Ontario trustworthy AI framework – May 2021 consultations – TBS comments.

Ontario is holding consultations on their [proposed Trustworthy AI Framework](https://www.ontario.ca/page/ontarios-trustworthy-artificial-intelligence-ai-framework-consultations), until June 4th. Below are the comments prepared by Benoit Deshaies and Omar Bitar of the Treasury Board of Canada Secretariat, Office of the Chief Information Officer, on 2021-06-02.

Comments

N.B. in this document, we use Ontarians loosely to mean either individuals or businesses.

General comments

1. This is a fantastic effort on the part of the Ontario government to advance the responsible development and use of AI in Ontario. This work will lay the foundations for a modernized policy and legal environment which will build public trust in AI applications in the public sector while boosting the competitiveness of Ontario’s AI ecosystem and supporting the growth of its digital economy.
2. The text of the commitments and actions is clear and simple, making this complex issue very accessible and easily understood.
3. The proposal is missing a clear definition of AI and algorithm, to help specify the types of technologies, techniques, and systems addressed in the framework. AI may be used as an umbrella term here, but it risks being interpreted as only including machine learning, and not including other forms of automated decision making such as conventional programming, decision trees, rules-based systems, etc. We recommend including a definition or clarification of intent.
4. The text in Ontario’s [Digital and Data Strategy](https://www.ontario.ca/page/building-digital-ontario) on the trustworthy AI framework suggests that this is a cross-sectoral initiative that would have implications for both the public and private sectors. It is not clear to us whether the framework would only impact government operations, or whether certain measures may apply to AI development and use by corporations as well. We suggest articulating the scope of the framework to clarify the conditions of its applicability. Section 5 (Scope) of the TB [Directive on Automated Decision-Making](https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592) (“the federal directive”) can serve as an example of a targeted approach to defining the scope of an instrument.
5. Mechanisms to monitor the progress of this initiative towards the proposed goals and desired outcomes will be needed. This includes compliance mechanisms for public sector organizations and other entities falling within the purview of the framework. As well, it involves crafting performance metrics for each commitment to gauge and report on progress over time.
6. A governance structure with clearly defined roles and responsibilities with respect to the framework is needed to oversee, advise on, and prioritize development and implementation.
7. The [Alpha Principles for the Ethical Use of AI and Data Driven Technologies in Ontario](https://github.com/ongov/AI-Principles) (Alpha Principles) and [Transparency Guidelines for Data-Driven Technology in Government](https://github.com/ongov/Transparency-Guidelines) (Transparency Guidelines) can be leveraged to further elaborate and expand on the framework. The link between these previous efforts and the proposed Framework is unclear.
8. Some measures proposed in the framework clarify that they apply for AI making decisions that affect people directly, but others not. We see two potential categories: those affecting specific individuals, such as automated service-delivery decisions, and those affecting a larger population, such as those supporting program or policy design. Both categories warrant attention, but with different measures.
   1. “Right to explanation” may not be appropriate for systems supporting policy setting, but ensuring quality, representative and inclusive data would apply to both.
   2. AIs supporting policy setting is not an area we have explored significantly in the federal responsible AI program, but this is an area with potential to ensure governmental efforts benefit all Ontarians equally. The use of disaggregated data to evaluate the impacts on gender, geographical, age and language categories, for example, could be helpful.
   3. Commitment 1 “No AI in secret” seems to target AI affecting specific individuals, and Commitment 2 “AI use Ontarians can trust” may be wider in scope, for example. We suggest clarifying when each category of AI use is targeted.
9. Given the rapid pace of change and evolution in this space, we suggest building in a periodic review process for the framework to ensure that it adapts to changing circumstances and is able to meet current and emerging challenges effectively.

