Guide on peer review (Draft)

Contents

Introduction

The Directive on Automated Decision-Making (directive) is a policy instrument that sets requirements for federal institutions to ensure the use of AI or other automated systems in administrative decision-making is compatible with the core principles of administrative law such as transparency, accountability, legality, and procedural fairness. Section 6 of the directive lists these requirements, one of which is to complete an Algorithmic Impact Assessment (AIA) that will determine the scaled requirements of the directive, based on the calculated impact level of an automation project. Projects assigned an impact level of 2 or higher are subject to the peer review requirement (subsection 6.3.5) that mandates publication of a complete review or plain language summary prior to the system's production.

Peer review is a quality assurance mechanism in which the project is subject to scrutiny by experts in the relevant domain. In the context of the directive, it involves an assessment of the AIA and supporting documentation to validate content integrity, technical soundness and ethical considerations. The completion and publication of a peer review can help departments have confidence in the quality of their automated system, ensure effective compliance with the directive and foster greater transparency.

This document supports federal departments commissioning a peer review and individuals undertaking the review. It defines a process, proposes roles and responsibilities, and identifies best practices to improve the consistency and robustness of reviews. The *Peer Review for Automated Decision-Making Tools Under Canada's Directive on Automated Decision-Making* (Bronson & Millar, 2020) report was a key source that helped to inform the development of this guidance.

Process

Prepare for review

Confirm need for peer review and seek support

To determine applicability of the peer review requirement, departments should first confirm that their project is within scope of the directive. The directive applies to any system, tool, or statistical model

used to make or support an administrative decision or related assessment about a client. See the Guidance on the Scope of the Directive on Automated Decision-Making for more information.

Projects within scope of the directive must complete the AIA, a mandatory questionnaire designed to help departments better understand and manage the risks associated with automated decision systems. The AIA is composed of weighted questions that assess factors such as a system's design, algorithm, decision type, impact and data. Responses to the questions contribute to a score that determines the impact level assigned to a project. The impact level ranges from 1 (little impact) to 4 (very high impact) and is based on criteria of reversibility and expected duration. Automation projects assigned an impact level of 2, 3 or 4 must undergo peer review.

Identify suitable reviewers

Departments commissioning a review are responsible for contacting and selecting potential reviewers. Appendix C of the directive identifies peer review requirements that are proportionate to the impact level. As a project's impact level increases, it is expected that departments consult a greater number of reviewers. Even when not required, the inclusion of multiple experts is strongly recommended to ensure diverse views and consideration of both technical and ethical issues.

Areas of expertise

The required areas of expertise will vary according to the project. Reviewers should be qualified subject matter experts with specialized knowledge and experience relevant to the project in areas including:

- technical: artificial intelligence, machine learning, data science, statistics, computer science, systems engineering or other related fields
- ethical: ethics, privacy, public policy, diversity and inclusion, human-centred design or other relevant areas

Qualifications and experience

To be considered a qualified expert, reviewers must have sufficient depth and breadth of expertise obtained from at least 5 years of work experience. Examples of relevant work experience could include:

- conducting research and analysis of sociological impacts of projects, programs or policies, for example by using a diversity, human rights or GBA Plus framework
- analyzing datasets to uncover insights, build models, test for biases, and inform decision-making
- designing, developing, and implementing IT services or solutions
- evaluating systems across lifecycles from a sociotechnical perspective

A combination of education and experience may serve as an alternative to work experience, at the discretion of the department.

Experts must also hold or be able to obtain the appropriate security clearance prior to conducting the review.

Experts

Experts should be aligned with the options in Appendix C of the directive:

- employees from a federal, provincial or municipal government institution
- faculty members of a post-secondary institution

- researchers from civil society organizations (e.g., non-governmental organizations, advocacy groups, labour unions, professional associations)
- third-party providers or individuals from other external organizations
- members of a data and automation advisory board specified by Treasury Board of Canada Secretariat

Where possible, it is recommended that stakeholders from impacted groups are included in consultations for all impact levels. In addition, consider intersecting identity factors such as gender, race and ethnicity to promote diversity and inclusion when selecting reviewers.

Departments opting to have a data and automation advisory board conduct the review should contact TBS (ai-ia@tbs-sct.gc.ca) and provide the completed AIA, information on the system, timelines and required security classifications of the reviewers.

