Comments on the Toronto Police Services Board Proposed Policy on AI Technologies

Montréal Society and Artificial Intelligence Collective (MoSAIC)

15 December 2021

Table of Contents

About MoSAIC

The Montréal Society and Artificial Intelligence Collective (MoSAIC) is a coalition of interdisciplinary and multi-sectoral experts who investigate the impacts of technology on society, communities, and equality-seeking populations. Members have expertise in disciplines including computer science, law, communications, media studies, geography, and sociology. Members come from academia, civil society, and industry. We believe that this diversity of viewpoints is fundamental to understanding the broader impacts of artificial intelligence.

Executive Summary

We welcome the Toronto Police Service Board's (TPSB) proposed policy to introduce guidance for the Toronto Chief of Police on the use of artificial intelligence (AI). The proposed policy and consultation is long overdue. Before addressing our specific recommendations focused on evaluation, explainability and interpretability, and use, procurement and development, we wish to raise four substantive points about the proposed policy.

1. Any implementation of AI technologies by law enforcement needs to begin from the assumption that it cannot reliably anticipate all the effects of those technologies on policing or policed communities and act accordingly in light of these impacts. Examples of algorithmic bias and other problematic behaviour surface in the press weekly. Computer scientists continue to refine modeling and algorithmic techniques every year, both raising new problematic behaviour in existing algorithms and proposing new approaches. Since the technology is constantly changing, predictions of how its deployment affects different populations must also constantly change.

2. Toronto Police Services (TPS) policy should approach AI with caution, reflecting a greater rigour and transparency than at present, and with greater rigour and transparency than currently proposed. AI has been in use by the TPS since at least 2016; yet, the proposed policy does not include evidence about existing practices or potential harms. The proposed policy includes a problematic spectrum of risk categories such as "Extreme Risk" categories (clearly referencing Clearview AI) that are illegal for police use in Canada. Proposals such as the Risk Evaluation Committee remain too underdeveloped in the proposed policy to allow effective feedback.

3. The TPS must prioritize consideration of and prevent any potential infringements of the *Canadian Charter of Rights and Freedoms* before they occur. We believe there to be sufficient national research on these risks (including Robertson, Khoo and Song, Robertson & Khoo, Yuan Stevens, as well as Tamir Israel) to justify this recommendation. Racial bias and discrimination on the basis of race as well as other protected grounds are constitutive features of biometrics; these biases cannot be programmed out. Predictive policing, meanwhile, problematically maps statistical regularities in groups onto individuals. Biases in AI pose distinct risks for Toronto, one of Canada's most multicultural cities with known anti-Black bias in its policing. The use of AI can exacerbate existing biases; for example, it is well-known that racialized, Indigenous and Black peoples are more likely to suffer violence at the hands of the police in Canada (and Toronto) and that systemic racism is pervasive in policing in Canada more broadly.

4. The Extreme Risk category is actionable immediately. We know the human rights risks related to predictive policing tools such as biometric recognition technologies are so great that the proposed policy must prohibit the use of these “Extreme Risk” technologies, which can be used for intrusive mass surveillance particularly when used in live settings. The long timeline to phase out Extreme Risk technologies (2024) is unacceptable and allows for the continued use of presumably illegal technologies in the Toronto Police Services. An immediate review should be undertaken by the Police Chief to identify Extreme Risk AI technologies in operation. Any Extreme Risk Technologies should be immediately retired.

With these comments in mind, we have identified the following three sets of specific issues with the TPS draft policy that we address as follows.

Evaluation. The TPS policy should:

- Treat bias as a systemic and constant problem in AI and act as if it will need to be accounted for in any application through internal audits, assessment of performance indicators and risk assessment undertaken as recommended below and in line with best practices.
- Align its timeline for and regularity of review the with speed of changes in AI technology
- Develop internal expertise on AI auditing.
- Develop performance indicators independent of, and prior to, the adoption of any particular technology and apply them consistently and publicly across specific applications.
- Issue regular public reports on its evaluation processes.
- Conduct risk assessments in cooperation with community members, including those who are most likely to be harmed by the use of AI and compensate them for their time and efforts.

