AI Ethics in Design

AI ethics is a multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes.

Some principles to consider for designing responsible AI interfaces:

1

Explainability

2

Accountability

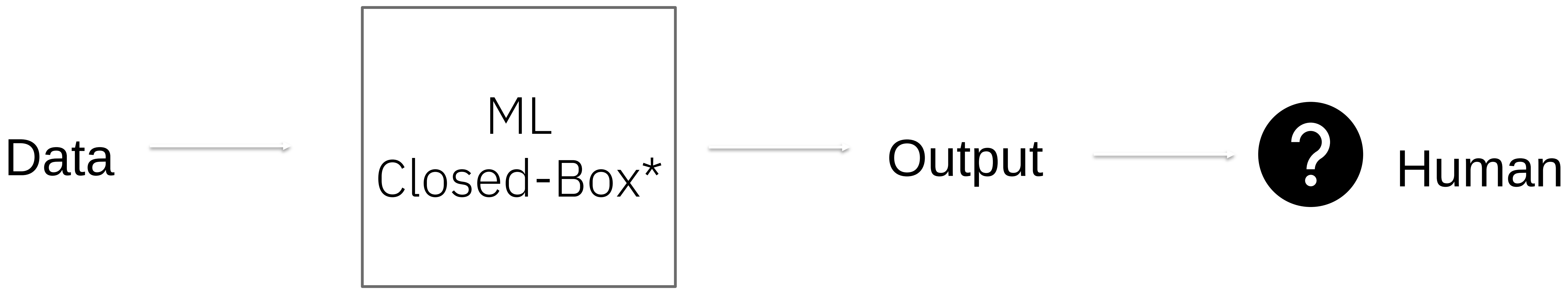
3

Privacy

4

Sustainability

Closed-Box Models



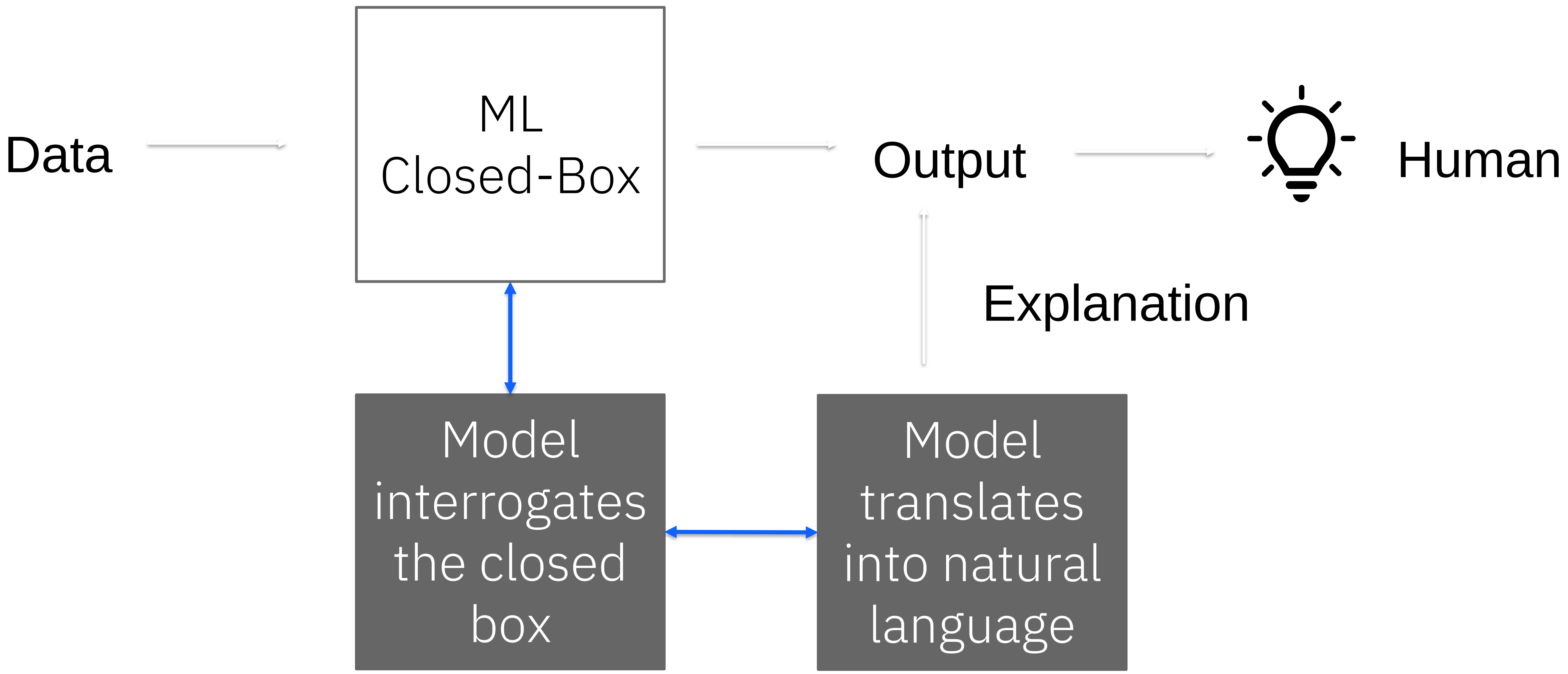
Data → ML Closed-Box* → Output → ? Human

Decision-making process cannot be explained in a way that can be easily understood by humans.

