Global Affairs Canada | Affaires mondiales Canada

UNCLASSIFIED

MEMORANDUM FOR ACTION

TO:	**The Minister of Foreign Affairs**
CC:	**The Digital Inclusion Lab, Office of Human Rights Freedoms and Inclusion**
SUBJECT:	A Strategy to Promote Transparency, Explainability, and Accountability in the Global Development and Use of AI

SUMMARY:

The rapid development of AI technologies is creating a global governance problem for policymakers. Governments around the world want to promote AI for its many benefits but are quickly recognizing the problems that it is creating today from a human rights perspective. Making matters worse is AI's black box problem, which makes it incredibly challenging to hold people or companies accountable when harm occurs.

This memorandum presents a strategy for the Government of Canada to leverage the current global momentum surrounding ethical AI. It recommends a soft law approach that will position Canada as the global thought leader on the social and ethical implications of AI.

RECOMMENDATION(S):

- That you leverage Canada's Presidency of the G7 to create an internationally agreed-upon list of principles and standards with regards to transparency and accountability in the development and use of AI.
- That you create a new global agency to monitor, evaluate and, coordinate global efforts to regulate AI.

Students: Tim Dutton, Vanessa Ko, Gabrielle Lim, Shamim Shabani

Advisor: Todd Foglesong, Munk School of Global Affairs

☐ I wish to discuss
☐ I concur ☐ I do not concur

Minister

Canada

BACKGROUND:

1. AI is developing rapidly, but regulations to govern it have not emerged in tandem with its rate of growth and adoption. This governance gap is particularly concerning when examined from a human rights perspective, since the growing prevalence of AI technologies has already exacerbated societal biases, manipulated customers and employees, and produced harms related to justice, equality, safety, and discrimination. Furthermore, due to AI's inherent "black box" problem, private companies and public bodies who use AI technologies to make decisions have found it difficult to explain why an AI made the decision it did, creating an *explainability* problem (see Appendix 1: Definitions). As a result, there is currently popular consensus that encouraging and regulating transparency and explainability within the development and use of AI would allow for better accountability and mitigate against potential harmful effects on society.

2. Few government agencies are addressing AI's global governance, especially as it relates to transparency and explainability of AI. Instead, there has been a proliferation of non-governmental actors who, collectively, govern AI in a piecemeal approach—academic consortiums, private sector companies, and non-profit organizations each tackle issues that relate to them in isolation. Due to this proliferation, a number of overlapping but separate principles to govern behaviour have emerged (see Appendix 2: Comparison of Principles). This presents an opportunity to Canada: the government should leverage this momentum and establish the first ever governmental agreement on transparency and accountability in AI.

CONSIDERATIONS:

3. Overall, most of the current governance methods are oriented towards future outcomes, such as the future of work and superintelligence. However, harms that arise from the use of AI, such a breach of privacy or discrimination against a marginalized community, are largely a result of issues during the development process of AI (see Appendix 3: Processes vs Outcomes). Each step in the development process can produce potentially harmful effects for end-users – AI applications that contain bias and incomplete data, for example, can reinforce sexist decision-making. Unfortunately, due to the lack of transparency and explainability in the process, stakeholders and end-users often discover the effects at the post-launch phase. At that point, it becomes complicated to effectively retrace what part of the process was at fault and how to correct it. Additionally, given the current proliferation of stakeholders, there is a lack of a coordinating body to influence and regulate explainability in AI. It is in addressing this governance gap, that Canada can have greater influence in AI governance.

4. There are multiple policy options for Canada to explore in the effort of addressing the current governance gap around explainability. First, the government could consider a *'hard law approach,'* which consists of international treaties and niche bans on specific AI technologies and products. For instance, countries could create a treaty to ban the use of black box AI in government agencies. Second, governments could consider a '*soft law approach,*' which would involve establishing a series of standards, audits, or publicly published impact assessments. Lastly, governments could take a *'self-regulation*' approach, whereby developers congregate to set a series of guidelines with minimal state-intervention. Self-regulation is the status quo and an example of this is the Partnership on AI, which represents Amazon, Apple, Facebook, Google, and Microsoft. In short, since the status quo has led to the explainability problem and a hard law strategy takes too long to implement, this memo recommends taking a soft law approach to AI governance (for full analysis, see Appendix 4: Analysis of Policy Options).

