UNCLASSIFIED

MEMORANDUM FOR ACTION

TO:	**The Minister of Foreign Affairs**
CC:	**The Digital Inclusion Lab, Office of Human Rights Freedoms and Inclusion**
SUBJECT:	Governance of Artificial Intelligence: Preparing for When AI Goes Awry

SUMMARY:

This memorandum presents options to position Canada's AI foreign policy stance as an ethical, responsible and proactive player.

RECOMMENDATION(S):

The following recommendations should be integrated into *The Montreal Declaration on the Responsible Development of AI* in support of the seven guiding principles of the declaration (well-being, autonomy, justice, privacy, knowledge, democracy, and accountability) if possible. Additionally, the example set by the European Union on preventative and proactive AI policy should be rigorously implemented into Canada's foreign policy on AI.

Recommendations:

- Create policies to ensure development agencies abide by the five-point platform for AI development based on Concrete Problems in AI Safety (2016). See Annex 1.
- Establish a multi-stakeholder advisory board comprised of industry, education, research and government personnel to update policy according to advancements in AI and to ensure that domestic industries abide by these policies. See Annex 2.
- Invest in education and training on ethics in the field of AI. Institute a mandated training course for developers and associated stakeholders in the development and implementation of AI. See Annex 3.
- Develop codes of conduct for AI engineers and others associated with the development of AI. See Annex 4.
- Incorporate, through the means of a protocol, AI-related policies in international treaties and agreements, such as the Wassenaar Arrangement (to deal with the dual-use nature of AI) and the Arms Trade Treaty (ATT, to deal with the human rights implications of AI). Potentially use these two treaties as the basis of future AI agreements. See Annex 5.

KRISTEN MYERS, RUXANDRA BADEA, SULAMITA ROMANCHIK

GLOBAL GOVERNANCE, BALSILLIE SCHOOL OF INTERNATIONAL AFFAIRS, UNIVERSITY OF WATERLOO

£ I wish to discuss £ I concur £ I do not concur

Minister

Background:

1. The field of Artificial Intelligence (AI) has progressed rapidly, both domestically and internationally. We are observing the transformations of current operations in many employment, development and governance sectors which demonstrate a need for greater understanding and preparedness in terms of these advancements. Concerns surrounding AI "accidents" have grown as these systems become more powerful and independent. Accidents are a situation where a human designer had in mind a certain objective or task, but the system that was designed and deployed for that task produced harmful and unexpected results (Annex 1).

2. Canada is a world leader in AI innovation. Accordingly, Canada should take a lead in promoting ethical AI. Recommendation 1 creates a baseline of ethical AI development and foresight to inform the advisory committee in recommendation 2, educational system (recommendation 3), and international treaties (recommendation 4). This policy brief is intended to guide Canada's proactive approach to ethical AI.

CONSIDERATIONS:

3. In November 2017, a preamble of a draft declaration on socially responsible AI was unveiled, titled *The Montreal Declaration on the Responsible Development of AI*. The declaration came to fruition through a co-construction process, wherein individuals from all fields were involved. The final draft of the declaration was to be ready sometime at the end of January or early February of 2018 but remains open to change and is revisable and amendable. While the Montreal Declaration can be seen as the first step towards a governance mechanism on AI, it would be more feasible to consider adding to existing frameworks rather than creating several new governance mechanisms. Having multiple governing mechanisms, agreements, frameworks, etc. does not strengthen sound governance but simply mitigates the seriousness of the issue at hand and makes compliance more difficult, as it is hard to keep up with the large number of governance outputs.

Investing in education and training on ethics will require a restructuring of the educational program within universities and colleges, as well as looking into certification programs or skills upgrading courses for all individuals already working in AI careers. Establishing sound ethics training will require the federal government to collaborate with the provincial governments across Canada and the Ministries of Education and to either increase, or reallocate, educational funds. It will also require Global Affairs Canada to collaborate with other departments within the Federal Government to ensure that all Public Service Workers are just as informed and trained on AI; its governors cannot be unaware of the implications of the emergent technology.