*"Closed-Box" – sometimes referred to as "black box"

Source: https://medium.com/quantumblack/making-ai-human-again-the-importance-of-explainable-ai-xai-95d347ccbb1c

Explainable AI (XAI)



Data

ML Closed-Box

Output

Human

Explanation

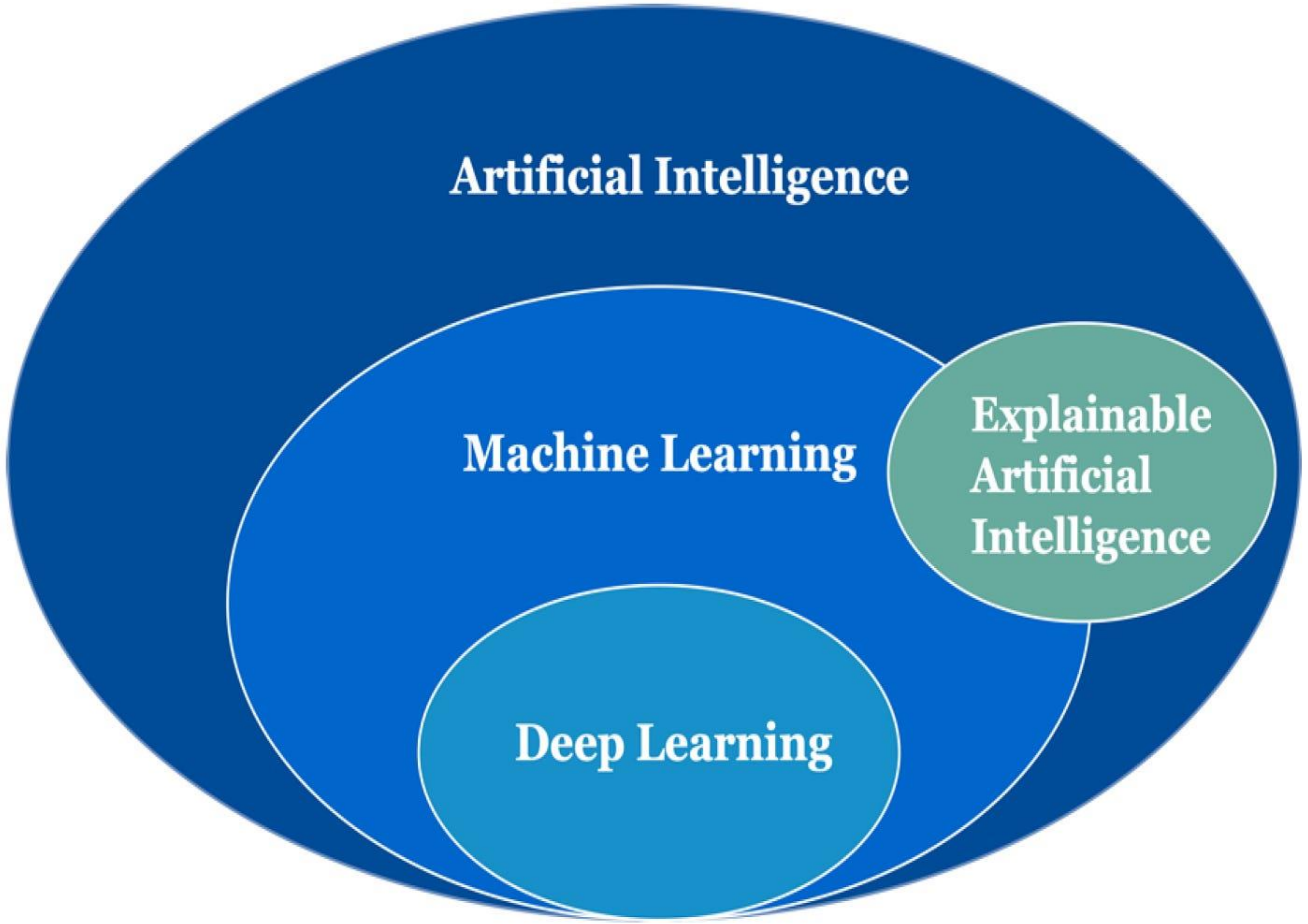
Model interrogates the closed box

Model translates into natural language

Reference: https://medium.com/quantumblack/making-ai-human-again-the-importance-of-explainable-ai-xai-95d347ccbb1c

Explainability

Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.