5. Canada should consider undertaking a soft law approach to governance, whereby the emphasis is on standard setting and evaluating potential harms through impact assessments. For example, by encouraging developers to publicly publish an impact assessment, multiple stakeholders can audit the development process for any unconscious bias or underlying, implicit discrimination. A soft law approach also enables a democratic, multi-stakeholder, participatory governance of AI that would directly tackle explainability: civil society actors, government agents, and regulators would all demand AI developers and private companies to follow the agreed upon global standards. Furthermore, with such a participatory system, developers can receive critical feedback to mitigate potential harms prior to launch. This soft law approach has been proven effective through other standard setting bodies such as, the Financial Action Task Force on Money Laundering (FATF), and the World Bank and International Finance Cooperation (IFC) Performance Standards on Environmental and Social Sustainability.

6. In the short term, Canada should leverage its Presidency of the 2018 G7 Summit to establish the first ever list of transparency and accountability principles agreed upon by governments. Using IEEE's principles and the Draft AI R7D Guidelines as a starting point, the G7 should jointly endorse a "High Level Principles on Ethical AI." Canada should then promote the list at other global forums, such as the G20 and OECD. The recently announced multi-stakeholder conference on AI this fall, which Canada will host, is another forum to champion ethical AI. The goal of the event should be to convince other countries to adopt the G7 list.

7. In the long term, Canada should lead a coalition of states to establish a new global agency with the purpose of fostering knowledge about the societal implications of AI, monitoring and evaluating national regulatory efforts, and coordinating global efforts to regulate AI. Modelled after the FATF, the new agency should develop a series of regulatory recommendations for the ethical use of AI to serve as best practice for governments. The agency should then monitor implementation, ensure recommendations are up-to-date, and convene stakeholders regularly to promote the adoption of recommendations globally.

8. If Canada does not take a leadership position on the ethical use of AI, then another country inevitably will. Such an outcome is unfavourable for several reasons. First, Canada risks losing its emerging position as a global thought leader on the ethical and social use of AI. Second, other countries may advocate for policies that are unfavourable or potential harmful to human rights. Third, Canada stands to lose global talent that it would otherwise have attracted, if it was home to the global agency on AI. To mitigate against these risks, the government should convene likeminded stakeholders in Canada to put together a strategy to build the global center for ethical study and use of AI.

COMMUNICATIONS IMPLICATIONS/ACTIONS:

9. The proposed course of action would generate significant domestic and global media coverage. Overall, the reaction should be positive, as many will see the creation of a global agency geared towards fostering ethical AI as a positive development. Nonetheless, the Minister should be prepared with talking points to respond to questions regarding the agency's regulatory powers, the extent to which the private sector is involved, and whether the agency will be effective at encouraging transparency and accountability. To guide the direction of the narrative, the Minister should write an op-ed in domestic and foreign newspapers, host a news conference with foreign counterparts involved in the creation of the agency, and participate in Q&As with industry actors to assuage fears of regulatory overreach.

APPENDICES

Appendix 1 - Definitions

Artificial intelligence: In its broadest definition, artificial intelligence refers to “The capacity of computers or other machines to exhibit or simulate intelligent behaviour” (Oxford English Dictionary). Within AI, however, there is narrow AI, which refers to specific tasks performed in specialized and well-defined domains, and general AI, which refers to the ability to simulate a wide range of cognitive skills, mimicking human intelligence, such as creativity, learning, and reasoning.

Stakeholders: Anyone involved in the development and implementation of AI (ex. research, design, manufacturing, marketing) as well as those affected by the use of the AI system (ex. public agencies, individuals, end users).

Transparency: The ability to discover how and why a system made a particular decision. Transparency also includes the concepts of explainability and traceability. There are varying levels of transparency and may include the disclosure of the use of AI to end users.

Explainability: The ability to provide the reasons and steps leading up to an action or decision being carried out by an algorithm or system of algorithms. Broadly speaking, a human would be able to understand the process and actions of an algorithm or AI system.

Traceability: Documenting the steps and actions taken by an algorithm to reach a decision.

Accountability: The liability to account for and answer for one's conduct. In the case of AI, this may mean an obligation to explain a decision or issue to the public or an oversight agency, accepting the blame and consequences of a decision made whether through an algorithm or not, or accepting responsibility for the care and use of personal data.

