Additional notes on OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS – PUBLIC CONSULTATION ON PRELIMINARY FINDINGS

Preface

Thank you for this opportunity to provide comments. I hope that these notes will aid in understanding some of the comments posted inside the main paper. Apologies if there is something that is still not clear. The lack of time didn't allow for more precise and concise feedback.

On the AI system lifecycle and the four dimensions for classification

RECOMMENDATION OF THE COUNCIL ON ARTIFICIAL INTELLIGENCE (https://one.oecd.org/document/C/MIN(2019)3/FINAL/en/pdf) defines AI system lifecycle as:

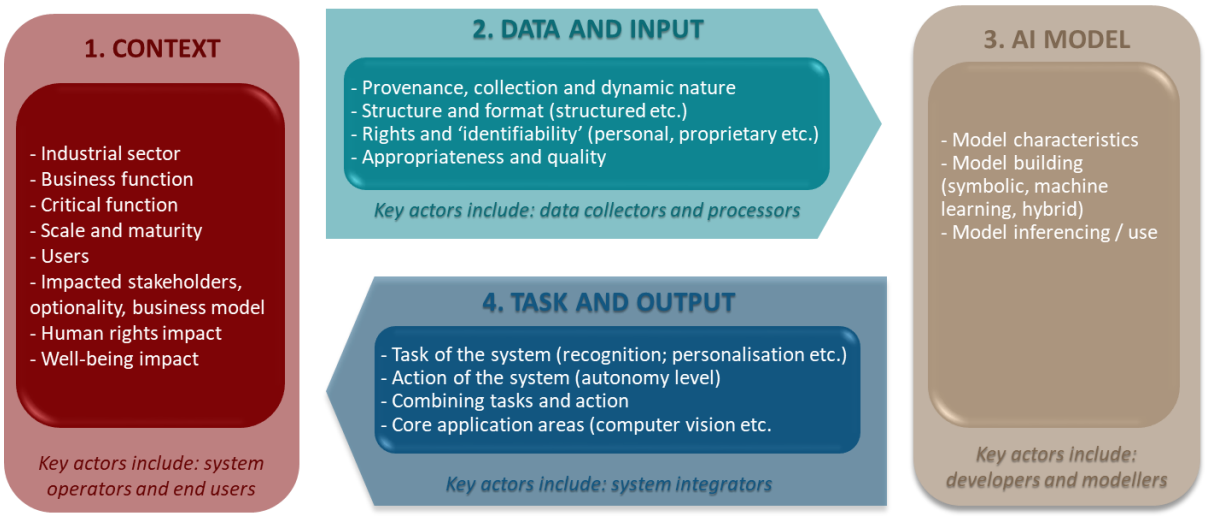
*–AI system lifecycle: AI system lifecycle phases involve: i) '**design**, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) '**deployment**'; and iv) '**operation** and monitoring'. These phases often take place in an iterative manner and are not necessarily*

sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.

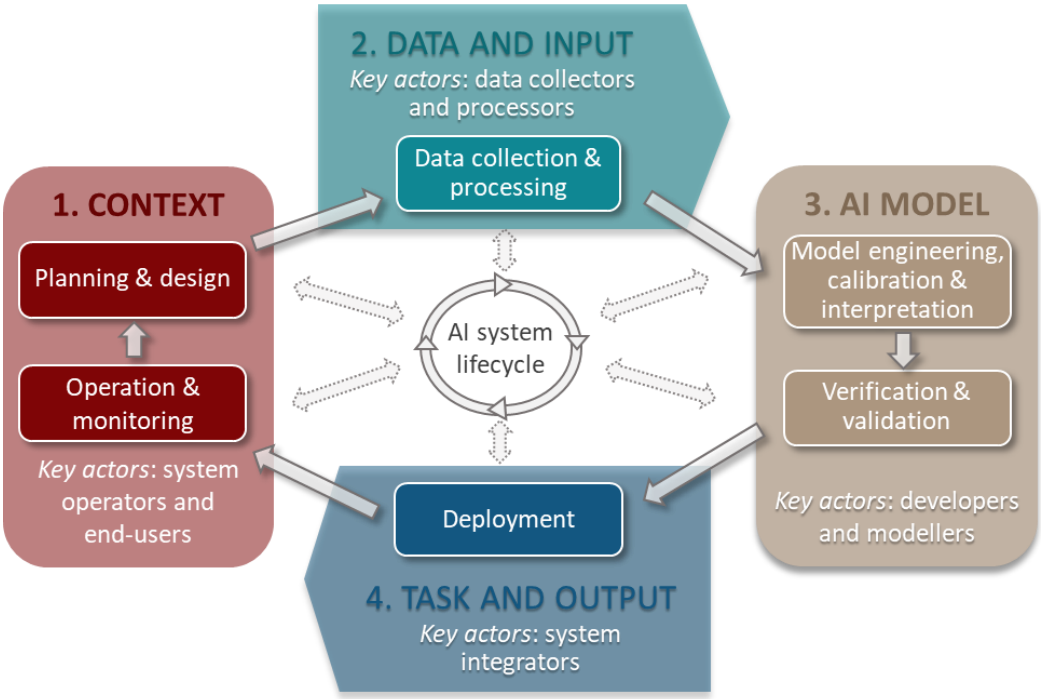
"AI system lifecycle" outlines the lifecycle without a clear delineation between development and run-time environments/contexts. For instance, data collection and processing can be performed **by AI at run-time** and **for AI at design and development time**. However, the terms **operation** and **design** indicate that the lifecycle covers both aspects. And the term **deployment** can only mean transition between the two.

