BALSILLIE SCHOOL
OF INTERNATIONAL AFFAIRS

Discord and Disruption

2019 Global Trends Report

An Anthology of Briefing Notes by Graduate Fellows at the Balsillie School of International Affairs

The Foreign Policy Research and Foresight Division at Global Affairs Canada is proud to support and be associated with the Graduate Fellowship Program/Young Thinkers on Global Trends Initiative. The challenges facing Canada today are unprecedented and truly global. Tackling those challenges require fresh ideas and engagement with new generations of thinkers, researchers, and activists to help create opportunities for a sustainable future. We would like to thank the students and professors of the Balsillie School of International Affairs for their time, effort and commitment throughout the year to make this initiative successful. The results of their work, which has been encapsulated in this anthology, will help inform the work of Global Affairs Canada as it relates to foreign policy, trade and international development.

Global Affairs Canada | Affaires mondiales Canada

BALSILLIE SCHOOL OF INTERNATIONAL AFFAIRS

67 Erb Street West
Waterloo, ON N2L 6C2 Canada
Telephone: 226 772 3001

Governing Artificial Intelligence: Preventing and Preparing Should It Go Awry

Ruxandra Badea, Kristen Myers and Sulamita Romanchik

Issue

Given that artificial intelligence (AI) is a moving target, governance measures have been outpaced by AI technological growth, thus posing potential threats to human rights and global peace and security. The current lack of AI governance calls for a robust governance framework to prevent and prepare for AI going awry — one that aligns with Canada's feminist foreign policy and its history and reputation for upholding human rights around the world.

Background

The growth and expansion of AI has progressed rapidly domestically and internationally, creating transformations in every aspect of our lives. As these systems have become more powerful and independent, concerns surrounding AI "accidents" have grown, thus calling for the need for greater understanding and preparedness for these technological advancements.

Many academics, policy makers, private sector leaders, and others have voiced concerns over AI's ability to reproduce human bias. The probability of automating inequality and perpetuating human rights abuses creates an increased and urgent need for determining what Canadian values can be used to emulate a "value-by-design" approach, starting with the AI design process. The discussion of guiding principles has been at the forefront of many high-level meetings on AI in order to confront the issues surrounding bias, human values, and ethics. For this reason, the importance of informed decision-making has been brought up in conjunction with ethics and bias.

Additionally, the Global Summit on Human Rights in the Digital Age (RightsCon 2018) revealed two primary concerns currently surrounding AI systems, which amplify the likelihood of AI systems going awry: first, the commercialization and ownership of data hinders transparency and auditability, particularly of corporations, and the question of governments holding greater access to the data that is inputted into AI systems is at the forefront of halting the black box problem of AI; and second, the value of maintaining human auditing and supervising of any AI system at any time and continuously. There is ample consideration of establishing a threshold as to when human judgement should be used, hence limiting an AI system from becoming entirely autonomous; should AI systems become fully autonomous, there are growing concerns that humans would be entirely removed from the loop. As these systems begin to evolve on their own, the ability to find an explanation for, or to pinpoint at which moment exactly and for what reason an AI system goes awry will become nearly impossible. Designing AI with auditability in mind is something that cannot be disregarded as the global community delves into its complexities.

Finally, over a handful of international treaties and agreements exist that address technologies, arms, and weapons similar in the magnitude of their impact as, but less sophisticated than, AI. Innovation, however, is moving in the direction where such technologies and arms are evolving into more complex and advanced products — AI products that the treaties and agreements do not explicitly cover. The first of such treaties is the

Arms Trade Treaty, which outlines national export and import controls of countries in alignment with human rights and international humanitarian law considerations, including those contained in the Geneva Conventions of 1949, and additional protocols (United Nations Office for Disarmament Affairs n.d.a). Second, the Wassenaar Arrangement promotes both transparency and greater responsibility in transfers of conventional arms and dual-use goods and technologies, taking into account the dual-use nature of AI (Wassenaar Arrangement Secretariat 2017).

The Biological Weapons Convention is the first multilateral disarmament treaty banning the development, production, and stockpiling of an entire category of weapons of mass destruction, as well as weapons, equipment or means of delivery designed to use such agents or toxins for hostile purposes or in armed conflict (United Nations Office for Disarmament Affairs n.d.b). Similarly, the Chemical Weapons Convention prohibits the large-scale use, development, production, stockpiling, and transfer of chemical weapons (Organisation for the Prohibition of Chemical Weapons n.d.). Both these treaties also augment the 1925 Geneva Protocol, banning the use of such weapons in international armed conflict. Biological and chemical weapons can reach AI capabilities but the aforementioned conventions do not include any provisions on AI, as AI advancements were not yet foreseen.

Likewise, the Nuclear Non-Proliferation Treaty aims to prevent the spread of nuclear weapons and weapons technology, to promote cooperation in the peaceful uses of nuclear energy, and to further the goal of achieving nuclear disarmament and general and complete disarmament (International Atomic Energy Agency n.d.). As such, the exclusion of AI in the aforementioned treaties, conventions, and agreements inadvertently allows for the exploitation of loopholes and/or the production of AI weapons for mass destruction and its combination with existing weapons.

