AI Standards Lab - Feedback on Risk Taxonomy

Feedback to the AI Office on Measure 6 Risk Taxonomy of the first draft of the Code of Practice.
Representative- ariel@aistandardslab.org
2024-11-27

Summary of feedback

- Signatories should not just “draw from the elements of this taxonomy;” they should either adopt this taxonomy in its entirety or create a taxonomy that includes all of the items in this taxonomy as a minimum.
- The types of systemic risks should reflect those listed in Article 3(65), and should be distinct from the nature and sources of systemic risks.
- The nature of systemic risks under each dimension should be further refined.
- The sources of systemic risks should also include model limitations or potential failures.

Our Interpretation of Article 3(65) of the Act

Article 3(65) of the AI Act defines “systemic risk” as “a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health,
safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.”

Here the Act references several “natures” of systemic risks:

- “having a significant impact on the Union market due to their reach”
- “can be propagated at scale across the value chain”

It also describes a risk source:

- “high-impact capabilities of general-purpose AI models”

And several types of risks, in terms of “actual or reasonably foreseeable negative effects” on the following:

- “public health”
- “safety”
- “public security”
- “fundamental rights”
- “the society as a whole”

We note that the “high-impact capabilities of general-purpose AI models” point is appropriately captured in Sub-Measure 6.3.1. However, the types of risks described in the Article are not reflected in Measure 6.1, as we will elaborate further below.

In general, we recommend that the (vice) chairs make edits to disambiguate the taxonomy text in the Code so that it becomes very clear that it does not override or replace Article 3(65). Specific suggestions will be made below.

Feedback on Measure 6

The draft mentions:

> Signatories commit to draw from the elements of this taxonomy of systemic risks as a basis for their systemic risk assessment and mitigation.

The text “draw from” implies that the signatories need not consider all of the systemic risks listed in the taxonomy, they just have to use some of it as a basis for their systemic risk assessment and mitigation.

We suggest for the above text to be revised with the following:

> Signatories commit to address all of the elements of this taxonomy of systemic risks as part of their systemic risk assessment and mitigation.

Feedback on Sub-Measure 6.1

The draft lists the following types of systemic risks

- Cyber offence
- Chemical, biological, radiological, and nuclear risks
- Loss of control
- Automated use of models for AI research and development
- Persuasion and manipulation
- Large-scale disinformation

We agree that some of the types of systemic risks listed are relevant to those listed in the act:

- Cyber offence - related to “public security”
- Chemical, biological, radiological, and nuclear risks - related to “safety”
- Loss of control - related to “the society as a whole”
- Large-scale disinformation - related to “the society as a whole”

However, several of the types of systemic risks listed do not seem to best fit this subsection of the taxonomy:

- Automated use of models for AI research and development - this is more relevant to Sub-Measure 6.3.3 on model affordances

- Persuasion and manipulation - this is more relevant to Sub-Measure 6.3.2 on dangerous model propensities, and is closely related to the item “tendency to deceive.”

The draft also mentions the following:

> Signatories may identify further systemic risks beyond those listed above, considering, for example, major accidents, large-scale privacy infringements and surveillance, as well as other ways in which general purpose AI models may cause large-scale negative effects on public health, safety, democratic processes, public and economic security, critical infrastructure, fundamental rights, environmental resources, economic stability, human agency, or society as a whole.

We recommend that the AI Office include these examples of systemic risks as part of the Code of Practice instead of leaving it up to the signatories. Of course, signatories may (and should) identify further risks beyond what is covered in the Code, but the Code should be sufficiently comprehensive that crucial risks are included, and the signatories must address them as part of their systemic risk assessment and mitigation.

We therefore suggest the following as a high-level taxonomy, taking into account the types of systemic risks listed in Article 3(65) of the Act, as well as the examples listed in the first draft of the Code:

- Public health risks
- Safety risks, such as
    - Chemical, biological, radiological, and nuclear risks
- Public security risks, such as
    - Cyber offense
- Risks to fundamental rights, such as
    - Risks to human agency
- Other risks to the society as a whole, such as
    - Loss of control
    - Large-scale disinformation
    - Risks to democratic processes
    - Environmental risks
    - Economic risks

Feedback on Sub-Measure 6.2

We note the dimensions of the nature of systemic risks as per the first draft of the Code of Practice below:

- Origin: Model capabilities, model distribution
- Actor(s) driving the risk: State, group, individual, autonomous AI agent, none (e.g., no clear actor can be identified)
- Intent: Intentional, unintentional (including misalignment)

- Novelty: Precedented, unprecedented
- Probability-severity ratio: Low-impact high-probability, high-impact low-probability, high expected impact
- Velocity at which the risk materialises: Gradual, sudden, continuously changing
- Visibility of the risk while it materialises: Overt (open), covert (hidden)
- Course of events: Linear, recursive (feedback loops), compound, cascading (chain reactions)

We would like to provide recommendations on several dimensions.