Manage conflict of interest

Reviewers are expected to disclose any conflicts of interest that could compromise the impartiality of a review. For example, the following situations could be considered a real, apparent or potential conflict of interest:

- previous or current involvement in the system design or implementation
- institutional affiliation, other professional or personal relationship
- direct remuneration in exchange for the review

The department is responsible for vetting appropriate experts and ensuring that any conflicts of interest identified have been assessed and managed prior to entering into an agreement. As well, all federal public servants must comply with the Directive on Conflict of Interest and Values and Ethics Code for the Public Sector.

To avoid conflict of interest, consider the following best practices:

- Select experts that are external to your department. If commissioning expertise within the public service, ensure that reviewers are not in a closely affiliated business line and have not been previously involved in the project.
- Be mindful of reciprocal reviews. For example, the risk of departments reviewing each other's work could be perceived as a biased agreement undermining the objectivity and integrity of peer review.
- If applicable, any remuneration provided should be fair for all external reviewers, clearly documented and received by the institution where possible, as opposed to directly compensating the reviewer.
- Reach out to the departmental values and ethics team for advice on specific cases.

Establish clear timelines

Early engagement is encouraged as the time required to complete a review will vary depending on the system's complexity, number of reviewers, and impact level. The department should ensure sufficient time for review (for example, 1-6+ months) in the project plan. The review should be initiated early enough that identified issues can be addressed before production, but far enough into the project lifecycle that sufficient information is available for assessment. For example, a review should ideally

occur after privacy and security assessments have been completed or at a minimum underway. Projects that follow an agile development process should aim to have the review completed prior to initial software production.

Clarify roles and responsibilities

The department should clearly define respective responsibilities in the agreement with the reviewer.

Department

The federal department planning to use an automated decision system:

- completes AIA and assembles supporting documentation (see Appendix B for full list)
- ensures necessary funds are available and that agreements involving remuneration are in accordance with federal policies
- develops statement of work that sets out key elements such as purpose, scope, timelines, deliverables and level of security clearance required
- contacts potential reviewers and selects the most suitable among them
- fosters open communication in which reviewers may engage with the project team and developers throughout the process
- provides response to feedback from reviewers that specifies changes or commitments made, or a rationale for not accepting suggested revisions
- coordinates approvals to publish the review
- produces a summary of the findings if opting to not publish the complete review, in consultation with the reviewer prior to publication

Reviewer

The individual undertaking the review with expertise in the relevant context:

- confirms having the expertise and availability needed to conduct the review
- discloses any potential conflicts of interest including financial, personal, professional or institutional relationships with the department
- critically assesses the automated decision system across areas under "complete the review". This includes validating the completion and quality of AIA responses and assessing the supporting documentation without the need to replicate testing
- provides updates in a timely manner
- maintains confidentiality of the peer review process in compliance with applicable policies and laws
- prepares written report, ensuring that comments are fair and recommendations focus on specific actions and areas for improvement
- discloses and describes any use of generative AI to support the review. Refer to the Guide on the use of generative AI for best practices and documentation requirements

Complete the review

Using the checklist in Appendix A, reviewers should provide comments and feedback across the following areas:

Accuracy and completeness of AIA

The AIA is a key document that should be carefully reviewed. As a first step, reviewers should validate the responses against available information to confirm the impact level. Reviewers should also identify any discrepancies in responses that would warrant the AIA to be updated. If there is a need for an updated AIA, the reviewer should not proceed further until the discrepancies have been discussed and addressed. This is essential as the impact level informs the applicable requirements under the directive.

Key questions**:**

- Have all of the questions been completed accurately?
- Does the documentation support the responses?
- Are there any discrepancies from the above questions that would indicate a different impact level than what is captured? If so, does it change the peer review requirement?

Readiness to comply with the directive

The peer reviewer should be familiar with the Directive on Automated Decision-Making and use the evidence provided to identify any potential gaps in compliance. In addition to verifying the completion of an AIA (6.1), the reviewer should validate that steps have been taken to meet requirements related to transparency (6.2), quality assurance (6.3), recourse (6.4), and reporting (6.5).

Key questions:

- Has the project prepared to meet all the requirements specific to the identified impact level?
- Will the timelines specified in the Directive be met? (e.g., publication of peer review and AIA prior to production)
    - If not, has the department advised TBS of a plan to meet the requirements?
- Have the necessary consultations been planned or taken place from the concept stage? (e.g., with legal services)

Data quality

Good data quality is a necessary foundation to build high quality systems. The peer reviewer should examine the processes that occurred and are in place to ensure that data quality is sufficient. This includes planning and decisions on data collection and use, as well as ensuring data governance is in place for any data generated by the system.