## Commitment 1: No AI in secret

**The use of AI by the government will always be transparent, with people knowing when, why, and how algorithms are used and what their rights are if harm occurs.**

1. A definition of harm in the context of AI will be needed along with criteria to enable consistent determinations in accordance with applicable policy and legislation. The criteria would be used to help classify the risk-level of a use-case should Ontario adopt a risk-based approach to AI governance.
2. Regarding the problem statement, we suggest considering that AI is being used not only to inform decisions but also to make them. This framing might also place disproportionate burden on the client to trigger assessments of bias in AI systems; this is further explored in the action “Create accountability...” below.
3. What type of fairness is envisioned in the goal statement and how does it differ from the notion of fairness implicit in equity? One of the key principles of the federal directive is procedural fairness, as understood and practiced in administrative law. This could differ from substantive notions of fairness, for example, in which the fairness of a decision (rather than the process through which it was reached and the capacity of the client to contest it) is scrutinized.

**Be fully transparent when using algorithms to interact with the public (e.g. rules to require the public be informed if they are interacting with a machine or have decisions made about them by an algorithm)**

1. It would be helpful to clarify this would also apply when automating part of the fact-finding or analysis process with the potential to influence the decisions of government officials. Slide 13 “Recommendations and decisions” of the presentation at (<https://tinyurl.com/tbs-ai-presentation>) may be helpful here.
2. This could be expanded to be transparent about AIs informing program design and policy setting.
3. There should be clear requirements about what information needs to be proactively disclosed.
4. Transparency measures may necessitate access to proprietary AI systems which the government should ensure through policy or legislation, taking the relevant caveats (e.g., trade secrets, trade agreement provisions, classified data processing, etc.) into account. Ontario’s Transparency Guidelines (‘Allow Meaningful Access’) expand on the need to assure access for oversight and accountability purposes.
5. Any exemptions to transparency measures (e.g., for reasons related to public safety) need to be clearly specified and elaborated.

**Create accountability for the use of AI in the government by giving people rights to address potential biases created by the AI (e.g. right to explainability, right to contest, and right to opt out)**

1. “giving people rights to address potential biases created by the AI”; we think it’s for governments to address biases. The rights proposed for explainability and contestability can be useful and should be kept, but presented differently. It could be, for example, “ensure Ontarians’ rights to a meaningful explanation and to contest decisions are preserved when decisions are automated”.
2. The right to an explanation for significant decisions that impact Ontarians is an important one and may be anchored in administrative law. AI explainability requirements could be differentiated based on the level of risk or impact of a use-case (especially if Ontario opts for a risk-based approach). The federal directive (Section 6.2.3, Appendix C) provides two approaches to providing meaningful explanations to clients: common explanations for impact levels I-II (e.g., via a FAQ or other common resource) and more individualized explanations for higher impact (III-IV) use-cases.
   1. Further, Ontario’s Alpha Principles highlight the importance of providing plain-language explanations that are meaningful to laypersons and experts alike.
   2. Explainability considerations can impact the selection of AI technologies in service contexts, as the outputs of some technologies (e.g., neural nets) are less amenable to explanation than others (e.g., linear models). Section 4.5 of the [Guideline on Service and Digital](https://www.canada.ca/en/government/system/digital-government/guideline-service-digital.html) provides guidance on the role of explainability in model selection and design, highlighting the importance of choosing models which yield interpretable outputs that can be explained.
3. The right to opt out is an interesting one; we don’t think it is necessary with a properly designed right to contest. Ontarians could try the automation first, and if the Ontarian is not happy with the outcome, then they can contest it. For this to be effective, we need to ensure that there is an easy way to request to have a person review the decision in a timely manner. The right to opt out of certain algorithmic decisions, such as being assigned a risk score, is difficult to implement, and could create challenges in delivering government programs.
4. In addition to creating or strengthening digital rights, there is value in considering designating officials (e.g., senior officials in ministries) responsible for the development and use of AI, and compliance with the framework. This point builds on an earlier point on the need for a governance structure to oversee the framework and is also noted in the ‘Accountable and Responsible’ principle in Ontario’s Alpha Principles.

