UN High-level Panel on Digital Cooperation:

A Proposal for International AI Governance

Authors: *Dr Luke Kemp,[1] Peter Cihon,[2] Matthijs Michiel Maas,[3] Haydn Belfield,[4] Dr Seán Ó hÉigeartaigh,[5] Jade Leung[6] and Zoe Cremer.[7]*

Summary

International Digital Cooperation must be underpinned by the effective international governance of artificial intelligence (AI). AI systems pose numerous transboundary policy problems in both the short- and the long-term. The international governance of AI should be anchored to a regime under the UN which is inclusive (of multiple stakeholders), anticipatory (of fast-progressing AI technologies and impacts), responsive (to the rapidly evolving technology and its uses) and reflexive (critically reviews and updates its policy principles). We propose some options for the international governance of AI which could help coordinate existing international law on AI, forecast future developments, risks and opportunities, and fill critical gaps in international governance.

1. Issues in Digital Cooperation

Digital cooperation will rise or fall by the use or misuse of rapidly developing artificial intelligence (AI) technologies. AI will transform international social, economic, and legal relations in ways that spill over far beyond the digital realm. Digital cooperation on AI is essential to help stakeholders build capacity for the ongoing digital transformation and to support a safe and inclusive digital future. Accordingly this submission will focus on the international governance of AI systems.

AI technologies are dual-use. They present opportunities for advancements in transport, medicine, the transition to renewable energy and lifting standards of living. Some systems may even be used to strengthen the monitoring and enforcement of international law and improve governance. Yet they also have the potential to create significant harms. These include labour displacement, unpredictable weapons systems, strengthened totalitarianism and destabilizing strategic shifts in the international order (Dafoe 2018; Payne 2018). The challenges of AI stem from both capabilities that already exist, or will be reached in the near-term (within 5 years), as well as from longer-term prospective capabilities. The two are intricately intertwined. How we address the near-term challenges of AI will shape longer-term policy and technology pathways (Cave and ÓhÉigeartaigh 2019). Yet the long-term disruptive impacts could dwarf other concerns. Both need to be governed in tandem.

Challenges from Existing and Near-Term Capabilities

- Maintaining effective human oversight in application of AI to military technology, decision support and infrastructure;
- Algorithmic bias and justice;
- Algorithmic transparency;
- AI-aided cybercrime;
- AI-aided cyberwarfare;
- Safety and regulation of autonomous vehicles;

[1] Research Associate, The Centre for the Study of Existential Risk, Cambridge University.
[2] Research Affiliate, Center for the Governance of AI, Future of Humanity Institute, Oxford University.
[3] PhD Fellow, Centre for International Law, Conflict and Crisis, University of Copenhagen.
[4] Academic Project Manager, The Centre for the Study of Existential Risk, Cambridge University.
[5] Executive Director, The Centre for the Study of Existential Risk, Cambridge University.
[6] Head of Research & Partnerships, Center for the Governance of AI, Future of Humanity Institute, Oxford University.
[7] Research Affiliate, The Centre for the Study of Existential Risk, Cambridge University.

- Privacy and surveillance; and,
- AI-enabled computational propaganda.

Challenges from Long-Term Capabilities

- Wide-spread labour displacement could heighten wealth inequalities, and fuel domestic and international political instability;
- Advances in the application of AI to military technology could overturn tactical or strategic force balances or lead to ambiguity over relative power, increasing the chance of strategic miscalculation and international conflict;
- The creation of high-level machine intelligence (HLMI). That is, an unaided AI system that performs as well as an average human across most cognitive skill tests and economically relevant tasks.If such an HLMI is not value-aligned with wider society it could cause catastrophic damage either by accident or strategic misuse.

While most of these challenges have not received sufficient attention, several have been mapped in *The Malicious Use of Artificial Intelligence* report (Brundage & Avin et al 2018), *AI Governance: a Research Agenda* (Dafoe, 2018), and in the Future of Life's (2019) 14 policy challenges. Greater attention is needed to forecasting these potential challenges. Both the foresight of policy problems and the magnitude of existing issues underline the need for international AI governance.

2. What Values and Principles Should Underpin Cooperation?

There are already over a dozen sets of principles on AI composed by governments, researchers, standard-setting bodies and technology corporations (cf. Zeng et al. 2019). Most of these coalesce around key principles of ensuring that AI is used for the common good, does not cause harm or impinge on human rights, and respects values such as fairness, privacy, and autonomy (Whittlestone et al. 2019). We suggest that the High-level Panel on Digital Cooperation compile and categorise these principles in its synthesis report. Importantly, we need to examine trade-offs and tensions between the principles to refine rules for how they can work in practice. This can inform future negotiations on codifying AI principles.

The international governance of AI should also draw from legal precedents under the UN. In addition to general principles of international law, principles such as the *polluter pays principle* (those who create externalities should pay for the damages and management of externalities) could be retrofitted from the realm of environmental protection to AI policy. Values from bioethics, such as autonomy, beneficence (use for the common good), non-maleficence (ensuring AI systems do not cause harm or violate human rights), and justice are also applicable to AI (Beauchamp and Childress 2001; Taddeo & Floridi 2018). Governance should also be responsive of existing instruments of international law, and cognizant of recent regulatory steps by international regulators on the broader range of global security challenges created by AI (Kunz & Ó hÉigeartaigh 2019). Finally, while some specialization of AI governance regimes for distinct domains is unavoidable, steps should be taken to ensure these distinct standards or regimes reinforce rather than clash with each other.

