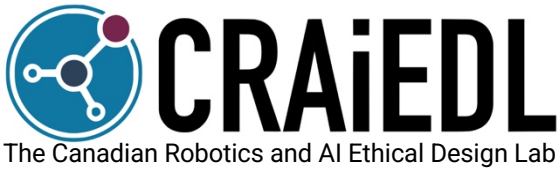
Peer Review for Automated Decision-Making Tools Under Canada's Directive on Automated Decision-Making

A Model Peer Review Process and Recommendations for Best Practices

October 2020

Dr. Kelly Bronson
Dr. Jason Millar



Contents

Executive Summary

Advanced algorithms (i.e. artificial intelligence, or AI) are increasingly being used to support and improve the efficiency, consistency and transparency of decision-making within governments, and to enable whole new categories of public decision-making. Used in these public-sector applications, algorithms raise a host of social and ethical concerns. Although most automated decision systems are largely in the conceptual, exploratory, or early pilot phases, Government of Canada (GOC) has seized the opportunity to be pro-active in critically assessing these technologies and implementing a policy response through its Directive on Automated Decision-Making (the Directive). The Directive sets out rules for how federal departments and agencies must develop and implement algorithms to inform (or make) service decisions.

The Directive specifies a peer review mechanism intended as an additional check to ensure that any risks associated with new algorithmic decision-making tools are properly identified and mitigated. To date, GOC has not received external guidance on what would constitute best practices for conducting such peer review.

This report details the results of a collaborative study between the authors, Treasury Board Secretariat and Canadian School of Public Service, and is aimed at providing additional guidance on how to conduct robust peer review under the Directive. Ultimately, the recommendations and a practical toolkit arising from this work will enable and promote the continued responsible development and implementation of AI within GOC.

Using an evidence-based approach, the report delivers a toolkit consisting of:

- 14 Key Recommendations for conducting peer review;
- A Model Peer Review Process;
- A Model Evaluation Grid for use in procurement of automated decision tools, completion of the online algorithmic impact assessment (AIA), and in peer review; and
- a Supporting Documentation Framework for use in completing AIAs and peer review processes.

The toolkit is designed to be practical and useful for:

- Policy officials to realize and even accelerate their goals for responsible AI in GOC;
- Algorithm developers (either external or internal) to better understand the core expectations of GOC and bring their projects into alignment with emerging standards;
- Policy officials overseeing the AIA process from online impact scoring to peer review; and
- Peer reviewers in their task of evaluating both the social and ethical risks of algorithms, and the various measures that were put in place during the development of the algorithm to mitigate those risks.

Moreover, the toolkit is designed to accomplish several specific goals common to analogous peer reviewing activities in other domains (e.g. Research Ethics Board reviews), including:

- Clarifying accountability in peer review (e.g. roles and responsibilities);
- Establishing consistent peer review practices (e.g. defining standard documents for assessment, identifying key decision points); and
- Establishing transparency in peer review to assist in procurement (i.e. with vendors), development, and engagement with external stakeholders or affected communities.

Introduction

Artificial intelligence is increasingly being used to drive decision-making within governments in order to increase the efficiency, consistency and transparency of decisions, and its implementation potentially raises a host of social and ethical issues. For example, the big datasets which "feed" into AI can be biased and lead to the reproduction of harms against historically marginalized social groups. As well, the implementation of AI can lead to the replacement of human labourers or what the World Economic Forum calls "technological unemployment."

To get ahead of the social and ethical challenges presented by the use of AI within the Government of Canada (GOC), Treasury Board Secretariat (TBS) is building management frameworks and policy instruments to assess and mitigate risks associated with the use of AI in government. TBS has published a Directive on Automated Decision-Making (the Directive) that sets out rules for how federal departments and agencies may use automation to inform service decisions. Under the Directive, any automated decision-tool that will be used in government to replace a human in making an administrative decision about a client must undergo a rigorous Algorithmic Impact Assessment (AIA), which results in a calculated impact level between 1 and 4, from least to greatest impact. Any automated decision tool that presents an impact level 2 or higher according to the AIA, is subject to a "peer review" process proportional to the risk, among other requirements. The application of risk mitigation strategies and peer review for automated decision-making tools are in their early days of being put into practice by government departments.

One GOC department, Canada School of Public Service (CSPS or the School) is an "early adopter" of an automated decision-making tool—the Regulatory Evaluation Platform (the REP)—which is intended to support federal regulators. According to the AIA, the REP does not raise significant risks (the AIA scored it as impact level 1). However, CSPS approached the authors to perform a "mock peer review" in response to the requirement from the Directive, in order to formally investigate the peer review process in more detail in anticipation of future automated decision-making tools with impact levels of 2+, and to establish best practices and recommendations for others tasked with peer reviewing automated decision-making tools for GOC. CSPS considers this mock peer review as an important next step toward ensuring ethical and responsible development and use of automated decision-making tools within CSPS and GOC more broadly. Additionally, CSPS "as a learning institution" is interested in creating learning products that build capacity across GOC for complying with the Directive. Consequently, the results of this project will further that mandate by creating a learning product for peer review of automated decision-making systems.

This report incorporates the findings of the mock peer review of the REP. The authors used an evidence-based participatory design approach to develop a model peer review process and recommended best practices for planning and performing peer review of automated decision-making under the Directive. The recommendations from this report are intended to enable and promote the continued responsible development and use of AI in GOC.

Methodology

Our research unfolded in two stages:

1) Environmental Scans of existing policies, practices, and literature on peer reviewing activities;
2) Participatory design of model peer review process and supporting documents.

Initially, our research plan focused entirely on using a systematic review of literature to search for existing best practices for peer review in the context of algorithmic assessments. However, once we completed an initial scan, we found only one report outlining existing policies and for peer review of algorithms or AI. GOC's activities on this front appear to be among the first.

With almost no guiding literature specific to peer review of algorithmic decision-making, we decided to scan for analogous peer review policies, practices, and literature specific to social and ethical impact assessments more broadly. Several examples emerged as candidate models for designing a peer review process specific to algorithmic assessment. We searched grey and academic literatures (see detail below) pertaining to these exemplary peer review practices. We were guided in our search of the literature by the following research questions:

a) How is meaningful peer review conducted according to existing policies, practices and literature on analogous ethics reviews and impact assessments?
b) What can we learn and adopt from best practices in responsible approaches to peer review, specifically in the context of ethics reviews and technology assessments?
c) What ethical principles and evaluation metrics are most appropriate to ensure responsible peer review processes are implemented for automated decision-making at GOC under the Directive?