Key questions:

- Has the testing and training data been assessed to ensure that it is of sufficient quality? (Refer to the GC Guidance on Data Quality for dimensions of data quality)
- Have the impacts of any remaining data quality issues been documented?
- Is data provenance well documented? Does the metadata capture context, history and ownership of data including sources, when and how data was collected, changes made to the data and by whom?
- Are the appropriate data governance roles, responsibilities and processes in place for the data used and generated by the system?
- Will data be managed in alignment with applicable federal, national or international standards such as the GC enterprise data reference standards or ISO standards on data management?

- Are methods and decisions on data labelling, sampling, and collection appropriate?

Fairness

Systems have the potential to produce inaccurate, biased or inconsistent outputs that could result in unfair outcomes. For example, biases or a lack of representation in the training data can be reflected in the outputs of the system, leading to amplification of those biases. As well, fairness captures risks related to procedural fairness where departments are obligated to provide clients with a meaningful explanation and recourse options in instances where a decision results in the denial of a service or benefit.

Key questions:

- Have the appropriate internal and external stakeholders been engaged?
    - Has feedback been addressed?
- Is the training and testing data representative of the clients being served?
- Is evidence of bias testing sufficient?
- Has the model been assessed for performance across different population groups?
- Are there mechanisms in place to monitor and address unfair outcomes or biases over time?
- Has the system been assessed to understand whether it would create barriers for persons with disabilities?
    - If barriers have been identified, have measures to remove or address them been put in place?
- Does the Gender-Based Analysis Plus address how the system may impact different population groups?
- Will the department be able to provide a meaningful explanation of how and why a decision was made to the client? For example, evidence could include audit trails and justification of model choice and model output explanation method. The explanation should contain the elements set in Appendix C of the directive.
- Are recourse options available for the client?

Privacy

Systems often rely on vast amounts of data and may be vulnerable to privacy breaches. When data collected, used or produced by a system constitutes personal information, it must be managed in accordance with the *Privacy Act* and related policy instruments. The privacy risks will vary based on the amount and type of personal information involved and how the system uses personal information to inform a decision. Refer to the Digital Privacy Playbook for further information on privacy considerations.

Key questions:

- Is personal information being collected and used in accordance with the *Privacy Act* and its related instruments? This could include:
    - consultations with privacy officials
    - completion of a privacy notice statement
    - preparation or updating of a personal information bank

    - establishment of an information sharing agreement
- Has a privacy impact assessment been completed or is there one underway?
- Have privacy safeguarding initiatives been undertaken? This could include the use of privacy-enhancing technologies such as data minimization, de-identification, anonymization and synthetic data
- Is there a plan to regularly review and update privacy deliverables and processes?

Security

The integration of security considerations from the onset and throughout a system's lifecycle is critical to protect sensitive information, build user trust and ensure business continuity. The reviewer should ensure that effective security safeguards have been implemented.

Key questions:

- Has the department undertaken a security assessment and authorization?
- Has an interim or full authority to operate (ATO) been issued?
- Is there a process to document and respond to cyber security events and incidents that aligns with the GC Cyber Security Event Management Plan?

Model development

The decisions made during model development impact the operation and outputs of the system. Models can also perform differently over time and across different population groups. Peer reviewers should be provided with access to the model where possible and review the model and documentation to validate model choices and ensure that the model is appropriate for the intended purpose.

Key questions:

- Has model development been sufficiently documented?
- Does the evidence indicate that the selection of proxies and features is appropriate for the intended use?
- Have appropriate model evaluation and performance measurement metrics been chosen?
- Does the system perform at an acceptable level to meet client and operational needs?
- Has the model been compared to any similar ones in terms of benefits and drawbacks?

Risk management

A robust risk management practice is essential to ensure that risks are identified and managed as they arise. Risks can occur pre-deployment as well as during system operation. The AIA helps to identify risks, however an ongoing approach to risk management that is integrated with the departmental risk management approach is also needed. Reviewers should confirm that risks have been considered and processes for continuous risk management are in place.

Key questions:

- Have risks been measured and accounted for at different stages of the system lifecycle?
- Has the department adequately considered whether the benefits of the system exceed potential short- and long-term harms?
- Has feedback from consultations and prior reviews been incorporated?

- Does the project indicate alignment with the GC Framework for the Management of Risk, Guide to Integrated Risk Management or other similar instruments?

Governance

Clear roles and responsibilities for the system and its outputs are important to ensure accountability for the decisions that the system makes or supports. Reviewers should validate that departments have established appropriate governance measures throughout the system lifecycle.

