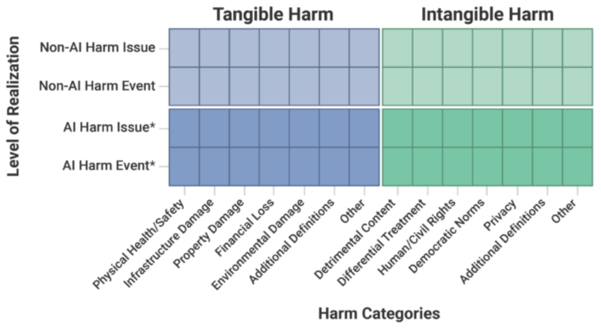
LINK 1: NIST.
https://www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf

- Proposed taxonomy:
    1. **Technical design attributes:** factors that are under the direct control of system designers and developers. Examples:
        - Accuracy.
        - Reliability.
        - Security and resilience.
    2. **How AI systems are perceived:** any assessment of an AI system that is made by a human falls into this category. Examples:
        - Explainability: white box model is explainable, contrary to a black box model.
        - Interpretability: for example, determine whether something is safe or not.

    3. **Guiding policies and principles:** rules about things in our society that we care about, like privacy, accountability, fairness, justice and equity. Difficult to measure them, since they depend on the situation. Examples:
        - Fairness: absences of harmful bias is necessary for fairness.
        - Transparency: by being transparent and sharing details about things like how the AI was trained, what it's meant to do, and how decisions were made in creating it, we can help people feel more comfortable and trust the AI.

LINK 2: CSET.
https://github.com/georgetown-cset/CSET-AIID-harm-taxonomy/blob/main/CSET%20V1%20AI%20Annotation%20Guide%20(with%20Schema%20and%20Field%20Descriptions)%2025Jul2023.pdf

- This paper has the main goal of updating the previous AI Harm Taxonomy of the CSET.
- Goal of the taxonomy: to create a structure for extracting AI harm information that will allow people to draw conclusions from the dataset without having to compile and interpret the individual incident reports themselves.
- Attempts to capture details about the AI system, sector, environment, entities, locations, dates and type of harm that were involved in the AI incident.



*An AI harm occurred when an **entity** experienced a **harm event** or **harm issue** that can be **directly linked** to a consequence of the behavior of an **AI system**.

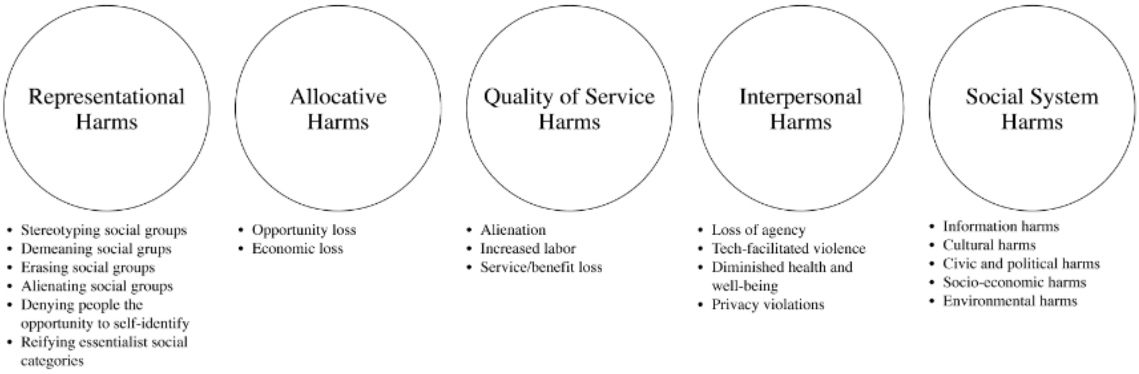
CSET distinguishes between tangible and intangible harm.

→ Tangible harm: observable (there are 3 categories inside: events, near-misses and issues. This is done to make it easier for different people to agree on what's harmful).
→ Intangible harm: harm that cannot, even with additional information, be observed.

LINK 3: Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. https://browse.arxiv.org/pdf/2210.05791.pdf

- They present an applied taxonomy of sociotechnical harms to support a more systematic surfacing of potential harms in algorithmic systems.
- Taxonomies:
  1. Representational harms: how socially constructed beliefs and unjust hierarchies about social groups are reflected in model inputs and outputs.
  2. Allocative harms: how these representations shape model decisions and their distribution of resources.
  3. QoS harms: how choices made to optimize models for particular imagined users result in performance disparities.
  4. Interpersonal harms: how technological affordances adversely shape relations between people and communities.
  5. Social system harms: how algorithmic systems impact the emergent properties of social systems, leading to increased inequity and destabilization.

Figure 1: Sociotechnical harms taxonomy overview.


Representational Harms
• Stereotyping social groups
• Demeaning social grups
• Erasing social groups
• Alienating social groups
• Denying people the opportunity to self-identify
• Reifying essentialist social categories
Allocative Harms
• Opportunity loss
• Economic loss
Quality of Service Harms
• Alienation
• Increased labor
• Service/benefit loss
Interpersonal Harms
• Loss of agency
• Tech-facilitated violence
• Diminished health and well-being
• Privacy violations
Social System Harms
• Information harms
• Cultural harms
• Civic and political harms
• Socio-economic harms
• Environmental harms