Stage 1: Environmental Scans of Existing Policies, Practices, and Literature on Peer Reviewing Activities

Given the novelty of this topic—peer review within the context of algorithmic (i.e. automated) decision-making—we first used an "environmental scan" of grey, legal and policy literature for any instances of peer review processes related to algorithms using the search terms "peer review" AND "automated decision-making"; AND "Algorithm"; AND "Technology." This search yielded 42 results only six of which were deemed directly relevant to this project.

At the same time, we performed a document analysis of grey literature on existing processes analogous to peer review for automated decision-making, charting how (and when) peer review is conducted. The analogous processes we analyzed were: 1) privacy impact assessments; 2) environmental impact assessments; 3) research ethics board reviews (sp. the "scholarly review" component therein).

Third, we used these analyses to narrow in on a set of precise search terms for a formal review of the academic literature. We reviewed two commonly used databases, Web of Science and Scopus and also two databases that compile legal case scholarship and secondary legal publications, Westlaw and LexisNexis. We used the following search terms: "Peer review" AND "privacy impact assessment"; AND "Impact assessment"; AND "Impact evaluation" AND "Impact N/3 evaluation";

"Research ethics board"; AND "Institutional Research board." Secondarily, we tried the same search but using synonyms for peer review that we derived from our early review of papers: "Peer evaluation" OR "Peer assessment" OR "Scholarly review" OR "Trusted external researchers." All of this yielded 195 articles, 31 of which were deemed directly relevant to this report.

Consultation was a key component of our method and was used to inform and validate findings. We had consultations with an international expert on AIA who is beginning work on peer review for automated decision-making (Berkman Klein Centre for Internet and Society, Harvard University); an expert on scientific peer review in the context of environmental impact assessment from the Impact Assessment Agency of Canada (Government of Canada), and an academic expert on inclusive innovation. We also engaged in several conversations with the GOC project sponsors/leads, who have been involved in the development of both the Directive, and the REP, and were able to provide detailed feedback during both research activities. Finally, both of the principal researchers on this project have considerable expertise having served on institutional Research Ethics Boards in academic and healthcare contexts, as well as the private sector. These consultations pointed us toward relevant case examples and helped us ensure that the data we surveyed was timely and grounded in current policy practice. Last, on July 29, 2020 we used a virtual participatory design workshop (methodology described below), which included a diverse set of key stakeholders in gathering critical input in the model process design. These various consultations are consistent with a scoping review method where scoping reviews (Arksey & O'Malley, 2005; Levac et al., 2010) are used to identify strengths and current gaps in research, and policy, by consulting with domain experts to inform and validate systematic literature reviews.

Stage 2: Participatory Design of Model Peer Review Process and Supporting Documents

Having identified and gathered the relevant evidence from analogous peer reviewing activities, we performed an *initial* process mapping exercise to identify key principles, steps, decision points, and other practices common among those peer reviewing activities. Based on the outcome of the mapping exercise, we drafted a *prototype* peer review process for automated decision-making. Because the peer reviewing activities we used as evidence for our prototype were not specific to automated decision-making tools, we also identified a set of key questions and gaps in our prototype which we believed, once addressed, would add the specificity required of a robust peer review process for automated decision-making per the Directive.

To refine and validate the prototype peer review process, we ran a participatory design workshop, drawing from additional expertise throughout GOC. GOC project leads identified and recruited workshop participants from various GOC departments; All participants had a working knowledge of the Directive, and some had additional experience conducting peer review of automated decision-making tools. The participatory design workshop was run virtually via Microsoft Teams wherein researchers facilitated the discussion to prompt critical feedback of the prototype peer review process.

Following the workshop, two research assistants independently identified common themes from the entire set of workshop notes. These findings were used to generate a final process map and a set of recommendations (see Appendices A through C and Key Recommendations section below).

Key Findings

Stage 1: Environmental Scans

Existing Policies, Practices, and Literature on Peer Review of Algorithms

The use of peer review in the context of algorithmic impact assessment is an incredibly novel domain in practice and in the literature. Indeed, our environmental scan of grey and policy documents revealed that *no other peer review processes specific to algorithmic decision-making have been put into practice by governments elsewhere*. However, there is indication that peer review of AI will soon be used in the US, New Zealand and across EU member states.

In 2019, the US launched Bill S.1108, or the Algorithmic Accountability Act to direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments. Section 3(b)(1)(c) under Data Protection Authority highlights the need to, if reasonably possible, include in impact assessments what the report refers to as "external third parties and independent technology experts." One key finding from a review of this legislation is that experts ought to specifically use "a detailed description of the automated system, its design, training data and purposes," to review these for the relative benefits and costs and benefits of the automated system (see Subparagraphs (A) and (B)).

The New York-based *AI NOW Institute* released a report in 2018, presumably in anticipation of Bill S.1108, titled, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." The report suggests that government's use of automated decision systems should provide a meaningful and ongoing opportunity for "external researchers" to review, audit, and assess these systems. Similar to Section 6.2.5.3 of the ADM Directive from the GOC, a "trade-secret claim" is not grounds for obstruction to peer review since using a proprietary license still allows for external parties to review and audit these components, as necessary. The report further states that if a vendor objects to meaningful external review, it would signal a *de facto* conflict between the vendor's system and public accountability (p.14). The report additionally recommends an internal agency self-assessment and a necessary increase in the capacity of public agencies to assess the fairness, justice, due process, and disparate impacts from AI. It states that, ideally, government agencies must have expertise on those automated decision systems applied within government, and they ought to provide both detailed reports to scaffold expert peer review and non-technical summaries for the public to participate in robust public engagement around AI in government.

In 2019, the European Parliamentary Research Service released *A Governance Framework for Algorithmic Accountability and Transparency*, which includes peer review of algorithms by "external or outside researchers" or by "trusted external experts" as a means to "achieve public accountability." Similar to the US legislation described above, the EU framework would grant "qualified researchers" "meaningful access" to the technical details of any algorithm or AI – notably to training data or records of past decisions made using the AI. Slightly different from the US Bill, the proposed EU framework lays out the goal of the peer review as monitoring any system to explicitly interrogate its potential risks (rather than risks and benefits) and specifically whether the AI raises possible harms to "the public interest." This framework appears to offer a helpful recommendation regarding how to involve community stakeholders in a feedback process that structures the peer review, and to do so in a way that is ongoing and transparent. For example, under this framework affected community

members are empowered to suggest researcher reviewers and to work with these researchers to publish peer review findings openly.

Like the EU, the government of New Zealand (Stats New Zealand and Internal Affairs) released a report on the ethical use of AI within government. The report recommends connecting expertise for the ethical assessment of AI across government, and also recommends "looking beyond government for privacy, ethics, and data expertise" in an external review of any new AI or any AI that is being applied to new use cases. Like the other jurisdictions, the government of New Zealand seems to be leaning toward a peer review process conducted by trusted external experts such as academics, and ones with a range of expertise spanning computer science, data analytics, law, sociology, and ethics.