Explainability and Interpretability. The TPS policy should:

- Define explainability in collaboration with community groups, especially groups representing over-policed populations such as Black and Indigenous peoples.
- Take a slow approach to AI adoption and work with community groups in the process of defining explainability.
- Work with explainability research experts who have no commercial or political interest in the adoption of a particular AI technology

Use, procurement and development. The TPS policy should:

- Develop and make public the privacy and algorithmic impact assessments the TPS uses in line with international best practices.
- Require suppliers to abide by assessment requirements, and take steps to mitigate against those risks or otherwise cease using the technology.
- Provide the public with a list of all the vendors and entities whose AI tools the TPS has procured immediately, rather than by the end of 2024.
- Apply transparency and risk assessment standards as described below to the procurement of AI technology by partner agencies such as the RCMP for tools that are used by the TPS.
- Adopt the open contracting data standard.

The remainder of this report explains these issues in more detail and provides links to references.

Review and Recommendations

1. Evaluation

While the draft policy document contains recommendations about measuring and monitoring impact, it could be substantially improved around development and robustness of evaluation.

Performance indicators (5(m), 5(n), 18(a)) are not clearly defined and, as a result, invite a lack of transparency. Performance indicators measure the degree to which a specific AI technology aligns with the expected use. Naïve or improperly-applied performance indicators can be misaligned with the expected use of the system and result in unmeasured or systematic under-performance. Even well-known performance indicators like accuracy or error rate have been demonstrated to give the false impression of an effective system. In the current proposal, indicators will be poorly-defined or inconsistent across technologies for a given application case.

The suggested five year review of the system is not sufficiently frequent for a reasonable system.The pace of change in AI can vary in a matter of days or months so the suggested five year review of the system is not sufficiently frequent for a reasonable system. Moreover, the efficacy of modern artificial intelligence systems can depend on changes to the data used, changes to “learned parameters”, or manual/engineering changes to properties of the system (e.g., software architecture, loss function). This is especially true of AI algorithms whose behaviour intentionally changes over time. In these situations, the performance of the system can vary dramatically depending on the deployment environment, algorithm design, or engineering organization.

The TPS needs to treat bias as a constitutive feature of, and effect of, the use of data-driven technologies in policing, one that must be named, and accounted for at every step. Currently, the proposal treats bias (s. 1(c)(i)(4) and s. 1(c)(ii)(1)) as a surface feature of training data sets that can be eliminated, analogous to an individual prejudice (e.g., by equating bias with “malicious” action or being of “poor quality”) rather an a systemic effect of institutions and structural inequalities. Data about arrests or criminal prosecutions in a neighbourhood may be inaccurate, which is one kind of error that is represented as a bias. But a dataset could also accurately reflect systemic biases in policing and law enforcement, therefore reinforcing an underlying injustice. In other words, accurate data about racist systems can have racist effects when operationalized in a machine learning system. Further, criminal justice experts have criticized predictive policing because it is unfair to make criminal judgments about individuals based on statistics applied to groups. AI-driven predictive policing relies on precisely this kind of probability-based reasoning. Here, a “bias” exists in the very application of the technology, independently of whether the data set is sufficiently robust.

Recommendations

Best practices in AI systems include rigorously developing performance indicators and periodically re-evaluating them, since the environment or system may change in behaviour. Moreover, best practices for deployed AI systems include the continuous tracking of performance indicators through dashboards and other interfaces to detect bias.

Based on these concerns and best practices in AI systems, we recommend the following:

- Conduct an internal audit of existing sources of bias as a ground basis for the assessment of bias introduced by AI and establish methods for contextual pre-assessment of bias in high risk applications.
- TPS should develop internal expertise in algorithmic auditing and AI oversight. Following Canadian best practices, internal expertise, ideally an independent team, should have the power to conduct investigations either on a routine or ad hoc basis based on complaints, evidence of non-compliance, or at its own discretion.
- Internal audits and internal investigations should be reviewed by an external and independent body, and could potentially involve civilian oversight.
- Before any technology is considered, clear performance indicators—including metrics for unfairness—should be publicly developed for general application cases, independent of a specific technology. Moreover, a performance indicator review should be conducted every three months to revalidate their reliability.
- To improve consistency of reporting before deployment, all technology for a general application case should share base performance indicators under comparable experimental conditions.
- Any artificial intelligence technology should be deployed with continuous evaluation, including the tracking performance metrics. In addition to aggregate metric performance, anonymized individual system errors should be included.
- The TPS should make public its risk assessment and performance indicator reports, including what and who cannot be measured and should at minimum produce a public version of its audits, assessments, investigations and reports.
- The process for risk assessment should be developed in cooperation with community engagement experts to ensure that community groups who stand to be the most harmed by this technology, including Black and Indigenous peoples, can effectively participate.
- Community participants should be compensated for their time and efforts.

