche. All of these problems derive from a lack of accountability in foundation model development and our current lack of interpretability tools. Advocates for safe AI of all backgrounds can work together on common solutions (see next section).
- 2. The GDC should advocate to balance expertise with democratic input. The UN system lacks sufficient expertise on the technical drivers of AI risks and opportunities. It is essential to engage with the private sector and academia, where almost all technical AI expertise is currently concentrated. At the same time, all global citizens will be impacted by AI developments, and need to have a legitimate say in designing the future.

- 3. The GDC should advocate for the **accountability** of foundation models. For at least the last decade, frontier AI developments have occured in a handful of large corporations whose actions are difficult to regulate for any individual nation. This, combined with AI's current black-box nature, makes it difficult to establish liability for the harms they can cause, making accountability a key ingredient of regulation.
- 4. The GDC should advocate risk-informed development when governing advanced AI at the international, national and local levels. While this submission is focused on the potential harms of misaligned AI systems, aligned AI offers huge potential for human development. Technology can enable development, while development can mitigate some of the potential harms of advanced technologies. The UNDP's 'Risk-Informed Development' framework, alongside work from the University of Oxford on 'Differential Technological Development' can serve as a basis for the sustainable development of AI innovation.

Key actions

- A capacity-building program for Al expertise within the UN. The UN Al Experts Board could be supplemented with mixed democratic assemblies of global citizens. The program could offer scholarships to experts from low- and middle-income countries to travel to leading Al centres such as Beijing, San Francisco and London to improve Al expertise across the globe. The proposed UN Futures Lab could hire Al forecasting experts to analyze opportunities and harms, similar to work done by Epoch (https://epochai.org/).
- 2. Build international consensus on independent auditing of advanced AI systems. Third-party audits are important tools to minimize harms and maximize democratic accountability. Auditors reduce the potential for AI systems to misfunction in the real world. Companies are eager to prove their trustworthiness before deploying advanced systems, making audits a win-win. The AI auditing landscape is rapidly growing but remains concentrated in the United States and United Kingdom. The GDC should advocate for international coordination through a multilateral forum similar to the IAEA's Regulatory Cooperation Forum.
- 3. **Prioritise international cooperation on sharing AI incidents.** Transparency about incidents of AI harm are a useful way of establishing trust between companies, the public and member states. Public sharing of incidents, such as the EU's upcoming European AI database, allows for consensus-building on AI harms and failure modes. Member states and companies should adopt interoperable standards for the reporting of AI incidents. Special consideration would be given to low- and middle-income countries, where incidents often go underreported. Human rights bodies could also take up the cause of whistleblowing rights for those who work on advanced AI systems.

4. Promotion of interoperable standards for safe AI. Standard-setting bodies, such as ISO and IEC, can help make safer AI systems by promoting international collaboration on the adoption of safety standards. Standards need not be uniform, and can respect member states' diverging values and regulatory frameworks while remaining interoperable-thus avoiding a race to the bottom. UN bodies such as the Tech Envoy's Office can use their convening and discursive power to promote multilateral standards cooperation. Standards development can also take input from private-sector innovation, such as Anthropic's work on 'Constitutional AI': a set of hard-coded rules to ensure safer AI systems.

Content overview

Principles	Suggested actions	What SI has done
Convergence of problems and solutions: diverse Al-related risks (e.g. bias or self-preservation) have similar origins and solutions.	 Build international consensus on third-party audits of foundation models Cooperate internationally through a multilateral auditing Brought together AI safety experts from around the globe in Geneva for a shared brainstorm on short-term and long-term AI-related risks and their overl 	 Brought together Al safety experts from around the globe in Geneva for a shared brainstorm on short-term and long-term Al-related risks and their overlap
Emphasise accountability of foundation models.	 forum similar to IAEA's Regulatory Cooperation Forum Prioritise international sharing of Al incidents, with common standards for reporting Human rights bodies should advocate for whistleblowing rights for advanced Al workers 	 Held discussions on the human rights/Al intersection with human rights NGOs in Geneva and an OCHA staff member Held a discussion on Al incident sharing with a Swiss think tank as part of a policy diffusion workshop
Risk-informed development: conceive of AI-related risks according to their interplay with development goals.	 International cooperation on safe Al standards via SSBs 	 Interviewed members of SSBs and their technical commissions Held a workshop with UNDP on risk-informed development and

		rapid technological change
Balance expertise & democratic input.	 UN AI Ethics Board, supplemented with democratic assemblies Offer scholarships for LMIC experts to improve AI knowledge transfer 	 Held a consultation with a leading NGO in the previous UN Digital Roadmap as part of a wider academic/NGO stakeholder engagement workshop
	 UN Futures Lab hires AI forecasting experts Incorporate private-sector expertise in development of safe AI standards 	 Consulted with the Futures Lab on institutional design

Conclusion

We congratulate the Office of the Secretary-General's Envoy on Technology for developing a thoughtful, inclusive consultation process. We recognize the difficulty of handling many issues simultaneously and hope to have demonstrated the convergence of current and emerging threats from modern technology. We wish the co-facilitators the best of skill for navigating the process ahead.