- <https://incidentdatabase.ai/taxonomies/>

***NIST, TAXONOMY OF AI RISK***

<https://www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf>

This white paper focuses on the preconditions of trust in AI and aims to further engage the AI community in a collaborative process to encourage consensus regarding terminology related to risk so that these types of risk may be identified and managed.

OECD (Organisation for Economic Co-operation and Development):

* Traceability to human values (laws, human rights, democratic values, diversity…)
* Transparency and responsible disclosure (understand AI outcomes and make the outcomes be based on information)
* Secure, safe and robust AI’s to manage risks better
* Responsibility to comply this points.

Similarly, the European Union Digital Strategyʼs Ethics Guidelines for Trustworthy AI2 identifies

seven key principles of trustworthy AI:

• Human agency and oversight

• Technical robustness and safety

• Privacy and data governance

• Transparency

• Diversity, non-discrimination, and fairness

• Environmental and societal well-being

• Accountability

Finally, US Executive Order 13960, Promoting the Use of Trustworthy Artificial Intelligence in the

Federal Government3 specifies that AI should be:

• Lawful and respectful of our Nationʼs values.

• Purposeful and performance-driven… using AI, where the benefits of doing so

significantly outweigh the risks, and the risks can be assessed and managed

• Accurate, reliable, and effective

• Safe, secure, and resilient

• Understandable…by subject matter experts, users, and others, as appropriate

• Responsible and traceable

• Regularly monitored

• Transparent

• Accountable

So we can divide in **categories of risk:**

* Technical design attributes: factors under control of designers and developers.
* How AI systems are perceived: transparency, explainability, interpretability. This means that some AI outputs need to be done by humans or interpreted by them rather than by an automated process.
* Guiding policies and principles: privacy, accountability, fairness, justice, equity, etc., which cannot be measured consistently across domains because of their dependence on context.

***CSET AI Harm Taxonomy***

<https://github.com/georgetown-cset/CSET-AIID-harm-taxonomy/blob/main/CSET%20V1%20AI%20Annotation%20Guide%20(with%20Schema%20and%20Field%20Descriptions)%2025Jul2023.pdf>

Taxonomy should characterize the harm associated with an AI incident.

Differentiate between tangible and intangible harm.

Tangible harm: incidents involving observable injury, loss, or damage. Differentiate between events(harm needs to have definitely occurred) , near-missses(inminent risk of harm), and issues (non-inminent risk of harm)

Intangible harm: non-directly observable. Depends on social norms, context.

We can separate intangible harm in 1)harm to civil liberties, rights, human rights, or democratic norms 2) detrimental content (hate-speech, misinformation) and 3) different treatment based on a characteristic.

**PROPOSED TAXONOMY:**

* Harm to collectives, nations, or against freedom of choice, biased answers…
* Failure of automatic tasks (self-driving, facial recognition…)
* Accidents related to robots usage or misuse
* Errors because not-enough development, more training needed …
* Not clear, transparent, based on information answers

**PROPOSED TAXONOMY 2:**

* Security and privacy breaches: data breaches, privacy violations…
* Ethical violations: biased-decision making, informed consent violations…
* Social harms: harm to collectives or races; stereotyping, demeaning, erasing social groups…
* Operational features: model-related incidents, algorithm-related incidents, failure of automatic tasks…

**Concrete categories**

**Autoclass methods, unsupervised methods, semi-supervised methods, open ai autoclass, to not classify manually each description**

**HUGGING FACE: NLP MODELS THAT CLASSSIFY SENTENCES WITHOUT LABELS (facebook/bart\_large-mnli)**

**PROBAR INCIDENTS DATASET EN VED DE REPORTS DATASET**