One best practice has recently emerged from a non-governmental context. In late 2019, the United Nations Office for the Coordination of Humanitarian Affairs released a consultant-prepared report on the use of AI for humanitarian efforts. This report provides what appears to be the only published guidance regarding the use of peer review for algorithmic assessments; specifically, the report includes a valuable *evaluation grid* that peer review committees can use to help assess AI. The evaluation grid is itself derived from Cathy O'Neil's matrix for ethical evaluation of AI (see https://orcaarisk.com).

Analogous Existing Canadian Impact Assessment Peer Review Policies, Practices, and Literature

In the absence of case examples of best practices for peer review of AI, we undertook an analysis of "neighbouring" impact assessments, which are designed specifically to inform government decision-making and which aim to identify potential risks and social implications of technologies or related policies or programs: Ethics reviews for research involving human subjects; Government of Canada Environmental and Social Impact Assessments; and Privacy Impact Assessments.

Research Ethics Boards (Tri-Council Agencies, Canada)

University research ethics boards (REBs), responsible for approving any research involving living human participants and human biological materials (including reproductive materials and stem cells), serve as a useful model of risk assessment. In a typical REB process, the initial REB application is prepared and submitted by the principal investigator before any research can begin. The initial application to an REB triggers a procedure similar to the AIA's Peer Review after the online AIA assessment.

In Canada, each research institution is expected to have its own standing peer review committee which operates under the clear policy standards laid out in the Tri-Council Policy Statement-2 (2018). The REB membership typically consists of at least 5 members, ideally with diversity from each of the underrepresented groups (women, visible minorities, Indigenous, persons with disabilities, LGBTQ+), and representation from a diversity of relevant research disciplines (e.g. ethics, law) as well as one member of the community with no affiliation to the institution. REB committees do additionally consult domain experts outside of the university on a needs basis, but when they do so the policy states that they must be cognisant of "undue influence or elements of power imbalance that would affect the REB review." Even REB members are required to disclose real, potential, or perceived conflicts of interest, and REB members sometimes recuse themselves from deliberations and decisions. Typically, REB members are not compensated for their work.

The REB committee uses the principal researcher's application to the REB to scaffold its review and ensure that it is proportional to the anticipated risk; application materials typically consist of descriptions of the nature of the research, and the researcher's own estimation of potential adverse

effects from the work. Proportional to the perceived risk of the research, REBs may require that the researchers provide more fulsome documentation about the research to support the peer review.

Additionally, the TCPS-2, among other guiding documents, puts forward a number of quality control mechanisms for peer assessment of research. One notable mechanism is the use of evaluation grids by assessors, which have been demonstrated to improve the quality and objectivity of peer assessments (see the San Francisco Declaration on Research Assessment, https://sfdora.org). Grids help make evaluation criteria explicit and transparent in that reviewers are forced to detail and explicate their judgements and justifications for decision-making by following the grid (often a table). Grids also help implement evaluation criteria to ensure that reviewers are not selective (biased) toward any particular assessment criterion. Grids have also been proven to help ensure accountability among reviewers, again mitigating against prioritizing any particular evaluation criterion by pushing reviewers to consider all relevant aspects of the algorithm's consequences. Last, in the same way evaluation grids help principal researchers adhere to TCPS-2, they could help in algorithmic assessments even as early as the procurement stage: developers could take steps toward ethical self-governance by following those evaluation criteria to which they may, down the road, be subjected. Another notable quality control mechanism used in research peer review is training modules for reviewers. The Tri-council agencies have a series of tools to guide awareness among peer reviewers about equity, diversity and inclusion, including an important training module for reviewers and/or those guiding the review process (https://www.chairs-chaires.gc.ca/program-programme/equity-equite/bias/module-eng.aspx?pedisable=true).

Privacy Impact Assessment (Office of the Privacy Commissioner of Canada)

The Office of the Privacy Commissioner of Canada recently created a procedure for federal institutions within the Canadian government to review prospective programs or initiatives for compliance with the *Privacy Act,* through a structured Privacy Impact Assessment (PIA), which came into effect on March 13, 2020.

Much like the REB, the first step in the assessment is to be completed internally by the team responsible for the proposed initiative and is a necessary preliminary step to instituting programs that may exploit the personal information of Canadians. Like the REB and the AIA, the PIA requires an assessment and assignment of risk or impact level, and directs specific reviews based on which provisions of the *Privacy Act* are engaged by the government program.

The OPC's PIA framework has thus far been applied to an algorithmic decision-making mechanism by the Canadian Border Services Agency (CBSA) to evaluate the Scenario-Based Targeting of Travellers. This algorithm assigns a level of risk to incoming travellers to Canada based on information contained in their passports.

While some insights can be drawn from the PIA framework, specifically around when peer review occurs, it is also important to note that the PIA recognizes that each instance of review must be case-specific; while a generic framework for compliance with the *Privacy Act* has been developed, the specific privacy considerations must be identified internally then reviewed by a consistent team of public servants who develop an expertise in PIA as a process.

Environmental/Social Impact Assessment (Impact Assessment Agency of Canada)

The Impact Assessment Agency of Canada (IAAC) requires impact assessments (IA) to be carried out for various types of proposed development projects in Canada. New legislation (IAA 2019) requires any assessment to consider not just environmental impacts but also a wider set of social and cultural impacts and it calls for "meaningful participation" of members of the public (and specifically Indigenous Canadians) in this process. IAAC has recently implemented guidelines for what they call

an expert technical review (ETR) of impact assessment submissions and this is analogous to peer review of AI.

To support this review, IAAC has put forward Science Advice for Government Effectiveness (SAGE) principles which were created by their Council of Science and Technology Advisors (CSTA). One notable principle is that of inclusiveness, operationalized in this context by stating "advice should be drawn from a variety of scientific sources, from experts in relevant disciplines and external and international sources. Due weight also needs to be given to the 'traditional knowledge' of local peoples." SAGE indicates that one best practice is to involve a wide range of scientific expertise from relevant disciplines inside and outside of government in the ETR. Moreover, like the recommendations from *AI Now*, SAGE recommends empowering affected community members to review, challenge, and contribute to the scientific information discussed. SAGE recommends that IAAC be responsible for selecting "independent" experts to populate the expert technical review panel with between 2-4 people, and also that one of these recruits be designated as Chair. Unlike REB committee members, ETR review panelists are paid an honorarium and their travel expenses are reimbursed.

SAGE recommends that IAAC, in consultation with proponents, relevant federal departments, relevant regulators, Indigenous groups, and others, develop a list of questions for the expert technical review panel to address and answer. In other words, wide participation is sought to scope the peer review. Like the REB committee, SAGE recommends mitigating conflict of interest when external stakeholders participate in the ETR by requiring participants to declare any vested interests in the outcome *a priori* to the review processes.