**Provide clarity and transparency to the public on how Ontario collects data for use in algorithms (e.g. explore options to update provincial notices of collection to inform the public if data collected is used to develop algorithms for decision-making.).**

1. This is a good measure. It could require, when it is known, to state the algorithm that the data will support. We should avoid generic disclaimer at the time of collection, such as “the data collected may be used to train algorithms in the future, maybe”, which may raise unnecessary concerns, and encourage data use beyond the purpose for which it was collected.
2. In addition, when informing Ontarians of a decision affecting them, any relevant decision-making algorithm that supported that decision could be noted. Also – on request? – an explanation of the data that informed those algorithms, the processing of that data, and a description of how it supported the decision, could be provided.

## Commitment 2: AI use Ontarians can trust

**Rules and tools are in place to safely and securely apply algorithms to government programs and services based on risk.**

1. Would the proposed risk-based rules apply to any instance of AI use in government (e.g., does this cover research, would it impact employees subject to AI use, is it limited to administrative decision making)? Further to an earlier point, there may be value in clearly articulating the scope of the framework in general and of risk-based rules in particular.

**Deliver recommendations on ways to update Ontario’s rules, laws and guidance to strengthen the governance of AI, including whether to adopt a risk-based approach to determine when which rules apply.**

1. A risk-based approach would be consistent with the federal directive, using the [Algorithmic Impact Assessment](https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html) (AIA) to determine risk and determine proportionate requirements. Another approach could be to have an additional set of requirements for pre-identified use cases deemed “high risk”, such as what is proposed in the draft EU regulation.
2. The proposed recommendations on ways to modernize Ontario’s AI rules could address the scope and aim of regulatory reform (i.e., what AI technologies, use-cases, sectors, and actors would be regulated, under what conditions, and to what end). The Law Commission of Ontario’s recently published [report on AI regulation](https://www.lco-cdo.org/wp-content/uploads/2021/04/LCO-Regulating-AI-Critical-Issues-and-Choices-Toronto-April-2021-1.pdf) may be of interest in this regard as it surveys the regulatory environment in Canada and proposes relevant recommendations addressed to the Ontario and federal governments identifying the proper form and scope of a comprehensive regulatory regime for AI.

**Assess whether to use an algorithmic assessment tool as a way to measure risk, security, and quality.**

1. An algorithmic assessment tool such as the federal AIA could support consistent evaluations of risk and justify corresponding mitigations in areas such as data quality, privacy, security, ethics, and legality.
2. Measures to strengthen algorithmic assessments include:
   1. Third-party review mechanisms to validate assessments of data, systems, risks, impacts or outcomes. The TB Directive on Automated Decision-Making's peer review requirement (6.3.4) can serve as an example of how such a mechanism can also foster cross-sector engagement and collaboration (e.g., with academia and civil society). Audits and peer review are noted in Ontario’s Alpha Principles as ways to ensure accountability for biased or otherwise harmful AI systems.
   2. the publication of the results from the assessment and the 3rd party review, as required by the federal directive (6.1.4, and appendix C “notice”), could help strengthen the commitment to algorithmic transparency. It can serve as a mechanism through which the public learns about how, when, and why the Ontario government collects data for use in an AI context. Ontario’s Alpha Principles can inform the type of information made available to a client or the wider public (e.g., impacts of personal data processing). This information can also be shared as part of a public AI registry documenting active and previous uses of AI in the public sector.
   3. Assessment tools can be developed in collaboration with the public, as suggested by Ontario’s Transparency Guidelines.

**Ensuring processes are in place so that algorithms are continuously tested and evaluated for bias/risk and whether audits or human oversight controls are needed.**

1. We agree with these requirements, and suggest they should be proportional to the potential risk of the system. Some automated decisions may not require human oversight, and imposing that would hamper efforts to make the operations of the government more effective. A risk-based approach can foster a more efficient allocation of human resources and lower the barrier to adopting AI in relatively low-risk cases. An example of a risk-based approach to human oversight is provided by the federal directive, which requires a ‘human in the loop’ during the decision-making process and for the final decision in use-cases with impact level III-IV.
2. Proposed processes to evaluate and address bias should account for both model and data bias throughout the AI system lifecycle. It could be clarified that the ongoing testing should also include looking for bias in the outcomes of the system, i.e. the decisions rendered, not only the in the input data. Such processes could help guard against harmful outcomes (e.g., discriminatory treatment of clients, inequities in the reach of programs or services, exposure of personal or sensitive information through use in AI contexts). Requirements 6.3.1 and 6.3.2 of the federal directive are relevant examples to consider in this area. The ‘Safe’ principle in Ontario’s Alpha Principles also underlines the need for ongoing monitoring and mitigation planning.