2. Explainability and Interpretability

The final policy should cooperate with community groups (e.g., overpoliced communities) in the process of defining and assessing explainability, and provide for continuous evaluation of explainability. Explainability (i.e., of inputs, processes, and outcomes) is important for ensuring that AI technologies remain accountable to users and affected populations. The draft policy describes explainability as a distinguishing characteristic between risk levels (1(c)(ii)(4)), but does not define it. The draft policy leaves several unanswered questions: Who decides what is explainable? What is explainability used for? How does explainability differ from transparency? What do we want explanations of?

In Section 1(c)(ii)(4) of the proposed policy, lack of full explainability is a sufficient condition for a technology to be considered as high risk. We commend that TPS's proposed policy explicitly connects explainability and its links to risk at s. 1(c)(ii)(4). Explainable systems can foster trust, enable accountability, improve evaluation, and facilitate modifications to problematic systems. At the same time, getting explainability wrong can result in misleading evaluations of what AI technologies are doing, hampering attempts to limit negative impacts and leading to misplaced trust.

Unfortunately, the proposal neither defines explainability nor discusses it elsewhere. Explainability is a variable notion whose precise definition results in different impacts. Let us explore some possible definitions for "fully explainable." One option is *interpretable*: a system is interpretable if it obeys a domain-specific set of constraints so that it is itself clearly understandable to humans. Interpretable systems need no additional methods to generate explanations. Examples of interpretable systems include scoring systems (e.g., used in medicine for diagnosis and prognosis of various conditions) and decision trees. Demanding interpretability amounts to restricting the types of AI technologies that may be deployed. For example, neural networks are typically not interpretable. Facial recognition technology that depends on neural networks would therefore be prohibited under this definition of "fully explainable".

One could also define "fully explainable" another way. According to Arrieta et al. (2020, p. 85), "[g]iven an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand." "Fully explainable" could therefore also mean that the details or reasons (1) are highly *faithful* to the AI technology and (2) are useful to the application under consideration. The *faithfulness* (or *fidelity*) of an explanation is the degree to which it accurately represents the way in which the AI technology maps inputs to outputs.

For example, suppose the TPS wanted to evaluate how much an AI system used racial attributes in its processing. We know that predictive policing relies on data that has a long history of racial discrimation, so let us assume that efforts are made to remove explicit racial identifiers. Algorithms exist to estimate which features were essential to the output. Unfortunately, in certain contexts, many such algorithms perform no better than random at identifying the essential features: the algorithms are not faithful. Moreover, attributions are susceptible to possibly imperceptible distortions in the data, and errors in the model and evaluation procedure. Even if the TPS did consider racial attributes (skin colour, ethnicity, immigrant or refugee status, or Indigeneity) as essential to the output, the explanation-generating algorithms may understate the role of race in determining patrols or arrests. The TPS can deploy a system that discriminates on the basis of race.

Selection of a definition for "fully explainable" must navigate a number of trade-offs. Explaining AI systems, whether by humans or algorithms, takes considerable time. There is a trade-off between the speed of development and the time taken to explain AI technologies. In some cases, there also is a trade-off between the accuracy of an AI technology and its simplicity. A simpler AI technology can be interpretable, or may provide explanations with higher fidelity than can more complex AI technologies. Certain applications may not easily allow for the use of a more simple AI technology that achieves a desired degree of accuracy. Given these trade-offs, there is benefit in a slow approach to AI adoption by TPS that could involve cooperation with over-policed community groups such as Black and Indigenous communities in order to define explainability and to begin addressing the risks associated with AI (for more information on the meaningful participation of communities regarding AI, see for example work by Sasha Costanza-Chock and Ngozi Okidegbe).