Based on the principle of transparency, elements of the full technical and other documentation submitted in the context of an impact assessment can be accessed by the review committee who can vet: the quality of evidence collection methods and procedures; the reasonableness of predicted impacts and judgements made from the available evidence; and the level of risk and/or the degree of uncertainty around this risk. Results from the ETR and review committee meeting summaries are released to the public. Differences of opinion among the experts are also to be included in the summaries. The peer review committee submits a final report to IAAC who uses this to inform a final decision about the development project.

Stage 2: Peer Review Process and Supporting Documentation

The data collected during the participatory design workshop yielded 28 unique participant feedback statements (the findings) which were translated into[1]: 14 Key Recommendations (see the Key Recommendations section); a finalized Model Peer Review Process (Appendix A); a Model Evaluation Grid for use in procurement of automated decision tools, completion of the AIA and in peer review (Appendix B); and a Supporting Documentation Framework for use in AIAs and peer review processes (Appendix C).

Each of these tools is designed to be practical for the following stakeholders: project and program managers both in GOC and the private sector (e.g. clients and vendors); policymakers developing

[1] Not all participant feedback was incorporated into these final documents as some was specific to particular use cases and could therefore not be generalized across GOC over time, or across all peer review activities.

guidance on peer review for automated decision-making tools in the public sector, or overseeing peer review; and peer reviewers charged with reviewing automated decision-making tools, who can use the process as they work through review activities. The tools are also designed to accomplish several goals that we found are common in analogous peer reviewing activities, in particular: defining guiding principles for peer review; clarifying accountability in peer review (e.g. clarifying roles and responsibilities); establishing consistent peer review practices (e.g. defining standard documents, identifying key decision points); and establishing transparency in peer review such that, for example, vendors and project managers are better able to anticipate the need for, and potential impacts of, peer review as part of their regular planning and procurement processes.

Distinct Peer Review Process Steps

Each step in the model peer review process (see Appendix A) is designed to focus peer reviewers' attention on four aspects of the peer review process, including:

1. Who is responsible for completing that step of the process;
2. What function each step is meant to accomplish within the overall peer review process;
3. Which of the guiding principles applies specifically to each step; and
4. What key questions should be asked in order to accomplish that step.

Guiding Principles for Peer Review

The eleven guiding principles we initially identified were derived from our scoping review. The guiding principles, in no particular order, include: Vigilance; Independence; Accountability; Transparency; Proportionality; Accuracy; Freedom from Bias; Consistency; Inclusion; Robustness; and Legibility. Each guiding principle is paired with a short description in the model peer review process to provide some explanation for, and meaning behind, its inclusion and importance in the peer review process (see Appendix A). These short descriptions are intended to aid peer reviewers in understanding how to apply each principle throughout the process and to aid policymakers in justifying the implementation of suggested steps. The principles themselves are intended to help frame some of the key ethical issues associated with each step in the peer review process. For example, Accountability is described as follows:

> "Clearly defined roles and responsibilities are essential for accountability. They should be established and consistently applied throughout the Peer Review process."

Similarly, Independence is described as follows:

> "Peer Review should proceed independent of undue influence. Peer Reviewers should NOT report to any party who is in an actual or perceived conflict of interest (COI) with respect to the development and/or implementation of the algorithm."

Evaluation Grid

The evaluation grid (Appendix B) is designed for use primarily by peer reviewers though it could support conversations with external clients/vendors at the procurement stage and throughout the development lifecycle. The grid was developed from our scoping review findings, and specifically a synthesis of the existing best practices from analogous domains—notably, the academic research ethics board, or REB. In the REB context, evaluation grids or matrices are used to review the risks arising from research with hazardous materials or human subjects. These forms typically consist of checklists for reporting on the technical details of the research and its likely consequences. Such checklists have been used to assess AI-related research (*e.g. Guidelines for Developing and Reporting*

Machine Learning Predictive Models in Biomedical Research), however their use is longer-standing and much broader.

Supporting Documentation Framework for AIAs and Peer Review

A key gap we identified early in this research is the need for a standard set of detailed documentation to support both the completion of the AIA and the peer review process. The collection of such documentation should begin even before the completion of the first AIA in anticipation of the need for peer review in cases where the AIA deems the automated decision-making tool of impact level 2 or higher under the Directive.

Existing peer review processes – again, notably REBs – tend to be supported by a consistent set of documents that scopes the review process and brings reviewers from different disciplinary and other contexts up to speed on the details of the project under review. The creation of, and responsibilities related to, those documents are typically clearly established in related policies. Our initial scoping review and early participatory workshop revealed that, although the Directive sets out a number of requirements regarding the need for AIAs and peer review, there are currently no precisely defined guidelines addressing what kinds of documentation might be needed to support either activity, nor what kinds of accountability practices should be adopted regarding such documents. Because robust peer review starts with the AIA, we developed a documentation framework (see Appendix C) that moves in a stepwise fashion through the GOC AIA, suggesting data and information that could be collected at the stage of initial AIA completion but that also supports robust peer review activities.

Key Recommendations

We offer the following recommendations for developing and implementing robust peer review under Canada's Directive on Automated Decision-Making. These recommendations are meant to support the development of consistent, efficient, effective, and ethical peer review. Many of these recommendations add clarity and precision to elements of the Directive and recommend practices and actions that meet or exceed the requirements in the Directive.

Recommendation 1. Work to implement and strengthen existing guidance on completion of the Algorithmic Impact Assessment (AIA), and to develop additional guidance specifying how and to whom the results of an AIA must be communicated, and how AIA documentation is to be stored for future reference (e.g. during peer review).

a. A team of at least two people should independently complete an AIA at multiple stages in the project lifecycle, ideally at the conceptual and implementation stages. These people should have different expertise.
b. Those completing the AIA should be, or become, familiar with the details of both the plans for the automated decision-making tool and the program in which it will be applied. They should understand the tool's objectives, the action that the output will inform, and the application setting.
c. When possible, a preliminary AIA should be completed prior to issuing a request for proposal, and the impact level should be indicated in the statement of work.
d. In the case where there is discrepancy in scoring between people completing the AIA, use a consensus-based approach to determine the impact level.
e. Legal services should be consulted for each AIA that is completed.

Recommendation 2. Develop a standard set of documentation to support the peer review processes.

a. At minimum, this documentation should provide descriptions, details, and supporting materials that flesh out explanations for each of the answers provided in the accompanying AIA.
b. Documentation should be legible. "Legibility" is an essential principle for ethical automated decision-making and refers to the understandability of the tool, and the ability of peer reviewers, some of whom may have limited technical capacity, to understand the mechanisms and outcomes of algorithms or models.
c. Ideally, documentation will include development code, methods or purpose documentation and sample training data. In the case of external clients/vendors, they may request this documentation be kept confidential.
d. This documentation should be archived per Recommendation 1 for reference during any subsequent peer review or audit.