Let's now consider two illustrations presented in the OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS.

In point 4. **Figure 2. Characteristics per classification dimension and key actor(s) involved**



And in point 7. **Figure 3. The AI system lifecycle**



And also some definitions, namely:

25. ***The Context dimension*** *can be associated with the '**planning and design**' stage of the AI system lifecycle as well as, following deployment, with the '**operation and monitoring**' phase. Planning and design of the AI system involves articulating the system's concept and objectives, underlying assumptions, context and requirements (OECD, 2019[14]). Planning and design currently involves expertise such as data scientists, domain experts, and governance experts.*

The term **context** is very ambiguous in this definition. If it means socio-economic environment (as defined in point 1.), then it's the **operational** context. Therefore, "*planning and design*" is not part of it. Instead, it is part of the **development** context. The key actors will consequently be different and could be identified more accurately, based on the operational vs. development context.

41. The Data and input dimension maps directly to the 'data collection and processing' stage of the AI system lifecycle (Figure 3), which includes gathering and cleaning data, possibly labelling, performing checks for completeness and quality, and documenting the characteristics of the dataset. Dataset characteristics include information on how a dataset was created, its composition, its intended uses, and how it was maintained over time (OECD, 2019[14]). Data collection and processing currently involves expertise such as data scientists, domain experts, data engineers, and data providers.

The "Data and input" can be related to both **development** and **operational** contexts. If it occurs during AI operation, it is an operational aspect (originally described as Perceiving in **Figure 4. Stylised conceptual view of an AI system (per OECD AI Principles)**. Note that this is not an AI lifecycle diagram, but rather an AI system **operation** diagram). If AI is designed and developed (including labeling of data, training, validation of models, etc.) it is a development aspect. The key actors will differ for these two relations. For example, during the run-time, the data can be gathered, and learning can occur with the help of a user or operator. If it happens during AI operation, it is an operational aspect if AI is designed and developed (including labeling of data, training, validation of models, etc.) If it happens during AI operation, it is an operational aspect if AI is designed and developed (including labeling of data, training, validation of models, etc.)

53. Model building and interpretation involves the creation or selection of models/algorithms, their calibration and/or training and inferencing (i.e. use). It also involved verification and validation whereby models are executed and tuned, with tests to assess performance across various dimensions and considerations. Model building and inferencing involves expertise such as modellers, model engineers, data scientists, domain experts. Model verification and validation currently involves expertise such as data scientists, data/model/systems engineers, governance experts.

Some terms here relate to the **operational** phase and some to the **development** phase. For example, for non-learning AI, training, verification, and validation occurs in the **design** context. On the other hand, inferencing (use) occurs in the **operational** context. If this criterion is factored in, the key actors can be identified more accurately.

62. The Task and output dimension can be associated with the 'deployment' stage of the AI system lifecycle (OECD, 2019[7]). Deployment into live production involves piloting, checking compatibility with legacy systems, ensuring regulatory compliance, managing organisational change, and

evaluating user experience. Deployment currently involves expertise such as system integrators, developers, systems/software engineers, testers and domain experts.

Lastly, the "*Task and output*" dimension cannot be associated with the **deployment** phase because what an AI system does occurs during an **operational** phase of the AI lifecycle, and **deployment** is a separate phase.

It could be beneficial to separately discuss and illustrate the **development** and **operational** contexts and their corresponding actors to alleviate the above issues. If necessary, these two viewpoints can be reconciled/merged for the cases where there is no difference between them. For example, when AI is trained and operated simultaneously, and deployment doesn't make any sense beyond the initial deployment. Another example is the case when AI itself is performing updates/deployments. In such cases, everything happens in the operational context, though qualitative differences between sandbox(limited) and fully functional contexts may still need to be made.

