OECD Classification Framework: Lessons Learned

January 11, 2021

This document includes notes and possible recommendations for OECD's framework for classifying AI systems, outlined in the interim report. Recommendations are based on lessons learned from turning the framework into a survey instrument and annotation guideline for the purpose of testing and validating the framework. Recommendations are also informed by preliminary analysis of more than 100 test system classifications completed by Amazon Mechanical Turk workers between Dec 22, 2020 - Jan, 8 2021. Additional test classifications are being collected - both via additional survey responses and through annotation of systems described news articles about AI incidents.

Overarching thoughts:

- A recurring concern with the data and AI model dimensions that came up during this process was the trade-off between available information / burden on framework user and policy relevance. For the purposes of this framework, is knowing some of the specifics (about the system and the technical terms) worth the increased information cost and decreased usability?
- Relatedly, we need to distinguish between 1) content that should be "in" the framework (where goal is classify the system), 2) content that is intended for other goal(s) and 3) content that can be deleted.
  - After some discussion with Dewey, we suggest breaking "other" content into concrete set of questions to ask about the system in parallel to classification (AI model dimension in mind here) and articulation of implications of certain classifications
- Still some lack of clarity around the unit of analysis. Beyond the definition, how do we know where the system starts and ends, the boundaries between the system and the context? Or two systems that are similar, but just have different developers?
- An important consideration we may be missing - is the decision reversible? Relevant to both context [impact] and task/output [action autonomy] is the question of whether the output-directed decision or action can be reversed. This is intuitive for policymakers and straightforward in terms of policy implications.

Preliminary findings from test classifications:

- 38 subjects completed 3 systems for 114 test classifications (data collection ongoing)
- Notable dropoff by subjects asked to use this framework to classify 3 example systems, especially compared to subjects asked to use an alternate framework, with two dimensions, to classify 5 example systems
  - 48% of all partial responses are subjects who read the framework and then stopped participating or started a test classification but stopped participating before completing it (compared to 10% of all partial responses that were dropoffs

after reading the other frameworks). In other words, 76% of subjects who read the framework instructions but then did not complete any classifications were for this framework.

- Dimensions with lower classification accuracy and/or consistency:
  - [context] end user
  - [data] data structure
  - [model] acquisition of capabilities
- Dimensions with greater classification accuracy and/or consistency:
  - [context] risk to individual
  - [context] system performance of a critical function
  - [data] system data collection (automated vs. human)
  - [output] action autonomy (exception of 2 systems, a majority accurately classified)
- Dimensions with high classification variation:
  - [context] sector of deployment, though depends on system
  - [output] system task(s), but most identified at least 1 correct task

Questions when using framework to classify example systems:

- When determining the sector, what if more than one sector seems to fit? For example, if an AV is used as a public shuttle, is this transportation or public administration? Or recreation? For the survey we opted to only allow one selection of one sector.
- Also when determining the sector, should we prioritize the sector for the specific task the system performs or the sector in which the system operates? For example, a recruitment tool used for hiring at a retail company - is that administrative and support service or retail?
- Data collection can be difficult to classify when it is hard to discern the input. For example, for a search engine or navigation tool, is the input the inputted search term or destination? Or the data searched/collected/aggregated by the system to perform the search or navigation task?

Specific edits and recommendations:

- To provide a clear, concise summary of the dimensions for subjects, it was helpful to “cut-and-paste” together text from different parts of the interim report, resulting in the following text:

