Treasury Board of Canada Secretariat | Secrétariat du Conseil du Trésor du Canada

Canada

Comparison of the OECD's AI Systems Classification Framework with the Algorithmic Impact Assessment

June 22, 2021

Author: Benoit Deshaies, Director of Data and AI Policy, OCIO

Summary

OECD AI Systems Classification Framework

- **Identifies characteristics of AI systems** along 4 dimensions.
- The framework can also be used to guide reflection on ethical and social implications of AI systems.
- The results of the characterization can be used to inform policy decisions or evaluations of specific AI tools and use-cases.
- The framework is a work in progress. The OECD has been testing the framework over the past year and is currently holding a public consultation to identify areas for improvement.

Government of Canada (GC) Algorithmic Impact Assessment (AIA)

- **Evaluates the risk of automated decision systems** in order to gauge their potential impact on Canadians.
- The results of the AIA determine mitigation measures required under the Directive on Automated Decision-Making. The mitigations are proportionate to the level of impact identified.

Comparison Table

Element	GC Algorithmic Impact Assessment	OECD AI Systems Classification Framework
Purpose	Assess risk and determine proportionate mitigation measures.	Identify characteristics of AI systems and guide reflection on ethical and social implications.
Scope	Automated decision systems subject to the TBS Directive on Automated Decision-Making (DADM).	AI systems as defined in the OECD Recommendation of the Council on AI.
Structure	Tool structured around risk areas and de-risking and mitigation areas.	Framework organized around context, data and input, AI model, and task and output (each dimension has multiple sub-dimensions; see notes).
Methodology	Survey covering wide range of weighted risk and mitigation areas that calculates a level of impact for an AI system.	Survey organized around the dimensions of the OECD framework that identifies key characteristics of an AI system.
Results	The impact level (I-IV) and corresponding mitigation measures inform applicable requirements under the DADM.	Responses to survey inform AI policy decisions, evaluations of specific tools or use-cases (e.g., credit scoring system, GPT-3, AlphaGo Zero), and reflection on risk.

Sample Output of OECD Framework

Example 1: AI-powered hiring system

OECD.AI Policy Observatory

Characteristics of AI system along 4 dimensions

1. CONTEXT

Sector: Administrative and support services

Business function: Human resources

Critical function: No

Scale: Narrow

System users: Amateur (job candidate and HR)

Impacted stakeholders: applicants / employees

Human rights impact: YES
Well-being impact: YES

2. DATA AND INPUT

Data provenance: Provided by candidate; Observed by algorithms; Derived

Collection: Collected by automated tools

Dynamism: Static; dynamic; real-time

Structure : Unstructured data

Rights & Identifiability: Proprietary personal data; identified

Appropriateness & quality: Unknown

3. AI MODEL

Model type: Hybrid

Model building / training: Semi-supervised

Discriminative model

Model inferencing / use: Probabilistic

4. TASK AND OUTPUT

Task: Personalization, interaction support, recognition

Autonomy level: Medium – provides recommendation for human decision execution

Composite system: Yes

Belongs to core application area: TBD

Sample Output of GC AIA Tool

- Section 1: Impact Level
- Section 2: Requirements Specific to Impact Level
- Section 3: Questions and Answers
  - Section 3.1: Project Details
  - Section 3.2: Impact Questions and Answers
  - Section 3.3: Mitigation Questions and Answers

Section 1: Impact Level : 2

Current Score : 36

Raw Impact Score: 42

Mitigation Score: 36

Will the system only be used to assist a decision-maker? **Points:** +1

Yes

Will the system be replacing a decision that would otherwise be made by a human? **Points:** +3

Yes

Will the system be replacing human decisions that require judgement or discretion? **Points:** +0

No

Is the system used by a different part of the organization than the ones who developed it? **Points:** +4

Yes

Are the impacts resulting from the decision reversible? **Points:** +1

Reversible

How long will impacts from the decision last? **Points:** +1

Impacts are most likely to be brief

Impact Level and Score

Assessment Details

(excerpt)

BENOIT DESHAIES

Hello,

I have looked at the framework for classifying AI and find it an excellent tool. I want to congratulate the OECD for leading this important and useful work.

Below are a few comments and suggestions, and attached is a comparison of the framework against Canada's Algorithmic Impact Assessment.

1. I thought the name of the OECD framework is misleading in using "Classification", and would recommend using "Characterisation" instead. With classification, I assume the AI system would be assigned a single label or category. The output of the framework as I understand it is a set of characteristics for the AI system. As such, "characterisation" would more accurately reflect its purpose.

2. The characterisation of AI systems proposed in this framework could be very useful to describe systems and their basic characteristics in algorithm inventories, or registries of automated decision systems, which are now being built in many jurisdictions.

3. Context. On system users, the label of "amateur" can have a negative connotation (perhaps this varies per country or culture). "End-user (untrained in AI)" could be an appropriate replacement to represent, for example, most users of Google Assistant, Siri, Alexa, etc.

4. Data and input. Many systems will combine data from multiple different sources, for which there would be different characteristics. It is not clear on how to represent that in the framework.

5. Data and input. For the structure of the data, it would be helpful to capture the type of unstructured data as text, audio, still image, video, other (e.g. sensor data).

6. Data and input. In assessing the scale of the data, consider whether a number of records might be appropriate. Tens of gigabytes of textual or structured data might represent many more entries than a single hi-definition video. Also, I wasn't clear on the goal of the static and real-time distinction.

7. AI Model. Consider adding a measure for how easily the model predictions or classifications can be explained (transparency/explainability).

8. Additional items that could be captured:

a. Motivation for the system (improve quality of decisions, lower costs of transactions, use innovative approaches, automate tasks that are not humanly possible, etc.)

b. Assessment of the duration of the impacts. The context currently captures impacts on human rights and well being, but we get no indication of the duration and reversibility of these impacts. Will they be small impacts or large? Will they last days, weeks, years, or indefinitely? How easily can they be reversed?

I look forward to see the evolution of this framework. I would be pleased to provide further input should there be an opportunity to.

Warmest regards,

Benoit