Given the variety of ways of defining "fully explainable" and the trade-offs, the most appropriate definition depends on the users and the context. The TPS must outline a process for defining "fully explainable" with the cooperation of community groups for each proposed deployment of an AI technology. All impacted community groups must be involved in this definition. In the TPS's proposal, if only law enforcement officials are involved in the deployment of a technology, applying the TPS's proposed risk classification might result in failure to classify a technology as high-risk when it poses significant risk to over-policed communities as well as communities in general.

Recommendations

Based on the concerns we have raised, we recommend the following:

- For each proposed deployment of an AI technology, meaningfully work with community groups to outline a transparent process for (1) defining explainability, (2) assessing the degree of explainability of any deployed models, and (3) continuously evaluating systems for explainability. This process should work directly with all impacted

community groups at each step. Any proposed explanation-generating algorithms in (2) should be demonstrated to community groups.
- Emphasize a slow approach to AI adoption that ensures that the explainability needs of community groups are fully explored.
- In applying explanation-generating algorithms to proposed AI technologies, ensure that experts specializing in AI explainability (who have no commercial or professional interest in the technology under evaluation) are consulted.

3. Use, Procurement and Development

While the TPS draft Policy states that it will not procure, utilize or deploy a new AI technology deemed to be of extreme risk, it is unclear how existing (or already in use) AI technology is being procured, and what the mechanisms for transparency and accountability are in place. Use or procurement of AI also includes the trialling, temporary deployment or testing of tools or freeware, as demonstrated by the testing of Clearview AI by police in Canada.

The pipeline for new AI technologies is unclear in the proposed policy including but not limited to further details about the suppliers and their capacity to work effectively in the policing areas. Will the TPS be developing its own AI technology or will they be obtained through procurement? Trade agreements and intellectual property law may prohibit proper transparency.

The TPS should elaborate its internal development processes for AI technologies. There are few details about the internal capacity, the conditions of development, the present standard for data collection internally and how data might be used for internal development.

The proposed policy should improve supplier accountability when procuring AI technology. We commend the proposed policy for requiring that assessment tools for AI including “a privacy impact assessment be conducted in advance of their deployment or use”. At a federal level, AI suppliers need to be vetted and the submission of an ethical statement is not required before being placed on a trusted vendors list. More so, AI might be included in larger systems (e.g. Microsoft Outlook) that might require changes to overall procurement to ensure that AI technologies do not accidentally enter into use at the TPS. Improved standards and capacity tests for suppliers would streamline the procurement process.

Recommendations

- TPSB should launch, consult, develop and publish its privacy and algorithmic impact assessments tools to be used for development and procurement, referencing international best practices.

- TPS should require suppliers to abide by its assessment protocols. TPS and suppliers should take proactive steps to mitigate against those risks or cease using the technology if risks cannot be adequately mitigated (e.g., biometric recognition tools, Extreme Risk technologies).
- TPS should curate and immediately make public a list of all the vendors/entities from whom they procure AI technology, similar to Public Services and Procurement Canada (PSPC) and the TBS's list of pre-qualified AI suppliers.
- Procurement of AI technology by partner agencies (e.g., RCMP) used by TPS must also be transparent and meet the standards required by the privacy impact assessment, the Algorithmic Impact Assessment tool, as well as the Board's oversight.
- TPS should consider adopting the open contracting data standard (OCDS).

Signatories

All signatories are included in alphabetical order.

Anonymous

Ana Brandusescu
PhD student
Department of Geography
McGill University

Alan Chan
PhD student
Mila, Université de Montréal

Fernando Diaz
Canada CIFAR AI Chair
Mila, McGill University

Andrés Ferraro
Postdoctoral Research Fellow
School of Computer Science
Mila, McGill University

Alex Ketchum
Faculty Lecturer of the Institute for Gender, Sexuality, and Feminist Studies
McGill University

Fenwick McKelvey
Associate Professor
Communication Studies
Concordia University

Jimin Rhim
Postdoctoral Research Fellow
Electrical and Computer Engineering
McGill University

Shalaleh Rismani
PhD Candidate, Electrical and Computer Engineering
McGill University

Renée Sieber
Associate Professor
Department of Geography
McGill University

Jonathan Sterne
James McGill Professor of Culture and Technology
McGill University

Yuan Stevens
Tech Policy & Legal Researcher and LL.M Candidate
University of Ottawa, Faculty of Law