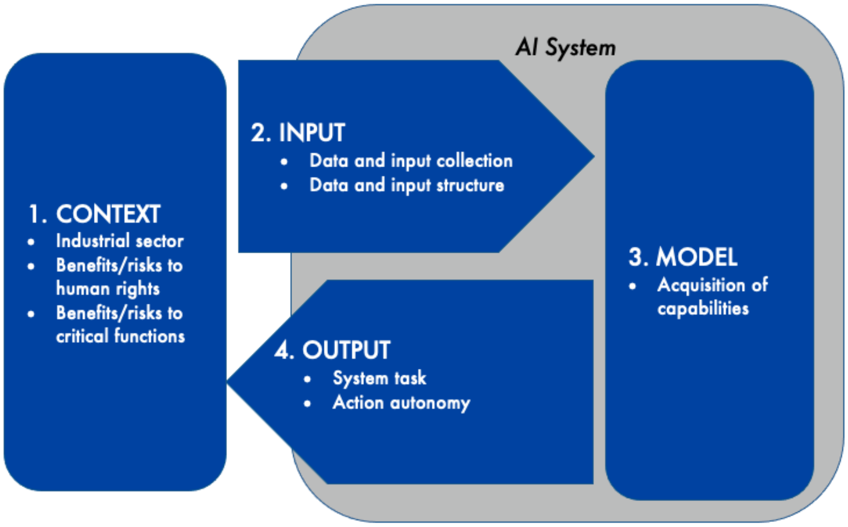
  > “The framework classifies systems along four dimensions: context, input, model, and output.
  >
  > **Context** refers to the socio-economic environment in which the AI system is deployed. Core characteristics of this dimension include the sector in which the system is deployed (e.g., healthcare, finance, defense), deployment impact and scale, effects on human rights, and its critical nature.

Input refers to the input or data used by the AI model to build a representation of the environment. Core characteristics of this dimension include data collection, data characteristics (e.g., form, structure) and data properties (e.g., type, access).

Model refers to the technical components that make up an AI system and represent "real world" processes. Core characteristics of this dimension include model type and acquisition of capabilities (e.g. expert knowledge, data).

Output refers to the tasks the system performs and the action it takes to influence the environment. Core characteristics of this dimension include system task and action autonomy."

- To minimize text and cognitive burden on the survey subjects, I slightly modified Figure 3 to include only the subdimensions we will ask subjects to complete:


AI System
1. CONTEXT
• Industrial sector
• Benefits/risks to human rights
• Benefits/risks to critical functions
2. INPUT
• Data and input collection
• Data and input structure
3. MODEL
• Acquisition of capabilities
4. OUTPUT
• System task
• Action autonomy

- Comments on the visual representation of the framework (from Ashley Llorens, JHU APL)
    - The top-level system diagram should include the user of the system. This will help convey/promote a human-centered approach to system design and deployment.
    - "Context" should be visually conveyed as including the system and users (i.e., context is not just something that the system interacts with, but rather something that it exists inside of). This is not just a philosophical point but a technical reality to consider within the system design and deployment.

- For the **context** dimension
    - We did not include any subdimensions for "deployment impact / scale / impacted stakeholders" because breadth of deployment requires context about system development and deployment that is usually not included in a system description; system maturity did not have any response categories; stakeholders impacted by

the system and For-profit or non-profit use included no instructions, were not sufficiently scoped, and seemed to overlap with information collected elsewhere.

