Comments to OECD framework for the classification of AI systems

Contributors: Songül Tolan - Isabelle Hupont – Fernando Martínez-Plumed, Emilia Gómez, Digital Economy Unit, Joint Research Centre, European Commission.

Disclaimer: these are comments based on our scientific-technical individual expertise but they do not represent any policy or official position of the European Commission.

Key questions during the consultation / survey testing phase include:

- *Should there be a core classification framework for less-technical audiences and additional considerations for more technical or informed users?*

The framework is user-friendly enough to be understood by less technical audiences. Having two different versions of the framework might be more misleading than beneficial.

- *Which characteristics should constitute core criteria for a user-friendly policy-oriented classification framework? Could you please comment on the tentative suggestions by the expert group for core criteria? (the criteria that are not marked as "optional criteria")*

See our comments below on this topic.

- *Can users consistently and reliably classify specific AI systems using these core criteria?*

If the optional points that should in our opinion be core are taken into account, we think specific AI systems could be consistently and reliably classified with this framework.

- *Which characteristics may be useful for a more detailed and technically-oriented framework? Could you please comment on the tentative suggestions based on the expert group's feedback for additional, more technical, "optional criteria".*

Validation issues (especially validation metrics) are just superficially mentioned in point 54. There should be more detailed and technical explanations (e.g. types of metrics to be used for each type of AI model: accuracy, ROC curves, confusion matrices...). Examples in section II do not provide any kind of insight about the robustness and accuracy of the systems under the "task and output" criterion.

- *Should there be industry or application domain specific criteria and classifications, e.g. depending on context?*

We miss specific criteria depending on the risk of the intended use of the system for fundamental rights.

Feedback/comments:

- **AI system maturity** AI system maturity [optional criteria] 18. Thanks for considering our proposal for TRL (Martinez Plumed, 2020) building on the NASA TRL framework. We think this is a very good step towards the mapping of AI systems in different maturity level and the consideration of the lab to market pipeline. However, we feel that the framework missed one of the most relevant contributions of our proposal, which is the readiness vs generality plots. There are AI systems that arrive to TRL-9 by specializing much those systems, and when the environment or setup change, the systems need to be redesigned or re-thought. Our paper shows that the concept of TRLs without the generality dimension is very problematic in AI, and this is different from other areas. The solution

we propose is to provide also the level of generality for the defined TRL level. We propose to add to the framework a scale of **AI system generality,** by using a very simple scale, e.g. from 1 to 3 or 1 to 5 from narrow to broad generality, e.g.

  - 1 – narrow context / task (e.g.: speech recognition for a reduced vocabulary in a single language, facial recognition under ideal situations, cleaning robots in specialized tasks)
  - 5 – broad context (e.g. speech recognition in unrestricted context, facial recognition in the wild, general cleaning robots)

- **Distinction between core and optional criteria:** Since the purpose of this framework is to highlight aspects relevant to policymakers due to their impact on public policy, it seems problematic that the distinction between core and optional is made based on access/assessment difficulty. For instance, in the dimension "Data and Input", the criterion "data appropriateness and quality" is set as optional despite its relevance to the OECD AI Principles 1.2 and 1.4. In fact, if certain data appropriateness and quality standards are not met, robustness (Principle 1.4) as well as fairness (Principle 1.2) cannot be ensured and thus the entire AI system may not be applicable. For example, an AI system developed in a controlled environment, may not work correctly in real world applications in unexpected contexts (impact on robustness). Another example, an AI system mostly trained on data from a majority group may not work correctly for minority groups (impact on fairness). Hence, "Data quality and appropriateness" should be a core criterion. The same happens for "H. Benefits and risks to well-being" in "Context", as it is relevant to principles 1.2 and 1.2. This is in fact a very relevant point that we think should be more developed in the framework.
- **"Task" aspect also relevant in dimensions 2 and 3:** Task is currently linked to "output" in the framework (dimension 4), but the task on which an AI system is constructed to perform on can have a relevant impact on data collection (dimension 2) as well as model formulation (dimension 3). Independent of whether data collection is conducted by sensors or humans, it matters if the data was collected for a specific purpose (task) or without any purpose. Collecting data for one purpose and then using it for another purpose (which was not intended during the collection stage) can cause a misalignment between the ideal objective of the task and the approximation of this objective by the data provided. Examples of consequences of such misalignments in the context of healthcare are described in Mullainathan and Obermeyer (2021)[1]. For instance, a risk-assessment tool, that was supposed to be used for assessing the risk of future exacerbation of chronic diseases, for the assignment of preventive treatments was optimized on the basis of medical consumption. For the same risk score, black patients had a higher likelihood of exacerbating a chronic disease than white patients. Thus, according to the risk assessment tool white patients with the same actual risk of developing a chronic disease were more likely to receive preventive care treatments than their black counterparts. That is, the tool was biased against black patients. Since label choice is also part of the model dimension (dimension 3), considering the task in dimensions 2 and 3 has important public policy implications. So one idea would be to consider "Task" as a separate dimension or link it with the AI model, which is operationalizing the task.
- **Level of optionality in criterion "What degree of choice do users have?"**: We think that there is an over-simplification on the levels, as for instance we think there should be an additional level allowing for a higher degree of optionality in the context of human-algorithm interaction. In this level, the AI system provides an intermediate (structured) output generated from unstructured

[1] https://pubs.aeaweb.org/doi/pdfplus/10.1257/pandp.20211078

information and the user formulates a final output/decision using the intermediate AI system output as a basis/support. For example, in the case of clinical decision making, machine learning could be used to generate better quality data or extract relevant features from unstructured data, as discussed in Sánchez-Martínez, Sergio, et al. "Machine learning for clinical decision-making: challenges and opportunities." (2019).[2]

- Inline with the EU AI Act, the definition of critical functions which the AI system performs should be informed by its risks to international human rights (context criterion: "What impact does the system have on human rights?").
- **Context:** in addition to the "users" of the AI system one might consider adding a description of the characteristics of the team of **"developers"** of the AI system, as it has been shown that the characteristics of the developers (e.g gender, country, cultural background) may impact the way AI systems are built as developers incorporate unconscious biases. That is the main motivation towards advocating for diversity of teams that create AI system (Freire, Porcaro and Gómez, 2020)[3]. It would be another aspect to add into the context. This criteria also carires social and policy implications.

[2] https://europepmc.org/article/ppr/ppr103512
[3] https://arxiv.org/abs/2001.07038