AI Standards Lab - Feedback on risk terminology and risk sources

Representative- ariel@aistandardslab.org
2024-11-27

Overall, we feel that the current first draft Code is not careful enough with how it defines and uses the word 'risk.' To clarify this concern, the document below reproduces some of the input we gave in the initial consultation before the CoP drafting process started.

We bring this input to the attention specifically of the WG2 (vice)chairs to inform improvements of the taxonomy text.

Specifically, we observe that the current taxonomy, as well as the rest of the draft Code, sometimes uses the term 'risk' to denote the concept that we define as 'risk source' below. This makes the Code more difficult to read and apply.

(start of reproduced input from consultation)

Input: clarifying the term 'risk' and the Risk Taxonomy

Disambiguating the term 'risk'

We observe that in everyday (non-technical) English, the word 'risk' has several somewhat overlapping meanings. The state of the art in risk management, and in building risk management taxonomies, is to use more precise technical terminology to refer to different aspects of the phenomenon 'risk.'

Following the text of the AI Act, we propose (and will use below) the technical terms:

> **Risk:** the combination of the probability of an occurrence of harm and the severity of that harm; [Source:AI Act article 3]
>
> **Harm:** *(noun) Negative event or negative social development entailing value damage or loss to people. [Source: ISO/IEC/IEEE 24748-7000:2022]*

where this 'Harm' definition is, in our opinion, broad enough to cover all the systemic risks considered by the GPAI provisions of the AI Act. We also propose:

> ***Risk source:*** *element which alone or in combination has the potential to give rise to risk [SOURCE: ISO 31000:2018. Some other international standards use the word 'hazard' to denote essentially the same concept: we have no strong opinion on the choice but we will use 'risk source' in this contribution]*

Risk sources are therefore causally upstream of harms. They span a wide range of phenomena: some are purely technical or physical, while others involve actions that are performed (or fail to be performed) by humans. In some cases, risk sources may also be inadequacies of risk management measures, where they describe ways in which risk management measures may not achieve their intended outcome.

Clarifying the nature of the risk taxonomy

The 'risk taxonomy' to be created by the CoP drafting process is described as follows:
'The Code of Practice should help to establish a risk taxonomy of the type and nature of the systemic risks at Union level, including their sources.'

Using the terminology we developed above, we conceptualise the risk taxonomy as consisting of two parts:

- A structured list of **harms**, aimed at clarifying which types and natures of risks that fall under the category 'systemic risk' as intended by the AI Act
- A structured list of **risk sources**, which lists phenomena that may give rise to systemic risk, and which has the purpose of providing useful information to parties who implement the CoP, information that can be used in the 'risk assessment' activities that will be required by the CoP.

This means that in our view, the taxonomy needed in the Code of Practice is not just a system of classifying different individual harms and risk sources: it must also *name* and *describe* these individual harms and individual risk sources.

Completeness of the taxonomy

We propose that the list of **harms** included in the taxonomy is exhaustive with respect to the scope of article 56, the scope of what the AI Act considers 'systemic risks from GPAI'. The discussion of what is a systemic risk (and what is not) is mostly found in the AI Act recitals – we propose that the drawing up activity of the Code of Practice converts these recitals into a list of harms,

However, we observe that the list of **risk sources** in the taxonomy can never be fully exhaustive with respect to the current and future state of GPAI technology, or with respect to all potential uses to which a GPAI model might be put. We expect all parties signing up to the Code of Practice to commit to also making an effort to identify all relevant risk sources that do not appear in the taxonomy included in the CoP document.

That said, the more risk sources are in the taxonomy, the easier it will be for SMEs (and start-ups and large companies) to fulfill the obligations required of them in the CoP, and required of them by the AI Act. We consider the inclusion of a long list of risk sources in the taxonomy to have beneficial effects on the market overall.

We consider the inclusion of a large list of risk sources in the taxonomy to be specifically beneficial also in light of article 1 of the AI Act, which describes that the aim to 'improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI)'.

We envisage that a list of well-named and clearly defined risk sources will be an important tool for communication and cooperation between market parties across the GPAI value chain. We also envisage it as a tool that can promote communication and cooperation across disciplines: our contributions are explicitly designed to cover not only the purely technical or 'hard science' aspects of AI as thought to many ML students, but also socio-technical or 'soft science'' aspects like risk sources flowing from human-machine interaction.

(end of reproduced input from consultation)

Reflections on the above consultation input

It is clear that the taxonomy in the current draft Code goes beyond just naming risk sources and harms, and identifies additional dimensions of analysis. We support this high level of dimensionally.

We consider our proposal above to have a simple list of harms in the taxonomy to be obsoleted by developments in the first draft of the code: for more about harms and article 3(65) in the act see our separate feedback document on the taxonomy.

We still propose to have an extensive list of risk sources in the taxonomy.

Proposals on risk sources (6.3)

We propose adding a lot more content to the taxonomy section of the CoP draft to improve the usability of the Code. Individual risk sources are preferably not described with a single line, but with a title followed by one or more descriptive paragraphs. We would be happy to assist the (vice)chairs in adding content based on our consultation submission or other sources if invited to do so.

We recommend consulting our initial free-text submission to the multi-stakeholder consultation for more detailed descriptions of various sources of systemic risks to be added to the taxonomy 6.3. These 'risk source' sections are contained in the WG2 and WG4 sections of that submission. You might also consult our public domain paper "Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems," currently available on arXiv.