On the risks, misuse, and abuse of AI

"Risk - the possibility of something bad happening at some time in the future; a situation that could be dangerous or have a bad result" - **Oxford English Dictionary**

AI is a new technology, and with time it may become more or, due to its evolving nature, less understood and predictable. There still be risks associated with intended applications due to known failures and errors. But there also be risks of unknown, usually adverse effects. The latter pose the most significant risks. So, it is more important to predict, search, discover, and plan for the unknown impacts of AI. It worth noting that the known and unknown risks would lie not only in the area of operation but also along all four identified in the framework quadrants and beyond, e.g., storing (physical and electronic), archiving, hauling, etc. It is during these less important states and transitions when the most unsanctioned actions and accidents usually occur.

But what if a policy would demand a proactive role from academia in finding adverse side effects of AI (algorithms, applications, etc.) and an appropriate mitigation plan. One choice would be to be passive like this professor and AI Ph.D.: "I'm only interested in getting the algorithm to work, suggesting how to use it is not in my interest, financial or otherwise." The absence of the necessary research could indicate higher risks and lower trust. The other choice would encourage risk-sharing between all involved parties, and investment will be not mainly for profit but also for safety, ethics, security, stability, etc. It will also promote knowledge circulation within academia before bad things happen in the field. A complimentary classification of the resultant AI based on these two choices would help the stakeholders better understand their proportion of the risks.

In his book, I, Robot, Isaak Azimov depicted nine different applications of AI. In all of them, the culprit was a human factor: error or negligence, or ambition or malice. One could argue this is fiction and that his AI was more sophisticated than what we have. But this is not the point of this remark. He wanted to say that the **weak link** with any knowledge or technology will always be human, whether it's a designer, operator, user, government, hacker, or thief.

Let alone AI, any technology has an upside and a downside potential, and it's up to the people to recognize the downside potential (a.k.a. Loopholes, bugs, etc.) impact and manage the risks associated with it soon as possible. It helps to raise awareness and reduces the number of loopholes that can be abused.

The lack of expertise in AI or the potential consequences of its use or misuse is precisely why this technology is riskier than others. It should also be emphasized that, unlike another tech, AI can be triggered remotely via the Internet or its sensors and has the potential to **initiate interactions with humans or other AI systems or non-AI systems and technologies**.

Besides prominent examples of both harmful and benevolent technologies like nuclear power and medicine, let me offer two examples of misuse or unintended consequences related to AI.

One was highlighted in The Great Hack documentary about the misuse of AI technology in the 2016 US election with two companies in the spotlight, Cambridge Analytica and Facebook. More details can be found in this wiki article and in the EPRS study.

The second was presented in the documentary called /the social dilemma. The testimony of the prominent engineers and scientists is about the technologies and ideas that were intended for the good of humanity but may turn out to be quite harmful to human and social health. See also research by EPRS. Especially chapters 2.1, 2.2, 3.2. Also, a few papers:
Computer-mediated communication and social identity
Can You See the Real Me? Activation and Expression of the "True Self" on the Internet
Effects of Prosocial Media on Social Behavior When and Why Does Media Exposure Affect Helping and Aggression?
New Technologies, New Identities, and the Growth of Mass Opposition in the Arab Spring
Facebook's emotional consequences: Why Facebook causes a decrease in mood and why people still use it
Collective Action in the Age of the Internet: Mass Communication and Online Mobilization

On "Task of the system" and the Seven Patterns

The following quotes are presented to clarify the scope and purpose of the Cognalytica research. They were edited for style and brevity. (see original here https://www.cognilytica.com/event/free-webinar-seven-patterns-of-ai-with-ai-today-hosts-cognilytica)

So we find that defining Artificial Intelligence can sometimes send you down the rabbit hole in conversations because everybody has a different opinion on what it is. So, for this webinar and in general, Cognalytica defines Artificial Intelligence as **machine behavior and function that exhibits the intelligence and behavior of humans**.
- Cognalytica

By understanding these seven patterns, you can greatly **simplify your AI projects** by understanding how those seven patterns are implemented with best practices and then

apply those patterns **individually or in combination with others** to achieve the desired end-goals of your AI and machine learning systems.
- Cognalytica