Learn more: https://www.ibm.com/topics/explainable-ai

Explainability

Closed Box AI ≠ Explainable AI

For example, hospitals can use explainable AI for cancer detection and treatment, where algorithms show the reasoning behind a given model's decision-making. This makes it easier not only for doctors to make treatment decisions, but also provide data-backed explanations to their patients.

UI Solution:

- Transparency: Clearly communicate the system's capabilities, limitations, and the sources of its information.
- Error Handling: Implement robust error detection and handling mechanisms, allowing users to report and correct errors.

Accountability

Common communication issues:

- Misinformation
- Errors
- Need for follow-up/resolutions

Example: Misleading words:

Air Canada is facing a lawsuit from a passenger who claims the airline's chatbot provided misleading Information, leading to financial loss. The passenger argues that the chatbot falsely assured them that their booking was confirmed, which later turned out to be untrue, resulting in missed flights and additional expenses.

Read more: https://www.cbc.ca/news/canada/british-columbia/air-canada-chatbot-lawsuit-1.7116416

Accountability

- **Algorithmic bias** occurs when a system acts in ways that reflect the prejudices inherent in its programming, data sources, or design.

- Gender bias, racial bias, or cultural bias, leading to discriminatory outcomes.
    - For example, a language model might generate stereotypical or prejudiced content, or it might perform less effectively for certain dialects or accents
- Speech recognition (speech disabilities, dialects, accents, stuttering, female voice.. etc)
    - Commonly used applications lack efficiency when exposed to misspelled words, different accents, stutters, etc. The lack of linguistic resources and tools is a persistent ethical issue in NLP.

Solution:

Bias Mitigation: Regularly audit the model for biases and correct them to ensure fairness and accuracy.

Privacy

Privacy

1.**Data Copyrights:** Traditional approaches to model training often involve the direct utilization of raw data, posing significant risks of data breaches or unintentional exposure.

2.**Generation of Sensitive Information:** Text generation models, trained on sensitive data, run the risk of generating output that inadvertently discloses confidential information. This can pose legal and ethical challenges, especially in regulated industries like healthcare or finance.

3. **Traceability:** Keep records of logs or summaries of interactions of the users with the interface.

Privacy

Example:
If I hover over a button, that might be recorded for a segment. The fact that my input is being processed (and maybe logged) is a new layer of concern

- Solution: Provide Consent Questions:

Do you want a log of all your interactions to review what happened?

- User Consent: Ensure users are aware when interacting with an AI and understand the extent of data collection.

Sustainability: Carbon Footprint

Environmental Impact **(measured by Hugging Face Code Carbon)**
Climate change is one of the greatest challenges that we are facing and reducing emissions of greenhouse gases such as carbon dioxide (CO2) is an important part of tackling this problem.