Government commitments towards sound AI and robotics governance have already been instituted by several countries--including Estonia, the US, Japan, China, South Korea, and the EU--and Canada cannot afford to lag behind. Examining how other governments approach AI governance will enable Canada to create its own governance mechanism that draws on the

strengths of existing mechanisms and fills any gaps that other states have missed. This will set Canada apart as the leader in AI governance that it strives to be.

Lastly, the Canadian government may face opposition from corporations and AI developers, on the basis that regulation of AI will stifle innovation. However, allowing innovation that violates national, and especially international, laws is futile and may result in action being taken against Canadian government and its companies. Involving developers in the creation of a regulatory process, and especially on the Advisory Board, is the ideal way to balance innovation and regulation.

Following the proposition of Moore's law, which states that technology experiences exponential growth rates and has the ability to outpace government policies, immediate implementation of the recommendations is necessary. A rapidly increasing number of AI applications are already being tested despite a lack of its governance mechanism, which is leading to serious implications for all stakeholders. As AI products enter the market as tradable goods, Global Affairs Canada must ensure that the products are not being exported to countries that do not support human rights or that do not align with Canada's feminist foreign policy. Thus, an AI governance mechanism is not for the near or far future but for today.

Communications Implications/Actions:

5. Media scrutiny surrounding the growing use of, investment in, and governance of AI and the potentially harmful effects it may pose is expected to continue. Furthermore, increased initiative and proactive leadership towards governing AI and preventing its potentially harmful effects will likely spark a wider public debate. As already mentioned before, private enterprises and corporations will likely raise concerns over the proposed governance mechanisms stifling innovation, while the wider public will raise concerns over the ethical use and innovation of AI, as well as its potential effects on job loss. An informed and comprehensive media output of the proposed strategy would reassure all stakeholders involved. Transparency will serve to give the public and stakeholders involved peace of mind in terms of what is being done to govern AI and the potential harm it may pose. It is pertinent that Canada takes a proactive stance towards implementing measures and mechanisms to govern AI while providing the public with reliable and accurate information in order to ease discontent. A strategic comprehensive communications strategy consisting of multi-department and multi-source release of information is the best way to achieve this goal.

6. Concerns about AI development underscore important links to pre-existing governance domains. This includes the government's procurement and job creation strategies, including the Innovation Superclusters Initiative, centered in the government's Innovation and Skills Plan, and the Innovative Solutions Canada program, part of the Innovative Skills Program. This also includes the government's Pan-Canadian Artificial Intelligence Strategy which includes the priority to develop global thought leadership on the economic, ethical, policy and legal implications of advances in AI. These projects should once again be highlighted as accompaniments to the wider AI media communication to demonstrate proactive actions taken.

Annex:

1. Related to Recommendation 1: The five-point platform from “Concrete Problems in AI Safety (2016).

- **Avoiding Negative Side Effects:** AI shouldn't disturb its environment while completing set tasks
- **Avoiding Reward Hacking:** AI should complete tasks properly, rather than using workarounds (like a cleaning robot that covers dirt with material it doesn't recognise as dirt)
- **Scalable Oversight:** AI shouldn't need constant feedback or input to be effective
- **Safe Exploration:** AI shouldn't damage itself or its environment while learning
- **Robustness to Distributional Shift:** AI should be able to recognise new environment and still perform effectively in them

Source: Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

2. Related to Recommendation 2: Related to the establishment of an Advisory Board

Canada’s neighbour, the United States (US), has introduced two identical, new bills in December of 2017 on ethical AI: bill H.R. 4625 in the House and bill S. 2217 in the Senate. One role of the bills is to establish a Federal Advisory Committee to provide guidance with the development and implementation of AI, including unbiased development of such technologies, accountability and legal rights, including matters relating to the responsibility for any violations of laws by an AI system and the compatibility of international regulations, and matters relating to open sharing of data and the open sharing of research on AI.