    - For “benefits and risks to critical functions / activities” we did not include critical sector or infrastructure because we already collect industrial sector, so can determine whether it is critical on the back-end. Instead, only ask whether the system does or does not perform a critical function.
        - *Larger concern that criticality is too broad to be useful - it essentially captures all things that could impact any economic or social service or harm an individual (e.g., where does one draw the line for security of citizens, how many citizens must be harmed).*
    - *Although currently a “related consideration” we did include “Users of AI system” as it is a low cost addition with defined response categories that can be collected with minimal system information. But we did add, after discussion when developing the annotation guideline, an additional “amateur” response category, meaning a user who requires no training (with non-expert practitioner being a user with some specific training and an AI expert practitioner being a user with specific training and knowledge of AI).*
- For the **data dimension**, we decided against including several subdimensions.
    - *We included “data and input collection - data collection” which had clear response categories and can be discerned from basic system descriptions.*
    - *We also included “data and input structure - structure of data” but had to provide definition of the subdimension and modify response category definition to make them more concise.*
    - We did not include “data and input collection - data provenance / dynamic nature of data / Scale of data” because defining these subdimesions and response categories requires a good deal of jargon and context, it is also unlikely sufficient information can be conveyed in a system description to classify a system along these dimensions.
    - Same concerns for “data and input structure - data encoding” and “data and input domains” - while more clearly defined, difficult to concisely convey definitions in a survey instrument and ensure adequate information in a system description.
    - “Data quality” and “data qualification” have no defined response categories (e.g., how know if representative versus non-representative or where the data is on spectrum of representativeness)
    - For data domain, concern about classifying the domain of training data versus data used in deployment context
- For the **AI model** dimension, we decided against including several subdimensions.
    - We included only “Acquisition of capabilities” with the following, slightly modified, response categories:
        - Acquisition from knowledge (e.g., learns from expert input or human-written rules)
        - Acquisition from data (e.g., learns from provided data)
        - Acquisition from data and system experience

    - “AI model type” was not included because to classify a system along all 5 sub-dimensions would require a good deal of instruction (i.e. providing definitions for terms) and system information, especially for non-technical users.
        - Specifically for Supervised vs. unsupervised vs. semi-supervised, any attempt to distill resulted in significant overlap with the data dimension, specifically data input structure. To elaborate here, using the text from the interim report, the distinction between supervised learning and unsupervised learning is only whether the “target data points” are “labelled” or “unlabelled” which boils down to a question of structure of data input structure.
        - Interim report reads: “In supervised learning, AI models are used to identify a relationship between input data points and labelled target data points. In unsupervised learning, AI models identify a relationship between input data points and unlabelled target data points. Semi-supervised learning blends supervised and unsupervised learning.”
- For the tasks and output dimension
    - We did not include “Combining tasks and actions into composite systems” because that is captured in the action of the system (autonomy level) subdimension. If a system is “high action autonomy” it means, per the provided definition, that the system combines task and action.
    - It seems the only new information collected here is whether the system generates new content, and given the policy relevance of this, it may be worth moving to a related consideration whereby a framework user could just “check” if the system is one that performs content generation.

Preliminary findings – OECD Framework for the Classification of AI Systems

Overview and goal of the Framework

- The top-level system diagram should include the user of the system. This will help convey/promote a human-centered approach to system design and deployment.
  Also “Context” should be visually conveyed as including the system and users (i.e., context is not just something that the system interacts with, but rather something that it exists inside of). This is not just a philosophical point but a technical reality to consider within the system design and deployment.

- For the survey, we crafted a condensed version of this text: “The framework classifies systems along four dimensions: context, input, model, and output.

Context refers to the socio-economic environment in which the AI system is deployed. Core characteristics of this dimension include the sector in which the system is deployed (e.g., healthcare, finance, defense), deployment impact and scale, effects on human rights, and its critical nature.
Input refers to the input or data used by the AI model to build a representation of the environment. Core characteristics of this dimension include data collection, data characteristics (e.g., form, structure) and data properties (e.g., type, access).
Model refers to the technical components that make up an AI system and represent “real world” processes. Core characteristics of this dimension include model type and acquisition of capabilities (e.g. expert knowledge, data).
Output refers to the tasks the system performs and the action it takes to influence the environment. Core characteristics of this dimension include system task and action autonomy.”

- A recurring concern with the data and AI model dimensions that came up was the trade-off between available information to / burden on framework user and policy relevance. For the purposes of this framework, is knowing some of the specifics (about the system and the technical terms) worth the increased information cost and decreased usability?