Recommendation 3. Following from above, develop clear procurement processes to ensure all potential AIA and peer reviewing activities are anticipated in the project plan and contracts prior to the start of development.

a. Develop consistent policies and procedures for including automated decision-making tool developers (e.g. data scientists, vendors) in the preparation for, and completion of, AIAs and supporting documentation.
b. Vendors or developers with demonstrated expertise in the ethics and governance of AI should be given preference.

Recommendation 4. Develop a clear peer review process to guide consistent peer review.

Recommendation 5. Develop clear accountability policies for peer review establishing, among other details, who is responsible for the proper oversight of peer review.
a. The person responsible for completing the AIA should be responsible for collecting all the standard documentation (see Recommendation 2) to support the AIA and peer review process.
b. The person responsible for completing the AIA ought to act as the "moderator" throughout the peer review process. The moderator would make no substantive contribution to discussions but act in an administrative and facilitative capacity: sending invitations to potential reviewers based on their domain expertise or stakeholder perspective; finalizing the committee 7 days from the time it is decided a peer review is required. The moderator would also provide the reviewers with the evaluation grid, the results of the AIA and supporting documentation.

Recommendation 6. Once a peer review committee process has begun, the committee itself should decide on the frequency of meetings. Timelines for the peer review processes will vary depending on the tool, context of application, and impact level.

Recommendation 7. Clear mechanisms should be adopted to increase equity, diversity, and inclusion in the peer review process.
a. Peer review committee members could use an evaluation grid for their independent assessment of algorithms. Completed evaluation grids ought to be submitted to the moderator in advance of peer review committee meetings. The moderator should then facilitate a discussion by focusing the group on areas of discrepancy.
b. Provide peer review committee training on issues relating to equity, diversity and inclusion.
c. The peer review committee should operate according to a consensus model of dialogue and decision-making with attention to implicit bias, and equity, diversity and inclusion issues which may arise in such settings.

Recommendation 8. Peer Review should involve a minimum of two peer reviewers at all impact levels.
a. Whenever possible, the peer review committee should be interdisciplinary, drawing on expertise specific to the nature of the automated decision-making tool as well as expertise with social and legal assessment.
b. When identifying peer reviewers, consider soliciting recommendations from other GOC departments and/or external academic partners.

Recommendation 9. For automated decision-making tools with impact level 4, stakeholders from impacted groups in the public/community should be included in the peer review process.

a. GOC should consider compensating members of the public/community for their participation in peer review.

Recommendation 10. Strive to create a standing pool of peer reviewers to build consistency in the review process and to build continuous institutional knowledge with respect to the peer review process.

a. Consider having these individuals perform peer review across departments, units or divisions thus helping to limit conflict of interest.
b. Consider ongoing training in ethical and social issues related to algorithmic decision-making to build internal ethics capacity among peer reviewers.

Recommendation 11. Develop clear Conflict of Interest (COI) guidelines for selecting Peer Reviewers.

a. A key consideration includes whether a COI (apparent or actual) exists due to remuneration for an external member of the peer review committee.

Recommendation 12. Develop a standard case study-based training program for peer reviewers.

Recommendation 13. Develop clear guidelines for the implementation of outcomes from peer reviews.

a. Results of the peer review should be considered in the decision to allow the system to go into production. Risks not identified or mitigated during the initial AIA and development should weigh heavily in the final decision.
b. Further develop guidance on: who is responsible for ensuring the implementation of peer review recommendations; timelines proportional to risk for completing their implementation (e.g. before the project goes into production; within a six-month window); and reporting and documentation mechanisms for "signing off" on their implementation.

Recommendation 14. Develop clear reporting requirements for peer review to clarify what is to be the output of the peer review process, and thus what "results" of the review ought to be reported upon publicly as per the Directive regarding tools with impact level 3 or 4.

Appendix A: Model Peer Review Process for Automated Decision-Making

Introduction

This document is intended to sketch a high-level process for conducting peer review of an automated decision-making tool within the government of Canada, beginning with the initial completion of the final algorithmic impact assessment (AIA). This process is evidence-informed and is intended to be refined through regular practice and use.

Guiding Principles for Peer Review

- **Vigilance:** The Peer Review process is the result of an impact level deemed significant enough for additional scrutiny of the automated decision-making tool ("the algorithm" hereafter). The AIA is expected to be updated whenever there is a change in the system or algorithm and therefore may affect the Peer Review. As such, Peer Review may be warranted any time there is a change to the algorithm, even in cases where the AIA-indicated impact level remains unchanged or decreases, as changes to the algorithm can introduce new unexamined risks.
- **Independence**: Peer Review should proceed independent of undue influence. Peer Reviewers should NOT report to any party who is in an actual or perceived conflict of interest (COI) with respect to the development and/or implementation of the algorithm. Similarly, peer reviewers should not be in actual or perceived COI in their capacity as reviewers.
- **Accountability**: Clearly defined roles and responsibilities are essential for accountability. They should be established and consistently applied throughout the Peer Review process.
- **Transparency**: Peer Review should be transparent, both in its process and outcomes. Consistent and complete documentation processes should be established to clearly communicate the key decision points, the reasoning behind those decisions, and accountability for those decisions.
- **Proportionality**: In keeping with the tiered impact levels described in Canada's Directive on Automated Decision-Making, the level of scrutiny applied throughout the Peer Review process should be reasonably proportional to the nature of the underlying risks associated with use of the algorithm. For example, higher impact levels could warrant additional peer reviewers or more detailed supporting documentation to support the peer review.
- **Accuracy**: Peer Review should be based on an application of the relevant program/technical/domain expertise as determined by the technical details of the algorithm design and the application for which it is intended.
- **Freedom from Bias**: Care should be taken to ensure Peer Review is conducted by individuals who are free from bias in their ability to assess the algorithm. As such, Peer

Reviewers should be selected who are not members of the team(s) directly responsible for the development and/or implementation of the algorithm.

- **Consistency**: In order to develop institutional knowledge and best practices for Peer Review at the Government of Canada, Peer Reviewers should be chosen in part, and to the extent practically feasible, based on their ability to serve in that role on multiple Peer Reviewing activities over a period of time.
- **Inclusion:** The AIA Peer review process should include a wide and diverse range of perspectives to ensure as many concerns about, and impacts of, automated decision-making systems can be raised and addressed. Representation of communities affected by automated decision systems provides an opportunity for challenge by members of the public and enhances accountability and transparency.
- **Robustness**: The Peer Review process should be adaptable to automated decision-making tools intended for any application and of any AIA-indicated impact level that would warrant Peer Review (Impact Levels 2-4).
- **Legibility**: Transparency is concerned with making information about the automated decision tool (e.g. data) available to peer review scrutiny, verification, whereas legibility relates to how easily AI systems and their decisions can be understood by non-AI experts. In the case where a peer review involves stakeholder groups or affected community members, legibility of documentation is of crucial importance.