## Commitment 3: AI that serves all Ontarians

**Ontarians benefit economically and socially from AI technologies that are rooted in individual rights and reflect the diverse communities across the province.**

1. The systemic nature of some of the risks posed by AI – to social cohesion, democratic legitimacy, the rule of law – indicates a possible need to ensure that the framing of this commitment accounts for such risks alongside those to individual rights. Ontario’s Alpha Principles (the ‘Good and Fair’ principle) highlight systemic risks.
2. The potential actions under this commitment do not address how the government might reap the social and economic benefits of AI. The [Pan-Canadian AI Strategy,](https://cifar.ca/ai/) relevant provincial investments in Ontario’s AI ecosystem, and Ontario’s Alpha Principles (‘Sensible and Appropriate’) could help guide the framework’s actions in this area.

**Embed equity and inclusion in the use of data and digital tools by requiring organizations to take steps to mitigate potential harms (e.g. data set requirements, documentation requirements for traceability, accountability provisions).**

1. Traceability and metadata on the provenance of data is important, as well as a description of the processing of the data to aggregate it and cleanse it. The federal directive requires in 6.3.3 that the data be “relevant, accurate, [and] up-to-date”, and that it be tested for “other factors that may unfairly impact the outcome”, which could be understood to mean that it’s representative of the target population and sufficiently diverse, among others. The AIA also includes questions on data quality.
2. Data requirements aimed at fostering equity and inclusion should prioritize data quality. Ensuring that ‘AI data’ is accurate and representative of the relevant population(s) is crucial for achieving these goals and minimizing the risk of harm (e.g., biased systems and outputs). Existing provincial data quality frameworks can be leveraged to guide quality assessments of data in the AI context.

**Engage with sector leaders and civil society to develop a standard for “trustworthy AI” and a process to certify that vendors are meeting the government’s standard.**

1. Sector-specific standards are a good idea. It may also be relevant to certify certain commercial products against the standards, rather than the vendors themselves, given the breath of offerings from certain vendors.
2. Since good and certified tools can be used poorly, and beyond their intended usage, this should supplement other measures rather than replace them; the government should still be held accountable for the tools it deploys and how it deploys them, and for meeting the other requirements of this framework.
3. Vendor certifications can be informed by the algorithmic assessment tool considered under the 2nd commitment: adhering to a standard in order to be eligible to provide AI services to the public sector will likely involve being capable of mitigating risks which the tool would help identify.

**Assess whether the government should prohibit the use of AI in certain use cases where vulnerable populations are at an extremely high risk.**

1. The federal approach is to impose strict requirements to ensure that high-risk systems are properly managed, rather than prohibit certain uses entirely (keeping in mind that certain uses would already be prohibited by virtue of being illegal as per other Acts). Systems automating decisions with a very high impact (level IV) would require approval from the ministers (the “Treasury Board” itself and not the Secretariat) to operate.
2. The European approach in its draft regulation to identify certain use cases as "prohibited” is worthy of attention and we note that it is not because they affect particularly vulnerable segments of the populations, but because they are dangerous to people’s safety or people’s fundamental rights. If Ontario is considering to move ahead with prohibitions, we recommend that it be along similar lines instead of use cases affecting vulnerable populations.
3. As noted earlier, determinations of potential harm (and therefore high risk) require a well-defined concept of harm in the AI context as well as criteria supporting consistent evaluations (which would include factors such as high-risk vulnerable populations). These determinations are critical to decisions on whether to prohibit certain applications of AI.