We love using the term AI, but just like the term Big Data. It's actually a little bit hard to wrap your arms around. When you say Big Data, are you saying the same thing I'm saying? What's Big Data? You might think, “if I have a lot of data.” But it turns out it's more than that. Big Data is about the “volume” and the “veracity” and “variety” and that sort of stuff. So yeah, it's all the so-called eight V's of Big Data. The same thing with AI. Are we talking about the same AI? Just like big data has got eight V's. We've got seven patterns.
- Cognalytica

One can see a few potential issues. First, Cognalytica’s research scope, purpose, and definition of AI are not the same as those defined by the OECD. Specifically, the research aims to find patterns of applying human-like behavior and intelligence currently in use, i.e., commercially viable and popular. Some solutions are simply not recognized or lack economic or technical support. For clarity, the two definitions of AI need to be reconciled, and the differences treated appropriately. The chosen definition of the AI system and other important terminology could be added or referenced in the framework document.
Secondly, two clauses can be added in the classification framework to aid the dynamic nature of classification based on Cognalytica research:

a. one, to provide a catch-all for cases whose task-based classification doesn’t fit predefined patterns;
b. two, classification based on these patterns should be regularly reviewed and broadened to include new types of applications.

Thirdly, pattern-based identification performs well in clear-cut cases but very poorly in boundary and mixed case scenarios. Therefore, the classification framework must offer some guidance on what to do when the solution cannot be cleanly fit into the predefined patterns of AI application or uses a combination of patterns simultaneously for the same task.

On composite AIs

The framework may benefit from refining the discussion and, subsequently, the classification of the composite AIs. The reason for this, though, is the following — a composite AI which is essentially an interlinked network of agents. A close second in the space of connections would be IoT. But composite AIs, or for brevity, *IoAI* will pose significantly more risks than IoT and AI. IoAI may inherit and magnify risks from AI and IoT. Check out a quick recap of quirks that pertains to the IoT: IEEE newsletter: Three Major Challenges Facing IoT, 7 Big Problems With the Internet of Things, and Top 10 Biggest IoT Security Issues More research is required to see how more precisely how AI issues would interact with IoT issues.

Different risks will be at various stages, e.g., context, information acquisition, decision making, action planning, and actuation. It will also be hard to anticipate and test composite AIs. For example, a hybrid AI can be planned and tested by the same developer, but the network of AIs cannot. Like any complex system, a network of AIs will have unpredictable bounds and impacts and will be hard or impossible to fit into a formula. We can foresee four

types of effects rooted in unexpected physical and electronic interactions (some other vocabulary can be developed):

1. Obstacles - would lead to the complete loss of some or all AI abilities.
2. Antagonists - would cause hindrance in some or all abilities, and some activities would be performed **below** the acceptable limits (e.g., slower).
3. Promoters - would induce a new ability.
4. Agitators - would cause stimulation in some or all abilities, and some activities would be performed **above** the acceptable limits (e.g., faster).

These effects are also of both physical and electronic nature. Needless to say that all if unexpected are adverse effects, notwithstanding that some could be beneficial in a controlled environment (e.g., optimization above expected limits, seizure of harmful but known and expected side effects, etc.). However, this aspect of AI technology will need to be further investigated.

Issac Asimov's "I, Robot" dedicated the chapter Catch the Rabbit to discuss **the uncertainty of complex solutions in boundary conditions**. It's always the poor capacity planning or buggy boundary condition code. Probably it is the exact root cause of problems with Dave, a composite asteroid mining robot with six subsidiaries — 'fingers.' So it goes that two AI engineers analyzed an issue with this AI. And it turned out to be in the module that coordinated the 'fingers,' and it only occurred under very few conditions. Here is the final dialog of these two engineers. Funny analogy.

"What were those queer shifting marches, those funny dance steps, that the robots went through every time they went screwy?"
"That? I don't know. But I've got a notion. Remember, those subsidiaries were Dave's 'fingers.' We were always saying that, you know. Well, it's my idea that in all these interludes, whenever Dave became a psychiatric case, he went off into a moronic maze, spending his time *twiddling his fingers.*"

This book is remarkable and needs to be examined, if not as proof of anything discussed in this paper, then opening up new angles for analysis.

On the subject of stakeholders

AI users, operators, and all other intentional parties can be easily identified. Without them, the system will either not function or will be useless. The story with the affected parties is more obscure. There could be passively affected parties. There could be once, twice, and N removed parties. There could be an effect of accumulation of impact not initially visible, etc. So many uncertainties warrant more elaborate methods of identifying impacts and stakeholders. Unfortunately, this section looks pretty barren present.

On the term Evolution vs. Learning

From Britannica: theory in biology postulating that the various types of plants, animals, and other living things on Earth have their origin in other preexisting types. The distinguishable differences are due to modifications in successive generations.