How should Canada Respond?

The recommendations proposed in this brief contain concrete measures that Canada can take to promote the ethical development of AI at home and abroad in the immediate and near future, and are in line with existing government initiatives and internationally agreed upon arrangements, conventions, and treaties. Moreover, Canada cannot claim to be an international leader if it does not reflect what it advocates for internationally. Thus, by leading by example at home, Canada can successfully position itself globally, while contributing to its international role as a convenor on international issues. This requires the consideration and examination of current AI governance initiatives and other government approaches to AI governance, enabling Canada to create its own governance mechanism that draws on the strengths of existing mechanisms and fills any gaps that other initiatives or states have missed.

Considerations for Global Affairs Canada

When implementing these policy recommendations below, it is important to consider the following: 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 ethical literacy 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 (GAC) to collaborate with other departments within the federal government to ensure that all public service workers are informed and trained on AI; its governors cannot be unaware of the implications of the emergent technology.

Scrutiny surrounding the growing use of, investment in, and governance of AI and the potentially harmful effects it may pose is expected to continue. Informed and comprehensive media outputs 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 take 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.

It is important to note that AI's integration into existing international agreements and treaties may entail a large political process. However, there is current political will and appetite, demonstrated through the various current national and international initiatives, in pursuing the governance of AI. Its integration into existing agreements, rather than the creation of new mechanisms, would be more politically feasible and would ensure existing players signed onto these agreements would be more empathetic towards these additional protocols. Further, by integrating AI into existing security agreements, the probability of states continuing to ignore the treaties is less likely, and formerly reluctant states may be more incentivized to sign and ratify the aforementioned agreements. Given the accessibility of AI to a greater number of states and actors, the question of inclusion and the importance of a variety of perspectives is necessary in order to minimize the risk of a global AI arms race. Inadvertently, these governance initiatives also contribute towards building a more inclusionary global platform within which a language can emerge that crosses diverse voices.

Existing Programs, Partnerships, and Initiatives

There are existing governance programs, partnerships, and initiatives in which the policy recommendations below fit well. This includes the government's $125 million Pan-Canadian Artificial Intelligence Strategy, led in partnership with the Canadian Institute for Advanced Research (CIFAR), which embodies the priority to develop global thought leadership on the economic, ethical, policy, and legal implications of advances in AI (CIFAR 2018). This also includes the government's procurement and job creation strategies, including the Innovation Superclusters Initiative, centred in the government's Innovation and Skills Plan, and the Innovative Solutions Canada program, part of the Innovative Skills Program (Government of Canada 2018).

Other private/academic sector governance initiatives include The Montreal Declaration for a Responsible Development of Artificial Intelligence, which sets out ethical guidelines for the socially responsible development of AI (Université de Montréal 2017). Drafted through a co-construction process, wherein individuals from all fields were involved, the final draft of the declaration was completed near the end of winter 2018 and is a living document that is revisable and amendable. Likewise, the Toronto Declaration on Protecting the Rights to Equality and Non-discrimination in Machine Learning Systems by AccessNow and Amnesty International focuses on discrimination in machine learning and is open for endorsement by different companies and corporations (AccessNow n.d.). Looking ahead, GAC can find the declaration to be another ideal platform within which to incorporate the policies this brief proposes.

Lastly, international governance domains include government commitments towards sound AI and robotics governance — including the EU's recent adoption of the General Data Protection Regulation and China's emerging data privacy system, the Personal Information Security Specification. Of particular importance is the EU's recent Declaration of Cooperation on Artificial Intelligence, signed in April 2018. The declaration is a commitment between 25-member states to *collectively* deal with the challenges of AI, while preventing states from engaging in an AI "arms race" (European Commission 2018). Keeping this in mind, Canada should explore the creation of its own governance mechanisms that draw on the strengths of these existing mechanisms, while filling gaps that these governance frameworks have overlooked.

Recommendations for GAC

Create guidelines to ensure development agencies abide by the five-point platform for AI development. A 2016 report, "Concrete Problems in AI" (published in conjunction with Google Brain, Stanford University, UC Berkeley and Open AI), sets out five practical points of accident risk prevention in terms of AI: avoiding negative side effects, avoiding reward hacking, scalable oversight, safe exploration, and robustness to distributional shift (Amodei et al. 2016). Implementing these into policy to be observed by developers, researchers, and corporations both domestically and internationally would provide Canada with a robust forward-looking safety plan.

Establish a multi-stakeholder advisory board comprised of industry, education, research, civil society and government personnel to update policy according to advancements in AI and to ensure that domestic industries abide by these guidelines. Involving multiple stakeholders in the ongoing discussion and governance process of AI is the ideal way to balance innovation and control. Simultaneously, the board can provide an ideal forum for building partnerships, addressing issues

surrounding AI and trade-offs between stakeholders, and for increasing trust and transparency between all stakeholders and the public. As Canada pursues a leadership role in AI, an advisory board provides the space for the Canadian government to take a greater role in the oversight of AI, which can redirect or instigate innovation of its own, and thus should not be seen as a threat to innovation.