3. Improving Cooperation on AI: Options for Global Governance

International governance of AI should be centred around a dedicated, legitimate and well-resourced regime. This could take numerous forms, including a UN specialised agency (such as the World Health Organisation), a Related Organisation to the UN (such as the World Trade Organisation) or a subsidiary body to the UN General Assembly (such as the UN Environment Programme). Any regime on AI should fulfil the following four objectives:

- **Coordination:** To coordinate and catalyse AI-related efforts under existing international treaties and organisations (both specialized agencies and subsidiary bodies);
- **Comprehensive Coverage:** To fill extant gaps in international governance, such as the use of AI-enabled surveillance technologies, cyberwarfare and the use of AI in decision-making;

- **Cooperation over Competition:** To encourage international cooperation and collaboration between AI groups on projects for the public good;
- **Collective Benefit:** To ensure benevolent, responsible development of AI technologies and the equitable distribution of benefits.

The Panel should consider the following options as components for an international regime:

- **A Coordinator and Catalyser of International AI Law:** there is already a tapestry of international regulations on AI being developed, including through the International Maritime Organisation (IMO), the Vienna Convention on Road Traffic, and the Council of Europe (such as the Budapest Cybercrime Convention and the Automatic Processing Convention). However, many of these initiatives are fragmented in membership and functions. We welcome the recent efforts of UN System Chief Executives Board for Coordination through the High-Level Committee on Programmes to draft a system-wide AI engagement strategy. This should be strengthened. Moreover, other avenues could be considered. For example, the creation of a coordinator for existing efforts to govern AI and catalyse multilateral treaties and arrangements for neglected issues. This would follow the precedent of the United Nations Environment Programme (UNEP) in synchronizing international agreements on the environment and facilitating new ones such as the 1985 Vienna Convention for the Protection of the Ozone Layer. New institutions could be also brought together under an umbrella body, as the 1994 World Trade Organisation (WTO) has done for trade agreements.
- **An Intergovernmental Panel on AI (IPAI):** There is a dire need for measuring and forecasting the progress and impacts of AI systems. This could include examining the future capabilities of AI across a range of cognitive domain and economic tasks, stocktaking how algorithms are used in decision-making, analysing emerging techniques and technologies and exploring potential future impacts, such as on employment. An IPAI could provide a legitimate, authoritative voice on the state and trends of AI technologies. We welcome the joint Canadian and French International Panel on AI. However, how it draws on expertise and accesses information needs careful design. If it proves successful it should eventually be expanded to become truly intergovernmental and encompass missing issues such as weapons control and AI. The IPAI could inform international governance and perform assessments every three years as well as quick response special issue assessments.
- **A UN AI Research Organisation (UNAIRO):** This organisation would operate from a pool of government funding. This UNAIRO could focus on building AI technologies in the public interest, including to help meet international targets such as the 2015 Sustainable Development Goals (SDGs) as called for by the 2018 UN Secretary-General's Strategy on New Technologies (Guterres, 2018). A secondary goal could be to conduct basic research on improving AI techniques in the safest, careful and responsible environment possible. The goal would be to channel AI talent towards cooperation in creating technologies for global benefit.

The outlined options for a regime should be anticipatory, reflexive, responsive and inclusive. This adheres to the key tenets of Responsible Research and Innovation suggested by scholars (Stilgoe et al 2013). To be inclusive we suggest following the ILO's innovative model of multipartite representation and voting. In this case voting rights could be distributed to nation states as well as other critical stakeholder group representatives. An ability to anticipate emerging challenges and respond to the quickly evolving technological landscape would be enabled by the IPAI. Responsiveness could be built into the body by having principles on AI reviewed and updated every three years. This would ensure that policies reflect the latest and in-country experiences.

With prudent action and foresight, the UN can help ensure that AI technologies are developed cooperatively for the global good.

References

Beauchamp, T. and Childress, J. (2001). Principles of biomedical ethics. *Oxford University Press, USA*.

Brundage, M et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *Future of Humanity Institute and the Centre for the Study of Existential Risk*.

Cave, S. and ÓhÉigeartaigh, S. (2019). An AI Race for Strategic Advantage: Rhetoric and Risks. In AAAI/ACM Conference on Artificial Intelligence, Ethics and Society.

Cave, S. and ÓhÉigeartaig, S. (2019). Bridging near- and long-term concerns about AI. *Nature Machine Intelligence*, 1: 5-6

Dafoe, A. (2018). AI Governance: A Research Agenda. *Future of Humanity Institute, Oxford University*.

Guterres, António. "UN Secretary-General's Strategy on New Technologies." *United Nations*, September 2018. http://www.un.org/en/newtechnologies/images/pdf/SGs-Strategy-on-New-Technologies.pdf.

Kunz, Martina, and Seán Ó hÉigeartaigh. "Artificial Intelligence and Robotization." In *Oxford Handbook on the International Law of Global Security*, edited by Robin Geiss and Nils Melzer. Oxford University Press, 2019. https://papers.ssrn.com/abstract=3310421.

Payne, K. (2018). Artificial Intelligence: A Revolution in Strategic Affairs? *IISS*.

Stilgoe, J., Owen, R. and Macnaghten, P. (2013). Developing a Framework for Responsible Innovation. *Research Policy*, 42(9): 1568-1580

Taddeo, Mariarosaria, and Luciano Floridi. "How AI Can Be a Force for Good." Science 361, no. 6404 (August 24, 2018): 751–52. https://doi.org/10.1126/science.aat5991.

Whittlestone, J., Nyrup, R., Alexandrova, A. and Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society. AAAI and ACM Digital Libraries.*

Zeng, Yi, Enmeng Lu, and Cunqing Huangfu. "Linking Artificial Intelligence Principles." *Proceedings of the AAAI Workshop on Artificial Intelligence Safety (AAAI-Safe AI 2019)*, 2019. http://arxiv.org/abs/1812.04814.