Canada could benefit from the creation of bills on AI governance which would be made possible and informed through a similar Advisory Board. The proposed Advisory Board would be comprised of all stakeholders, Global Affairs Canada included, to promote open dialogue and advice among partners. The Board would serve as the base of an AI governance mechanism for Canada domestically and abroad.

3. Related to Recommendation 3: Ethics training and education

The aforementioned bills that have been introduced in the U.S., delegate the Advisory Committee with providing advice on the following tasks among others already mentioned:

> (A) Education, including matters relating to science, technology, engineering, and mathematics education to prepare the United States workforce as the needs of employer’s change.
>
> (B) Ethics training and development for technologists working on artificial intelligence.

A plan for the inclusion of ethics in development and implementation of AI already exists in the U.S. and if Canada is to become a leader in ethical AI then the Canadian government must

likewise ensure that every individual implicated with AI technologies has gone through similar ethical training. This includes political figures and diplomats that may be involved in future treaty negotiations concerning AI technologies, as Canadian foreign policy is a reflection of the individuals who formulate it.

4. Related to Recommendation 4: Codes of Conduct and Licences:

Following in the steps of the European Union, who has so far accomplished the most comprehensive work on regulating AI and robotics of the future, Canada should establish a Code of Ethical Conduct for Robotics Engineers and for others implicated in AI (researchers, corporations, etc.). This task could be delegated to the proposed Advisory Board, whose multi-stakeholder constituency would ensure a wholesome code is presented. Likewise, a code of ethical conduct should be established for the Advisory Board itself, to ensure that there are no conflicts of interest between the developers and licensees. In order for conflicts of interest to be avoided, processes must be done through as much transparency as possible.

The following code of ethical conduct has been proposed by the EU and could serve as a guide for Canada's own code:

Beneficence- robots should act in the best interests of humans

Non-maleficence- the doctrine of 'first do not harm' whereby robots should not harm a human

Autonomy- the capacity to make an informed un-coerced decision about the terms of interaction with robots

Justice- fair distribution of the benefits associated with robotics and affordability of homecare and healthcare robots in particular

*please note: EU's ethical code of conduct is voluntary; however, Canada should take a more ambitious approach as a leader and propose a mandatory code of conduct.

5. Related to Recommendation 5: International Treaties

As concern about autonomous weapons rises, incorporating AI provisions into the Arms Trade Treaty (ATT) through a Protocol is a necessary step forward. The ATT explicitly outlines national export and import controls of countries in alignment with Human Rights and international humanitarian law considerations (ex. Geneva Conventions of 1949, and Additional Protocol I of 1977). As such, countries with potential to use weaponry for the violation of international human rights law, are prohibited from arms trade. Being that AI has the potential to inflict similar Human Rights violations as other weaponry, it cannot be excluded from existing international law but must be regulated in like manner to avoid it ending up in the hands of those who would exploit it (ex. Human rights abusing regimes, terrorist groups, potentially non-democratic countries, etc.).

Additionally, the ATT addresses Risk of Diversion, particularly regarding Gender-based violence. This aligns well with Canada's feminist foreign policy, as the treaty prohibits weapons

that commit or facilitate this form of violence. For prospective future international agreements, the ATT provides a solid reference point as a framework to address exports and imports of AI technologies between states.

Furthermore, the incorporation of AI provisions in the Wassenaar Arrangement through a protocol is also an important and crucial step forward. The Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies was established in order to contribute to regional and international security and stability, by promoting transparency and greater responsibility in transfers of conventional arms and dual-use goods and technologies, thus preventing destabilising accumulations. Integrating AI provisions into the arrangement would ensure that these transfers are not diverted to support purposes which could threaten regional and international peace and security, while further safeguarding the public by preventing potential human rights violations.