On the dimension of "origin," we interpret "model capabilities" and "model distribution" as referring to the stages in which the risk emerges. For example, a model may develop the capability of creating novel bioweapons during the training phase, in which case the risk has emerged during the model development stage of its lifecycle. On the other hand, a model may not have any dangerous capabilities, but is being integrated into an AI system which allows it to take certain actions in the world that are potentially harmful, in which case the risk has emerged during the model distribution stage of its lifecycle. We recommend that this dimension is rephrased as "stage of risk emergence," which includes "model development stage" and "model distribution stage."

We also recommend the inclusion of an additional dimension called "technical attributes," as per the recommendations in our initial consultation survey submission. This dimension would include "model capabilities" and "model inadequacy (technical failure)." This helps to differentiate if a risk arises due to the model developing certain dangerous capabilities or the model failing at certain tasks. For example, a model may be capable of developing novel bioweapons, creating a safety risk; or a model may fail at providing reliable medical advice, creating a risk to public health. Items related to "model capabilities" are further described in Sub-Measure 6.3.1, while we propose that items related to "model inadequacies" are discussed in an additional Sub-Measure under 6.3.

On the dimension of "Actor(s) driving the risk," we recommend including an item in which multiple actors are identified. This includes a combination of human and AI, as well as multi-agent AI systems.

Additionally, we recommend an additional dimension called "reversibility" which includes "reversible" and "irreversible." It is worth noting that the reversibility of reversible risks also exist on a spectrum, where some are more easily reversible and some are more costly to reverse.

Lastly, we recommend removing the dimension on probability-severity ratio. We believe this is more appropriately captured under the risk assessment section of the SSF, where given the results of model evaluations, each risk should be assigned a certain probability and severity.

In summary, we recommend the following dimensions of the nature of systemic risks:

- Stage of risk emergence: Model development stage, model distribution stage
- Technical attributes: Model capabilities, model inadequacy (technical failure)
- Actor(s) driving the risk: State, group, individual, autonomous AI agent, multiple (e.g., combination of humans and AI, and multi-agent AI systems), none (e.g., no clear actor can be identified)
- Intent: Intentional, unintentional (including misalignment)
- Novelty: Precedented, unprecedented
- Velocity at which the risk materialises: Gradual, sudden, continuously changing
- Visibility of the risk while it materialises: Overt (open), covert (hidden)
- Course of events: Linear, recursive (feedback loops), compound, cascading (chain reactions)
- Reversibility: Reversible, irreversible

Feedback on Sub-Measure 6.3

We note the categories under Sub-Measure 6.3:

- Sub-Measure 6.3.1 Dangerous model capabilities
- Sub-Measure 6.3.2 Dangerous model propensities
- Sub-Measure 6.3.3 Model affordances and socio-technical context

We believe this categorisation serves as a good starting point for identifying risk sources.

We also propose an additional Sub-Measure 6.3.4 called “model inadequacies”, to include items related to shortcomings and potential technical failures of a model. This may include limitations or failures related to various stages of the model lifecycle, such as training dataset curation, training, fine tuning, and model evaluations. These relate to technical issues that may have specific mitigation measures, in contrast to “model propensities” which relate more to behavioral aspects of AI models. We have provided an extensive catalogue of risk sources in our initial free-text submission to the multi-stakeholder consultation.

We propose the following changes:

- Include “Resistance to shutdown” under Sub-Measure 6.3.2.
- Include “Access to computational power (e.g., increasing speed of operations)” under Sub-Measure 6.3.3
- Move “Lack of reliability and security” from Sub-Measure 6.3.2 to our proposed Sub-Measure 6.3.4 “model inadequacies.”
- Move “Lack of explainability or transparency” from Sub-Measure 6.3.3 to our proposed Sub-Measure 6.3.4 “model inadequacies.”