Push for investment into ethics education and training for AI developers and associated stakeholders. Without a sound ethical background, all stakeholders involved risk violating Canada's human rights stance. The United States has already solidified its commitment to ethical education through the introduction of several bills in 2017, such as bill 4625 in the House and bill 2217 in the Senate (GovTrack 2017; JDSUPRA 2018), which Canada should also pursue. Ethical education should include the training of current public servants and diplomats, as these are the individuals directly formulating and negotiating treaties and agreements impacting AI governance.

Develop a code of conduct for AI engineers and others associated with its use and development. Drawing upon the EU's Code of Ethical Conduct for Robotics Engineers (Castel and Castel 2017), the Canadian government should establish a similar framework for all stakeholders involved in AI's use and development. By referencing this code, Canada can exert its leadership by including its world-respected values into its own code and then promulgate it world-wide. It is important to note that the EU's ethical code of conduct is voluntary. As Canada pursues a global, ethical AI leadership position, it should call for a more ambitious, mandatory ethical code of conduct. The task of developing the code of ethical conduct could be delegated to the proposed advisory board, whose multi-stakeholder constituency would ensure more inclusive standards.

Call for the incorporation, through the means of protocol(s), of AI-related policies into international treaties and agreements. As several AI applications can be used for both peaceful and non-peaceful means, and are thus dual-use, integrating AI provisions into the Wassenaar Arrangement would ensure that these transfers are not diverted to support unpeaceful means. Further, as concerns about autonomous weapons rise, incorporating AI provisions into the Arms Trade Treaty, Biological Weapons and Chemical Weapons conventions, and the Treaty on the Non-Proliferation of Nuclear Weapons, all of which do not take into account the possible or potential integration of AI systems, would prevent the unethical integration of AI into technologies that pose a threat to global peace and security. Not only does the integration of AI into these existing global agreements help to uphold global peace and security, but it also aligns with Canada's feminist foreign policy stance by prohibiting any potential AI weapons that could facilitate or promote violence against women and other vulnerable populations.

About the Authors

Ruxandra Badea is a student in the University of Waterloo's Master of Global Governance program based at the BSIA.

Kristen Myers is a student in the University of Waterloo's Master of Global Governance program based at the BSIA.

Sulamita Romanchik is a student in the University of Waterloo's Master of Global Governance program based at the BSIA.

Acknowledgements

The authors would like to thank their supervisors David Welch, Scott Janzwood and Jinelle Piereder, as well as Global Affairs Canada for their support throughout this project.

Works Cited

AccessNow. n.d. "The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems." www.accessnow.org/cms/assets/uploads/2018/05/Toronto-Declaration-D0V2.pdf.

Amodei, Dario, Paul Christiano, Dan Mané, Chris Olah, John Schulman, and Jacob Steinhardt. 2016. "Concrete Problems in AI Safety." https://arxiv.org/pdf/1606.06565v1.pdf.

Castel, E. Matthew and Jean-Gabriel Castel. 2017. "Should Canada Adopt some of the Proposals Listed in the February 2017 European Parliament Resolution on Civil Law Rules on Robotics?" *The Advocate's Quarterly* (47): 500–12.

CIFAR. n.d. "Pan-Canadian Artificial Intelligence Strategy." www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy.

European Commission. 2018. "EU Member States sign up to cooperate on Artificial Intelligence." https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence.

Government of Canada. 2018. "Innovation Superclusters Initiative."

www.canada.ca/en/innovation-science-economic-development/programs/small-business-financing-growth/innovation-superclusters.html.

GovTrack. 2017. "H.R. 4625: FUTURE of Artificial Intelligence Act of 2017." www.govtrack.us/congress/bills/115/hr4625/text.

International Atomic Energy Agency. n.d. "Treaty on the Non-Proliferation of Nuclear Weapons (NPT)." www.iaea.org/publications/documents/treaties/npt.

JDSUPRA. 2018. "Multiple Artificial Intelligence Bills Introduced in House and Senate." www.jdsupra.com/legalnews/multiple-artificial-intelligence-bills-22422/.

Organisation for the Prohibition of Chemical Weapons. n.d. "Chemical Weapons Convention." www.opcw.org/chemical-weapons-convention/.

RightsCon. May 16-18, 2018. Toronto, Canada: Beanfield Centre at Exhibition Place.

United Nations Office for Disarmament Affairs. n.d.a. "The Arms Trade Treaty." www.un.org/disarmament/att/.

———. n.d.b. "The Biological Weapons Convention." www.un.org/disarmament/wmd/bio/.

Université de Montréal. 2017. "The Montreal Declaration for a Responsible Development of Artificial Intelligence: A participatory process." https://nouvelles.umontreal.ca/en/article/2017/11/03/montreal-declaration-for-a-responsible-development-of-artificial-intelligence/.

Wassenaar Arrangement Secretariat. 2017. The Wassenaar Arrangement On Export Controls for Conventional Arms and Dual-Use Goods and Technologies WA-DOC (17) PUB 001. www.wassenaar.org/app/uploads/2015/06/WA-DOC-17-PUB-001-Public-Docs-Vol-I-Founding-Documents.pdf.