The Peer Review Process

Step 0: Complete the AIA and arrive at impact level

Responsibility: Depending on the assumed impact level of the algorithm this will vary, but will likely include more than one person from inside the department and specifically the design or management team for the algorithm.

Description: This first step concerns completing the AIA that will a) determine both whether a peer review is necessary (by indicating an impact level) and b) scaffold a peer review process. However, any changes to the tool or its training data, including changes that decrease the impact level or result in no change to the impact level, could warrant revisiting this step.

Key Guiding Principle(s): Vigilance; Diversity; Inclusion

Key Questions:

1. Is the impact level 2 or above?

If "yes" then a peer review is warranted.

Step 1: Populate a peer review committee

Responsibility: The people who complete the AIA will be responsible for this step. The "moderator" from GOC will ensure that the diversity and number of reviewers are proportional to the impact level. They will contact potential reviewers and attempt to populate the committee within 7 business days.

Description: This step is concerned with the assignment of the peer reviewers who will first be responsible for an objective, secondary assessment of the AIA and supporting documentation required to interpret it. Reviewers will adjudicate its accuracy and integrity.

Key Guiding Principle(s): Accuracy; Freedom from Bias; Inclusion; Proportionality; Accountability; Consistency

Key Questions:

1. Is the composition of the peer reviewers proportional to the impact level?
    a. Are the reviewers appropriately diverse according to domain or disciplinary expertise as well as gender, racial and ethnic diversity? OR Have reviewers been deemed capable of speaking appropriately to any relevant issues of equity and diversity?
    b. Have members of potentially affected populations been included where impact level (3,4) warrant such inclusion?
2. Do any of the reviewers have a conflict of interest or a potential perceived conflict of interest?
3. In the case that stakeholders, community members or end users are included among peer reviewers, are they being remunerated appropriately?

Step 2: Gather supporting documentation

Responsibility: The people who complete the AIA will be responsible for this step.

Description: Key supporting documents must be archived and maintained by the individual(s) who completed the AIA. These are necessary supports for a robust and independent peer review. Establishing which documentation is sufficient in each case will depend on the prior expertise of the peer reviewers. At minimum each peer review, regardless of the impact level, ought to be supported by a standard set of documents, including the AIA (see Appendix C report).

Key Guiding Principle(s): Proportionality; Transparency; Legibility

Key Questions:

1. Is the minimum set of documents available?
2. Are there additional requested documents required to complete a proportional peer review given the tool and/or the expertise of the peer review committee?

    If "yes" to (2) please gather necessary documentation before proceeding to Step 3.

Step 3: Validate the AIA

Responsibility: Peer Reviewers.

Description: The AIA Impact Scoring Algorithm (ISA) generates a predicted impact level of the automated decision-making tool; however, the ISA is imperfect and may not capture the actual impact. Additional scrutiny should be applied to the AIA and ISA output to validate both. [NOTE: this step might also uncover deficiencies in the supporting documentation in cases where the peer reviewers find that the AIA indicates a higher impact level than the supporting documentation would suggest].

Key Guiding Principle(s): Accuracy; Proportionality; Consistency; Robustness

Key Questions:

1. Based on a quick analysis of the available documentation (including the AIA) was the AIA completed accurately?
2. Has the AIA impact level for the algorithm changed (increased or decreased) based upon changes in the tool, the training data, legislation or application environment?
    a. Are there any obvious discrepancies between the AIA and the Detailed Functional Description and other requested documents that would seem to warrant a different impact level score (either higher or lower)?
3. If "yes" to (1) or (2), then the discrepancies may require the parties responsible for Step 0 to update the AIA before proceeding with a full Peer Review.

Key Decision Point: Peer reviewers must establish the impact level that will be used for subsequent Peer Review activities (Steps 4+) prior to proceeding.

Step 4: Review the algorithm using the AIA, supporting documentation, and Evaluation Grid

Responsibility: Peer reviewers.

Description: The peer review committee will refer to the supporting documentation, complete the Evaluation Grid and submit it to the Moderator. The Moderator will compare results across the committee and facilitate a discussion based upon any discrepancies among reviewers. If the reviewers believe that additional information or documentation is required to complete the peer review, the Moderator will assist in obtaining these additional resources. This step may require additional input from individuals responsible for the development and implementation of the tool (e.g. vendors). Algorithm developers and GOC project managers should anticipate and accommodate peer reviewers' requests for input during a review. This possibility should be accounted for in the project plan to avoid undue delays during the review. The peer review ought to proceed in a consensus-based approach and the deliberation structured by following the AIA and referring to the evaluation grid.

Key Guiding Principle(s): Proportionality; Accuracy; Freedom from Bias; Independence; Legibility

Key Questions:

1. Is the quality of information, collection methods and procedures sufficient to support an accurate and comprehensive proportional review of the algorithm?
2. According to the expertise of the reviewers, are there risks which can be anticipated which were not covered by the AIA?
3. Does the available information gathered to complete the peer review confirm the judgments that have been made about the algorithm's predicted impacts?
4. Have uncertainties around impact levels, and near- and long-term risks, been adequately accounted for in the development of the algorithm?

Step 5: Assess the adequacy of risk mitigation and ongoing risk monitoring strategies

Responsibility: Peer reviewers.

Description: Having examined the risks raised by the algorithm, the details of the Directive that pertain to the algorithm will be reviewed for any gaps in risk assessment, mitigation and governance. Note: the determination of "adequate/sufficient risk mitigation" rests in the expertise of the peer

reviewers. Peer review processes are necessarily subjective assessments which underscores the importance of Steps 1 and 2: choosing a review committee and gathering robust documentation to support thorough deliberation.

Key Guiding Principle(s): Accuracy; Freedom from Bias; Independence

Key Questions:

1. Are the predicted impacts and judgements made from the available evidence reasonable?
2. Is there an acceptable level of uncertainty around long term risks?
3. Considering responses to (1) and (2), what is the sufficiency and reasonableness of the proposed risk mitigation and ongoing monitoring strategies?
    a. Are the risk mitigation and ongoing monitoring strategies operational for the context in which this algorithm is being applied?
4. Is there a clear plan among those overseeing the application of this algorithm for dealing with risks which may arise, and which have not been identified by the AIA?
5. Have all the requirements of the Directive been met to a sufficient degree?
    a. Is a notice given to users that decisions will be automated?
    b. Are the explanations generated about the decisions sufficiently meaningful?
    c. Are the testing and monitoring processes sufficient and proportional to the level of risk?
    d. Are data quality measures sufficient?
    e. Is adequate employee training planned?
    f. Is there an easy to access recourse mechanism for clients to challenge the administrative decisions?