- Structuring elements: Still need to distinguish between 1) content that should be “in” the framework (where goal is classify the system), 2) content that is intended for other goal(s) and 3) content that can be deleted. We suggest breaking “other” content into concrete set of questions to ask about the system in parallel to classification (AI model dimension in mind here) and articulation of implications of certain classifications

- *The proposed classification aims to be simplified and user-friendly (see illustrative matrix approach in Annex E) rather than exhaustive, covering the most relevant but not all cases or exceptions.:* A comment on usability based on prelim test data – even using a shortened version of the framework summary, saw notable subject dropoff when asked to apply this framework to system examples, suggesting the instructions were not clear or otherwise leading subjects to walk away from the task.

- *Similarly, all four dimensions raise questions related to jobs and skills*(…): This bullet and the previous one are examples of text that can be consolidated into a section on the implications of specific classifications. Maybe a “Next Steps” section (title is a WIP) that outlines what to consider now that have a classification.

1) CONTEXT

- In ongoing testing, subjects did well classifying risk to individuals and critical functions (C & D) but not well for end-user (F) and varied in performance for sector (A)
- ***Industrial sector***: Questions that arise when trying to classify example systems:
  - what if more than one sector seems to fit? For example, if an AV is used as a public shuttle, is this transportation or public administration? Or recreation?

  - should we prioritize the sector for the specific task the system performs or the sector in which the system operates? For example, a recruitment tool used for hiring at a retail company - is that administrative and support service or retail?
- ***Deployment impact / scale / impacted stakeholders / reversibility***
  - An important consideration we may be missing - is the decision reversible? Relevant to both context [impact] and task/output [action autonomy] is the question of whether the output-directed decision or action can be reversed. This is intuitive for policymakers and straightforward in terms of policy implications.
  - Consider that identifying the correct breadth of deployment may require system information beyond scope of what general description of system can provide, need some specific information about deployment context
  - Need response items (i.e. levels of maturity)
  - Low-risk specifically in terms of risk to individuals
  - Critical sector or infrastructure: we already collect industrial sector, so can determine whether it is critical on the back-end. Instead, suggest only ask whether the system does or does not perform a critical function.
  - Some concern that criticality is too broad to be useful - it essentially captures all things that could impact any economic or social service or harm an individual (e.g., where does one draw the line for security of citizens, how many citizens must be harmed).

2) DATA AND INPUT

- In ongoing testing, subjects did well classifying data collection/origin (A) but struggled to classify data structure (B)
- ***Data collection***: For a search engine or navigation tool, is the input the inputted search term/destination or the data searched/collected/aggregated by the system to perform the search or navigation task?
- ***Data provenence***: May be difficult for a general user
- ***Data encoding***: May be difficult for a general user
- Definition and response items clear, but concern about classifying the domain of training data versus data used in deployment context
- ***Data quality***: Need defined response categories (e.g., how know if representative versus non-representative or where the data is on spectrum of representativeness)

3) AI MODEL

- In ongoing testing, subjects struggled to classify acquisition of capabilities (B)
- ***AI model type***: For a general user to classify a system along these 5 model type sub-dimensions would require a good deal of instruction (i.e. providing definitions for terms) and system information, especially for non-technical users.
- ***Supervised vs. unsupervised vs. semi-supervised***: Attempting to distill resulted in significant overlap with the data dimension, specifically data input structure. Using this text, the distinction between supervised learning and unsupervised learning is only whether the “target data points” are “labelled” or “unlabelled” which boils down to a question of structure of data input structure.
- ***Acquisition of capabilities / model building***: For testing we added a “acquisition from data and system experience” – think Jack’s revisions strengthen this subdimension

4) TASK AND OUTPUT

- In ongoing testing, subjects did well classifying action autonomy (B) and varied in classifying task (A)

- ***Combining tasks and actions into composite systems:*** Whether the output involves an action is captured in the action autonomy subdimension. If a system is "high action autonomy" it means, per the provided definition, that the system combines task and action. It seems the only new information collected here is whether the system generates new content, and given the policy relevance of this, it may be worth moving to a related consideration whereby a framework user could just "check" if the system is one that performs content generation. Another solution would be moving this to AI model dimension with slight alteration per Jack's revisions.