The first international treaty on AI governance – a basis for convergence or dissention?

The transforming technologies contained under the AI label have prompted a flurry of action by governments but international cooperation on AI governance is likely to remain elusive.

By Paul Meyer

With little fanfare this September in Vilnius, Lithuania the first international legally binding agreement on the governance of Artificial Intelligence (AI) was opened for signature. Negotiated by the 46 members of the Council of Europe and 11 observer states, the treaty is entitled "Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law". As of this month the treaty has already been signed by 10 states including the United States and the United Kingdom as well as the European Union. Canada was engaged in the negotiation (and even conducted a public consultation to inform its approach) but has not signed the convention to date, although there is no reason to suspect it will not.

The transforming technologies contained under the AI label have prompted a flurry of action by governments to generate norms to rule its development and use. These have tended to take the form of various non-binding Codes of Conduct or sets of principles without a common legal basis or inclusive scope. The report of the UN's High-level Advisory Body on AI has described the normative situation in this manner: "There is today a global governance deficit with respect to AI. Despite much discussion of ethics and principles, the patchwork of norms and institutions is still nascent and full of gaps. There is no shortage of documents and dialogues focused on AI governance. Hundreds of guides, frameworks and principles have been adopted by

governments, companies and consortiums, and regional and international organizations." In short, a lot of text, but little coherence.

The inequities in adherence to the existing normative statements is also striking. The UN report noted that only seven states had adopted all of the existing documents on AI governance, whereas 118 states had adhered to none of them. It will not come as a surprise that the majority of adherents are from developed Western states and the Global South is largely absent.

In theory the Council of Europe Convention is intended to bridge this gap by providing a legally binding treaty that is open to all. However, the laudable initiative to develop an AI treaty may confront the same problem encountered with the 2001 Council of Europe Convention on Countering Cybercrime which also was the first international treaty dealing with this subject. While this treaty currently has 76 states parties, its adherence in Asia and Africa is spotty. Notably, states such as China, Russia, Egypt, India, Indonesia have not signed on.

As a general rule, states wish to be at the negotiating table for treaties they adhere to. Eventually, the unequal status of states in the elaboration of the Council of Europe Cyber Crime treaty led to a decision to re-launch negotiations under UN auspices. This treaty was adopted this August after a multi-year negotiating process. The fact that the treaty was elaborated at the UN imparts a legitimacy and inclusiveness to the result that the product of a regional organization cannot match. Therefore, there is a risk that the current Council of Europe convention may encounter a similar problem in gaining the adherence of states beyond the largely European and North American states engaged in its negotiation.

This concern aside, what have the negotiators produced in the Convention and how effective will it be in governing AI? The principal objective of the Convention as set out in Article 1 is "to ensure that activities within the lifecycle of artificial intelligence systems are fully consistent with human rights, democracy and the rule of law". Each party is enjoined to "adopt or maintain appropriate legislative, administrative or other measures to give effect to the provisions set out in this Convention".

In other words, the states party are responsible to implement the Convention as they see fit. The scope of the Convention is also limited to AI activities "undertaken by public authorities, or private actors acting on their

behalf". The activities of the private sector are subject to a looser form of state oversight, basically only requiring the state to address "risks and impacts" from activities by private actors in light of the purpose of the Convention. Given the dominance of the private sector in the field of AI development and management, this light oversight stipulated by the Convention may be all that the negotiators could achieve while still ensuring support by the participating states, but could disappoint those looking for a more constraining regime.

Another feature of the Convention is its exclusion of anything related to security. Article 3 stipulates that the Convention does not apply to any activities "related to the protection of its national security interests" a particularly elastic formula as a state might determine almost anything as relevant to its "national security interests" (e.g. tariffs on imports of steel and aluminum). As if this exclusion was not sufficient, Article 3 also excludes "matters relating to national defence". Those stakeholders with concerns over military applications of AI will have to look to another instrument to address this matter.

On the goals of protecting human rights, democracy and the rule of law (all of which are enumerated in the Convention's title) the treaty provides the expected assurance that AI activities will need to conform to existing obligations in international and domestic law. However, operationalizing this area will require more than broad affirmations. For example, Article 5 stipulates that AI systems "are not [to be] used to undermine the integrity, independence and effectiveness of democratic institutions and processes…".

How this will apply to specific activities will need elaboration by state authorities and may result in major divergences in practice. In the context of an election for instance, one state may consider AI-enabled attack ads problematic whereas another would view them as a legitimate campaign tool.

Beyond support for these top order goals states are required to "ensure that adequate transparency and oversight requirements tailored to the specific context and risks are in place…". Similarly, states are to put in place "measures to ensure accountability and responsibility for adverse impacts on human rights, democracy and the rule of law" and offer "accessible and effective remedies for violations". Again, these measures will be left up to

the state to determine, but the Convention does enshrine expectations regarding state performance to this end.

The institutional support for the Convention is also relatively light. Instead of a dedicated implementing organization, a Conference of the Parties is envisioned assisted by the Secretariat of the Council of Europe. Parties, the Convention notes, "shall consult periodically" with a view to inter alia "facilitating, where necessary, the friendly settlement of disputes related to the application of this Convention.

" No fixed periodicity for meetings of the Conference of Parties is specified with convening authority left to the Secretary General of the Council of Europe "whenever necessary" or when a majority of the Parties requests its convocation. Incentivization for compliance is further restricted to a reporting obligation "within the first two years after becoming a Party and then periodically thereafter with details of the activities undertaken to give effect" to the chief obligations of Article 3.

In the absence of a common template, much will be left to the state as to the contents of such reports.
The entry into force requirements for the Convention also set a relatively low threshold of ratifications (five signatories including at least three member states of the Council of Europe). This should ensure entry into force for the Convention in the early new year. The denunciation of the Convention is also provided for upon three months notification.

Overall this is a relatively light institutional overlay that does not require establishing new forums, relying on a virtual Conference of the Parties and the existing Secretariat of the Council of Europe. By way of comparison, the UN report mentioned earlier had called for the establishment at the UN of an "inclusive policy forum" and suggested that it could serve as a "clearing house for AI standards that would apply globally". To support this envisaged forum the report proposed setting up an Office of AI affairs located in the Secretariat and reporting to the Secretary General.

The arrival of the first legally binding international agreement for AI governance is undoubtedly a significant development. It remains to be seen, however, whether the new treaty will serve as a rallying point for states seeking a governance framework or prove (as was the case with the Council of Europe's Cyber Crime Convention) a divisive factor within the

international community. States from the Global South that were not represented at the negotiating table for the Convention or those who might object to its focus on the human rights and democratic dimension of AI activity will not be easily persuaded to come on board. This will not affect Canada's likely adherence in the near term but signals that international cooperation on AI governance may remain limited to a subset of like-minded states for the foreseeable future.

This article has been first published in Open Canada.

Geneva, 20 December 2024.