If "yes" to all of the above, proceed to Step 6. If "no", repeat Step 4.

Key Decision Point: Peer reviewers should prepare a set of findings describing whether or not they deem the risk assessments, risk mitigation strategies, and ongoing risk monitoring strategies sufficient to support the release of the decision-making tool into production.

Step 6: Report results

Responsibility: Peer Reviewers and Department Management

Description: Having conducted a fulsome algorithmic assessment—from initial AIA to peer review of recommended risk mitigation—the reviewers shall deliver a report to the department's management, who will review and publicize documents in an open and transparent way as soon as possible per the Directive and/or other governing policies. For example, the peer review report could be published on the "open gov" portal in the spirit of transparency. However, depending on risk and contention of the tool, and intellectual property agreements, certain technical and other sensitive details about affected communities may best not be shared widely.

Key Guiding Principle(s): Vigilance; Accuracy; Independence; Transparency

Key Questions:

1. Has there been open and widely accessible communication about the process of the review, as well as its limitations: assumptions, contingencies, and exclusions?

2. Have peer review recommendations—e.g. processes for continuous monitoring proportional to impact level and type of algorithm—been implemented within the department?

Appendix B: Model Evaluation Grid for Peer Reviewers

This document is intended to be used by peer reviewers alongside the AIA results and supporting documentation. The grid is intended to clearly prompt peer reviewers (and others) to consider more deeply the links between various stakeholders and an array of socioethical concerns that might arise in the development or use of the automated decision-making tool. Each empty cell in the grid is meant to be filled with questions, concerns, and evidence from supporting documentation related to each stakeholder-concern pairing. A completed grid is meant to help guide a discussion among peer reviewers whether or not the algorithm has been designed in such a way as to anticipate the relevant socioethical concerns and mitigate any associated risks. It is intended that each member of the peer review committee use the grid independently and then submit to the moderator of the peer review committee for future discussion, especially around areas of discrepancy among completed grids.

Peer Review Evaluation Grid

		socioethical concern					
		accuracy	transparency	gaming	bias	accountability	other
stakeholder	affected popultion						
	decision-makers						
	host department						
	vendors						

Appendix C: Supporting Documentation Framework

This suggested list of documentation is intended to help the key government contacts in completing the AIA impact scoring tool and in their conversations with clients or design teams. The amount of information included in this documentation should be proportional to the anticipated impact level to help support potential peer review activities.

Documentation Collected Prior to AIA

This documentation is intended to provide the detail needed to support binary (yes/no) answers when completing the AIA and will provide peer reviewers a reference to help them understand what informed the answers in the AIA.

General Documentation

Documentation should describe in some detail:

1. List of departmental areas of public scrutiny and areas of frequent litigation.
2. Is the algorithmic decision high-stakes (including potential impact on environment, clients, staff, etc.)? Why or why not?
3. Is the algorithm legible to non-AI experts? Why or why not?
4. List of potential functions of the system; i.e. "assisting" a decision-maker versus "replacing a decision-maker".
5. List and description of "lengths" of impact of decision, to more concretely define a "brief impact" versus a "long-term impact", that does not justify brevity with the decision being reversible.
6. List of information considered "personal information" that is included in input data.

Algorithm-Specific Information

This information may be taken directly from business case documents or similar documents describing the algorithm, its development, and assessment. This will support the completion of the AIA and the Peer Review process, if required, by providing more details about what, specifically, informed the completion of the online impact scoring algorithm or AIA. We recommend documenting the following information, in detail proportional to the anticipated impact level, related to the respective sections of the AIA Questionnaire:

Risk Profile

7. Description of who the "clients" are, including evidence of their vulnerability (if applicable).
8. Description of the stakes of the decision and reference to the spectrum of low-to-high stakes described above, including justification for the level of impact it has been assigned in the AIA.
9. Description of expected impact on staff.

Project Authority

10. Note existing legal or policy authority for the project and add justification as to why it does or does not fall within that scope.

About the Algorithm

11. Justification for why the algorithm will or will not be a trade secret, with reference to the document defining such algorithms suggested above.
12. Justification for the level of "difficulty" or complexity of the algorithm itself.

Impact Assessment

13. Report of environmental scans including research on potential or expected impacts on:
    a. Rights and freedoms of individuals;
    b. Health of individuals;
    c. Economic interests of individuals; and
    d. Ongoing sustainability of environmental ecosystem.

About the Data

(A) Data Source

14. List of input data into system and data sources.
15. Description of ownership of, and access to data, including any intellectual property agreements that have been put in place to govern access to data.
16. Detailed description of who controls the data mechanically.
17. List of IT systems with which the system will interface.
18. Name of the federal institution that collected the data which will be used to train the system including collection methods. If separate from the institution using the algorithm, description of the relationship between the two institutions.
19. Name of the federal institution that collected input data including collection methods. If separate from the institution using the algorithm, description of the relationship between the two institutions.

(B) Type of Data

20. Description of input data in terms of structure.

Mitigation Questions and Answers - Consultations

21. List of internal and external stakeholders contacted for consultation and a description of their feedback.
22. Log of feedback categorized by date, name and title of the personnel, and a summary of their responses.

De-Risking and Mitigation Measures

Data Quality

23. Report of processes used to test against biases and other unexpected outcomes (this may include frameworks, methods, and/or guidelines), and the results of those processes.
24. URL where it is publicly available, if applicable.
25. Justification of resolved data quality issues.
26. Documentation of the Gender Based Analysis.
27. Name and title of the individual and their department responsible for the design, development, maintenance, and improvement of the system.
28. Proposal for managing risk if outdated or unreliable data is used for the automated decision.

Procedural Fairness

29. Name of the authority responsible for the audit trail as identified in legislation.
30. Sample record of recommendations or decisions made by the system, or, if in pre-production, a sample of anticipated recommendations or decisions (and details in 31-33).
31. Information about the log of all the changes made to the model and the system.
32. Log of which version of the system was used for each decision.
33. Sample log of instances where overrides were performed.

Privacy

34. Documentation of the completed privacy impact assessment.

Mapping to the Requirements in the Directive

35. Mapping of Directive requirements and explanation of how they were met.

References

Aviv Gaon & Ian Stedman, "A Call to Action: Moving Forward with the Governance of Artificial Intelligence in Canada" (2019) 56:4 Alta L Rev 1137 (HeinOnline).