Key questions:

- Has a clear accountability framework been developed to communicate roles and responsibilities, decision-making authorities and performance monitoring? (e.g., how often performance reports will be developed and by whom)
- Have specific human intervention points been identified during operation and monitoring?
- Has the department established an information management approach to document model versioning and decisions about the system?

Operational readiness

The system should be fully equipped to perform its intended functions effectively and reliably in a real-world operational context. Evaluating operational readiness involves reviewing evidence of thorough testing and verification to ensure functionality, robustness and scalability. As well, reviewers should assess whether appropriate communication and change management practices have been established to support clients and employees.

Key questions:

- Is implementation monitoring in place to ensure the system continues to operate as expected?
- Will information on effectiveness and efficiency of the system in meeting program objectives be published?
- Will the system be updated regularly based on performance or user feedback?
- Is there evidence to support that the system has been adequately stress tested? (e.g., handling large request volumes without significant degradation in functionality)
- Has a decommissioning approach been incorporated into the product lifecycle?
- Has the department made training and documentation available to support employees in using the tools and outputs responsibly?
- Has the department identified system limitations that will be communicated to users?
- Has an approach been developed to communicate the use of these tools to impacted clients?

Final report

For impact levels 2-3, the reviewer is responsible for developing a single report but if multiple reviewers are involved, the department should assign responsibility to consolidate the different areas of assessment. However, independent reports should be produced for impact level 4 where at least two evaluations are required.

The final report should include:

- Date of review, model version, reviewer names and affiliations
- Background, methodology and list of documents reviewed
- Major issues – significant concerns regarding the validity and quality of the project. Often requires substantial changes to be made before the system can launch
- Minor issues - relatively less critical concerns that don't undermine the overall quality of the system. For example, best practices that could be undertaken to supplement the project including use of different metrics or testing as well as additional clarifications to improve writing flow and data presentation
- Recommendations – areas and specific actions for improvement, as well as a conclusion on whether to proceed with system production
- Annex – references and supplementary materials used to support the review

Publish the review

The project lead develops a response that includes corrective actions and commitments to the peer review findings and recommendations. The final report and response are presented to the Assistant Deputy Minister responsible for the program using the system for consideration in advance of proceeding to launch the system. The department then coordinates publication of the peer review report on openly available sites such as the Open Government Portal or departmental websites prior to the system's production. In cases where there are limitations on full disclosure due to security, intellectual property or privacy considerations, departments can opt to publish a plain language summary of the report instead, with clear justification provided.

Enquiries

Please contact the TBS Responsible Data and AI team (ai-ia@tbs-sct.gc.ca) for any questions.

Appendix A – additional resources

- Peer review report template
- Complete the review checklist
- Peer review statement of work template

Appendix B – required documentation

The following should be captured in the documentation provided. In many cases, a well completed AIA will include much of the information below.

- Roles and responsibilities for the design, development, deployment, use and monitoring of the system (e.g., policy and legal authorities, confirmation of approvals)
- Description of system functionality (e.g., reasons for automation, anticipated benefits to the client and organization, points of human intervention during the decision-making process, limitations of use)
- Fairness assessment
    - Analysis of the impacts on clients including evidence of bias testing of the data and model and mitigation measures, recourse options and a completed GBA Plus
    - Evidence of transparency measures such as notice and explanations, publication of AIA and any supporting information including release of source code and reporting

- Information on the data (e.g., data provenance, data sharing agreements, approach to assess and resolve data quality issues and impacts of any remaining issues on the system, data governance measures for input and generated data)
- Access to data: When data is required for the review, if data has been manipulated (e.g., de-identified), the department in collaboration with the reviewer should determine whether this would allow for a sufficient review. More information on de-identification is available in Privacy Implementation Notice 2023 01: De identification
- Information about and access to the model (e.g., model type, other models considered or tested, hyperparameters chosen and approach to tuning and optimization, model performance and metrics, implementation readiness, intellectual property/license restrictions)
- System documentation such as requirements, data model, source code and architecture design
- Stakeholders consulted and summary of feedback received (e.g., What we heard report)
- Audit trails and information on the processes that support their use
- Information on privacy measures undertaken (e.g., privacy enhancing technologies, completed privacy impact assessment)
- Business and IT continuity management strategies and plans for impact levels 3 and 4
- Interim or final authorization to operate, based on the results of a security assessment
- Training and system instructions/procedures provided to employees and information on potential impacts to staff
- Procurement details for systems developed by a third party