Human-in-command: Where the ultimate decision of when and how AI is used lies in the control of humans.

Hard law: Formal binding agreements such as legislation, international treaties, and niche bans on AI in specific industries.

Soft law: Non-binding agreements, such as UN General Assembly resolutions and declarations, action plans, or codes of conducts.

Self-regulation: Lack of government involvement in the governance and development of AI. Instead, industry leaders, developers, and other AI-involved actors would create their own set of standards and norms, such as the *Partnership on AI*, which was founded by Amazon, Facebook, Google, DeepMind, Microsoft, and IBM.

Appendix 2 - Comparison of Explicitly Stated Principles

	PRINCIPLES								
ORGANIZATIONS	Creator responsibility	Fairness	Explainable	Traceable	Human-in-Command	User-Owned Data	Education	Whistle-blower protection	Disclosure that AI is in use
IEEE Ethically Aligned Design, v2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
UNI The Future World of Work	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
FAT/ML	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Asilomar AI Principles	Yes	Yes	No	No	Yes	Yes	Maybe	No	No
Canada Digital Disruption White Paper	Yes	Yes	Maybe	Yes	Yes	No	No	No	No
Draft AI R&D Guidelines	Maybe	Yes	Yes	Yes	Maybe	No	No	No	No
Association for Computing Machinery	Yes	Maybe	Yes	Yes	No	No	No	No	No
Partnership on AI Tenets	Maybe	Yes	Yes	No	No	No	Yes	No	No
Montreal Declaration	No	Yes	No	No	Maybe	Maybe	Maybe	No	No

Yes: Organization explicitly states it is in favour of the principle.
No: Organization has no explicit statement that it is in favour of the principle.
Maybe: Organization has used wording that suggests they may be open or in favour of the principle.

Clarification of Principles:

Human Responsibility: A human should ultimately be responsible for any algorithm or AI system in use. The excuse, “the robot did it” will not hold.

Fairness: The development and use of AI should promote justice and seek to eliminate all types of discrimination.

Explainable: The use of AI should be understandable and interpretable by people.

Traceable: The logic and technical process that led to a decision being made or action being taken by an algorithm or AI system should be made available.

Human-in-Command: The ultimate decision of when and how AI is used should lie in the control of humans.

User-owned data: The end user should be able to access and control who gets to use their personal data.

Education: There should be an increase in AI education for the general public.

Whistleblower protection: Those who draw attention to AI misuse should be protected legally.

Disclosure that AI is in use: Any use of AI should be disclosed to the user.

Organizations Listed:

IEEE Ethically Aligned Design, v2: IEEE is the world's largest technical professional organization, representing more than 400,000 engineers. Version 2 was released in December 2017 and was created by more than 250 cross-disciplinary thought leaders

UNI The Future World of Work: UNI Global Union represents more than 20 million workers from over 900 trade unions working in skills and services. Released in December 2017, the "Top 10 Principles for Ethical Artificial Intelligence" provides a framework to address transparency in AI.

FAT/ML: Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) is an organization of researchers and practitioners who are concerned with the unethical development of AI. Each year FAT/ML hosts a conference on the topic and recently released its own list of principles and best practices.

Asilomar AI Principles: These principles were developed in conjunction with the 2017 Asilomar conference, which was hosted by the Future of Life Institute – an organization that supports beneficial AI research and initiatives.

Canada Digital Disruption White Paper: Written by the Treasury Board of Canada Secretariat, this paper examines the political, ethical, technical, and legal considerations surrounding the use of AI within the Government of Canada. It provides a set of seven principles to guide the ethical application of AI by Canadian federal institutions.

Draft AI R&D Guidelines for International Discussions: As a result of 2016 G7 ICT meeting, the Japanese government held the "Conference Toward AI Networked Society" in October 2016. The conference led to the publication of this report in July 2017.