Berryhill, J., Heang, K., Clogher, R. and McBride, K. (2019). "Hello, World: Artificial Intelligence and its Use in the Public Sector," OECD Working Papers on Public Governance No. 36, November 2019. Online at https://www.oecd.org/governance/innovative-government/working-paper-hello-world-artificial-intelligence-and-its-use-in-the-public-sector.htm

Denise Avard, "Research ethics boards and challenges for public participation" (2009) 17:2 Health LJ 66 (HeinOnline).

Derrick, G. (2018) *The Evaluator's Eye*. London, UK: SpringerLink.

Derrick, G. and Samuel, G. (2017). "The future of societal impact assessment using peer review: pre-evaluation training, consensus building and inter-reviewer reliability," *Palgrave Communications*, 3. Online at https://doi.org/10.1057/palcomms.2017.40

Diakopoulos, N. (2015). "Algorithmic Accountability: Journalistic investigation of computational power structures," *Digital Journalism: Journalism in an Era of Big Data: Cases, Concepts, and Critiques, 3*(3), 398–415. Online at https://doi.org/10.1080/21670811.2014.976411

European Parliamentary Research Service Scientific Foresight Unit (EPRS STOA) (2019). "A governance framework for algorithmic accountability and transparency." Online at https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf

Government of Canada. (1999). A Framework for Science and Technology Advice: Principles and Guidelines for the Effective Use of Science and Technology Advice in Government Decision-making. Online at https://www.dfo-mpo.gc.ca/ae-ve/evaluations/18-19/96175-Summary-eng.html

Impact Assessment Agency of Canada (IAAC). External Technical Reviews. Online at https://www.canada.ca/en/impact-assessment-agency/services/policy-guidance/external-technical-reviews.html

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). "A governance framework for algorithmic accountability and transparency," Brussels: European Parliamentary Research Service. Online at https://doi.org/10.2861/59990

Mantelero, A. (2018). "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment," 34:4 Computer L & Sec Report 754 (HeinOnline).

McKelvey, F. and MacDonald, M. (2019). "Artificial Intelligence Policy Innovations at the Canadian Federal Government," *Canadian Journal of Communication Policy Portal* Vol 44: 43-50. Online at http://doi.org/10.22230/cjc.2019v44n2a3509

Michael McDonald, "From code to policy statement: creating Canadian policy for ethical research involving humans" (2009) 17:2 Health LJ 12 (HeinOnline).

OCHA. The Centre for Humanitarian Data. (2019). A peer review framework for predictive analytics in humanitarian response, draft report. Online at https://centre.humdata.org/wp-content/uploads/2019/09/predictiveAnalytics_peerReview_updated.pdf

Office of the Privacy Commissioner in Canada (2020), "Expectations: OPC's Guide to the Privacy Impact Assessment Process" March 2020. Online at https://www.priv.gc.ca/en/privacy-topics/privacy-impact-assessments/gd_exp_202003/#toc5.

Reisman, D., Schultz, J. Crawford, K and Whittaker, M. (2018). "Algorithmic Impact Assessments: A practical framework for public agency accountability," a report prepared for the AI Now Institute April 2018.

San Francisco Declaration on Research Assessment, online at https://sfdora.org

Sonia K Katyal, "Private Accountability in the Age of Artificial Intelligence" (2019) 66 UCLA L Rev 54 at 110 (HeinOnline).

Stats NZ (2018). Algorithm assessment report. Online at https://data.govt.nz/use-data/analyse-data/ government-algorithm-transparency.

Thelwall, M. (2019). "Artificial Intelligence, Automation and Peer Review," a briefing paper for Joint Information Systems Committee UK. Online at http://repository.jisc.ac.uk/7614/1/AI_and_peer_review_briefing_paper.pdf

Tri-Agencies. (2018). Tri-Council Policy Statement on the Ethical Conduct for Research Involving Humans, Online at https://ethics.gc.ca/eng/documents/tcps2-2018-en-interactive-final.pdf

UncannyAI. (2019). Designing Legible AI. Online at http://imagination.lancaster.ac.uk/update/designing-legible-ai/

US, Bill S1108, *Algorithmic Accountability Act of 2019,* 116th Cong, 2019 (not enacted).

Wachter, S., Mittelstadt, B., and Floridi, L. (2017). "Transparent, explainable, and accountable AI for robotics," *Science Robotics*, 2(6)

World Economic Forum. 2020. AI Procurement in a Box: Project overview. Online at http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_Project_Overview_2220.pdf

Author Information

Dr. Kelly Bronson holds a Canada Research Chair in Science and Society at University of Ottawa in the School of Sociological and Anthropological Studies with affiliation to the Institute for Science, Society and Society (ISSP.uottawa.ca) and Centre for Law, Technology and Society. She studies and intervenes into science-society tensions that erupt around technologies–GMOs, fracking, big data & AI—and their governance. Her aim is to bring community values into conversation with technical knowledge in the production of evidence-informed decision-making. Dr. Bronson's policy experience involves advising government and serving on expert panel committees like the Council of Canadian Academies and her policy-relevant research has been funded by the Social Sciences and Humanities Research Council of Canada, as well as a wide variety of private foundations. She has published her work in regional (Journal of New Brunswick Studies), national (Canadian Journal of Communication) and international journals (Science Communication, Journal of Responsible Innovation, Big Data and Society). Before her graduate training in the social sciences (PhD, York University), Dr. Bronson earned degrees in biology and worked as a lab bench scientist practicing genetics (Queen's University, Kingston, Canada).

Dr. Jason Millar holds the Canada Research Chair in the Ethical Engineering of Robotics and AI, and is an Assistant Professor at the University of Ottawa's School of Electrical Engineering and Computer Science with a cross-appointment in the Department of Philosophy. He leads the Robotics and AI research cluster at uOttawa's ISSP. He has authored book chapters, policy reports, and articles on the ethics, ethical design, and governance of robotics and AI. Dr. Millar consults internationally on policy, and ethical engineering issues in emerging technology. His work is regularly featured in the media, including articles in publications such as *WIRED* and The Guardian, and interviews with the BBC, CBC and NPR. He recently authored a chapter titled *Social Failure Modes in Technology and the Ethics of AI: An Engineering Perspective*, for the *Oxford Handbook of Ethics of AI* (OUP), and a chapter on *Ethics Settings for Autonomous Vehicles* in *Robot Ethics 2.0* (OUP). He co-authored a chapter on *Hacking Metaphors* in technology governance for the *Oxford Handbook on the Law and Regulation of Technology* (OUP). His research interests include developing tools and methodologies engineers can use to integrate ethical thinking into their daily engineering workflow, and focuses on applications in automated vehicles, artificial intelligence, healthcare robotics, social, and military robotics.