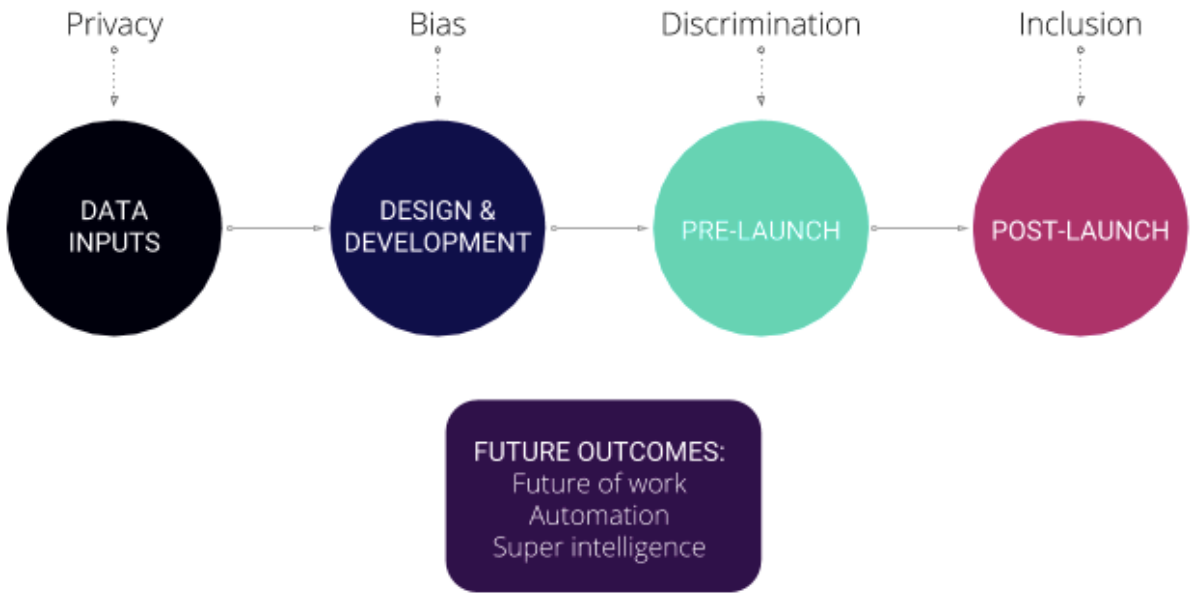
Association for Computing Machinery: ACM is the world's largest educational and scientific computing society. In January 2017, the organization released a public statement on algorithmic transparency and accountability that contained these principles

Partnership on AI: Created by Amazon, Apple, DeepMind, Facebook, Google, IBM, and Microsoft in September 2016, the Partnership was established to study and formulate best practices on AI technologies and serve as an open platform for public discussion.

Montreal Declaration: The Montreal Declaration for a Responsible Development of Artificial Intelligence was announced in November 2017 at the conclusion of the Forum on the Socially Responsible Development of AI. The Declaration is intended to start a large dialogue between the public, experts, and government decision-makers

Appendix 3 - Processes vs Outcomes Approach to Policy

While much debate has focused on the future outcomes of increased use of AI, many of the negative effects we've seen, such as discrimination or privacy infringement, occur within the development and implementation cycle of AI-based programs.



Appendix 4 - Analysis of Policy Options

	Pros	**Cons**
Hard Law	• Higher level of compliance for signatories • More uniform compliance across signatories	• Politically, economically, and socially difficult to achieve enough buy-in for an international law to work • Time horizon is not responsive enough for how rapidly AI is developing and

		how pervasive it is ● AI use and development is not globally uniform, which may marginalize much of the Global South. ● Even if a treaty were signed, reservations may be made by actors that would render the obligations ineffective ● Hard law works best with simpler issues, such as a mines ban. AI is an overly complex issue in comparison.
Soft Law	● Allows for regional or national development and evolution of policies that can adapt to real and changing environments. ● Easier to achieve buy-in ● More flexibility for a wide range of organizations (ex. a mature corporation versus a start-up will have varying degrees of capacity and financial resources) ● Allows for more sector-specific flexibility (ex. healthcare versus transportation will have different and unique requirements)	● Soft commitments with lack of legal accountability may lead to ineffective compliance ● In order to gain buy-in, regulations may be watered down to appease as many parties as possible.
Self-Regulation	● Private sector often has greater awareness of new developments and are at the forefront of technological innovation	● Profit tends to trump social responsibility, which may limit private sector's incentives to work in the best interest of society ● May exclude civil society actors and other communities, which can impact the diversity of voices ● Often does not allow for redress or other legal accountability mechanisms