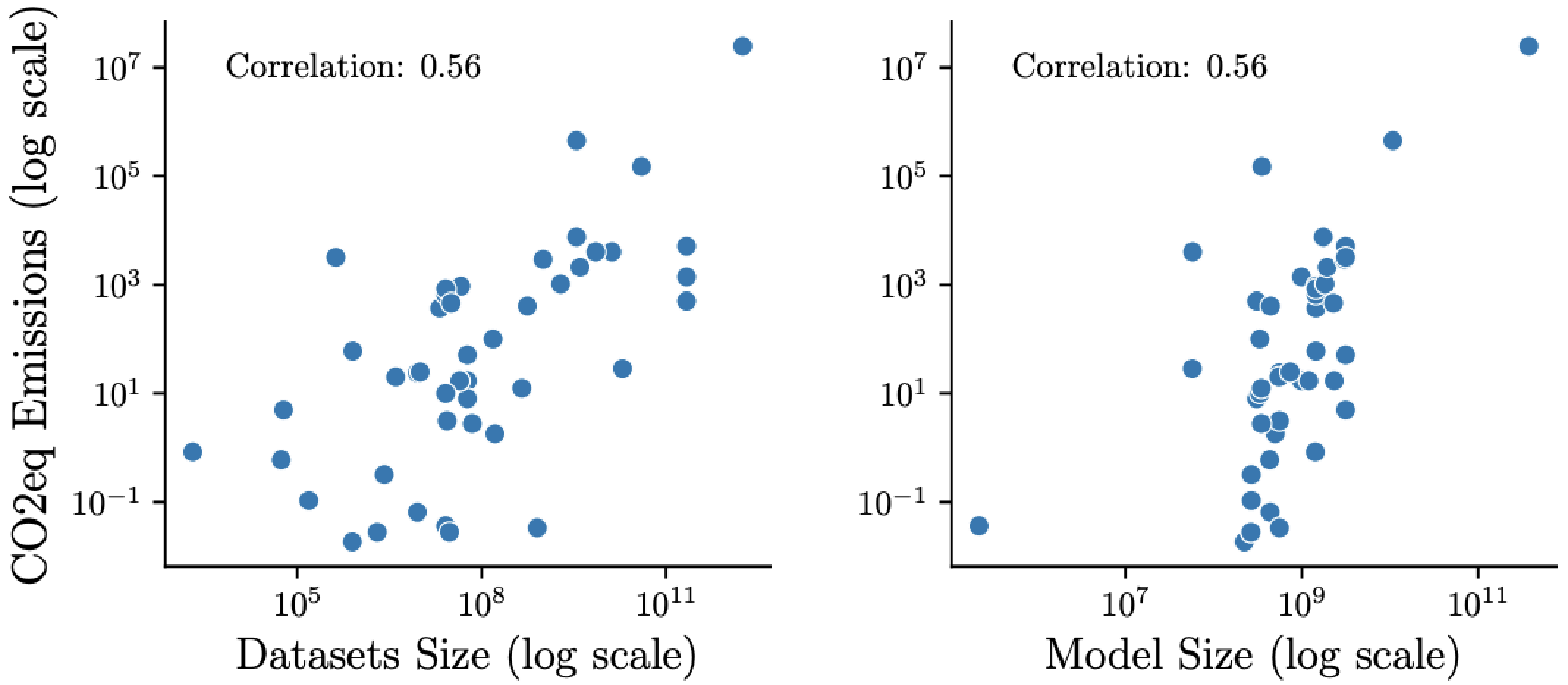


Fig. 8. Correlation between carbon emissions and dataset / model size

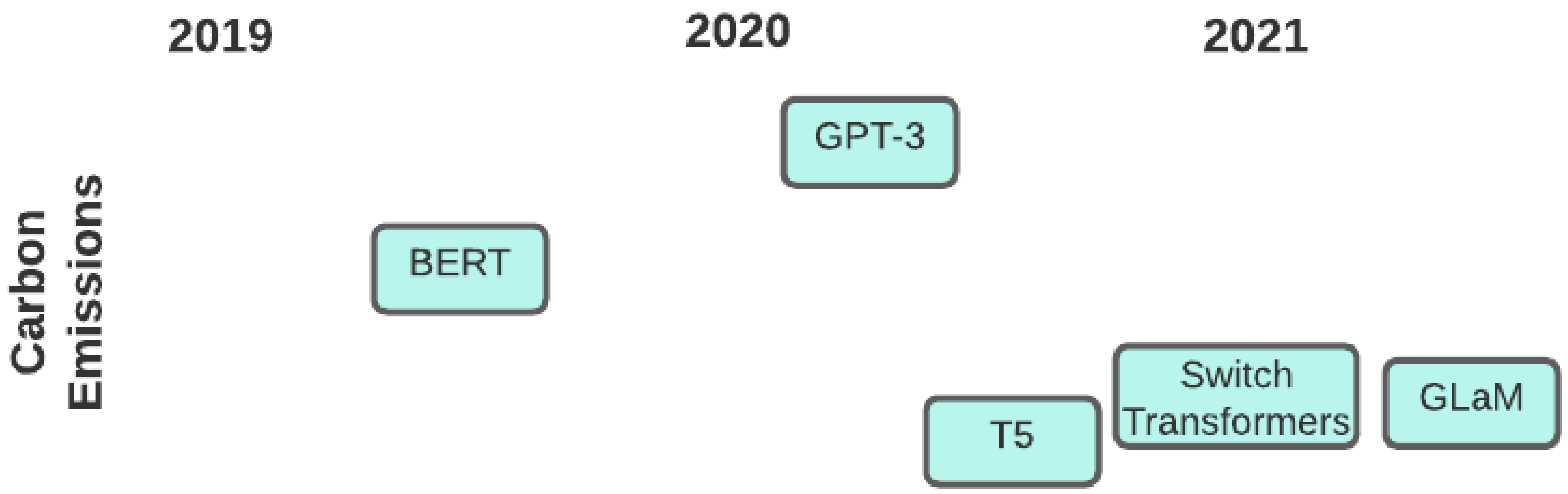
Learn more: https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/

Sustainability: Carbon Footprint

Small Language Models (SMLs):

- Do not have the same reasoning skills as large language models (LLMs).

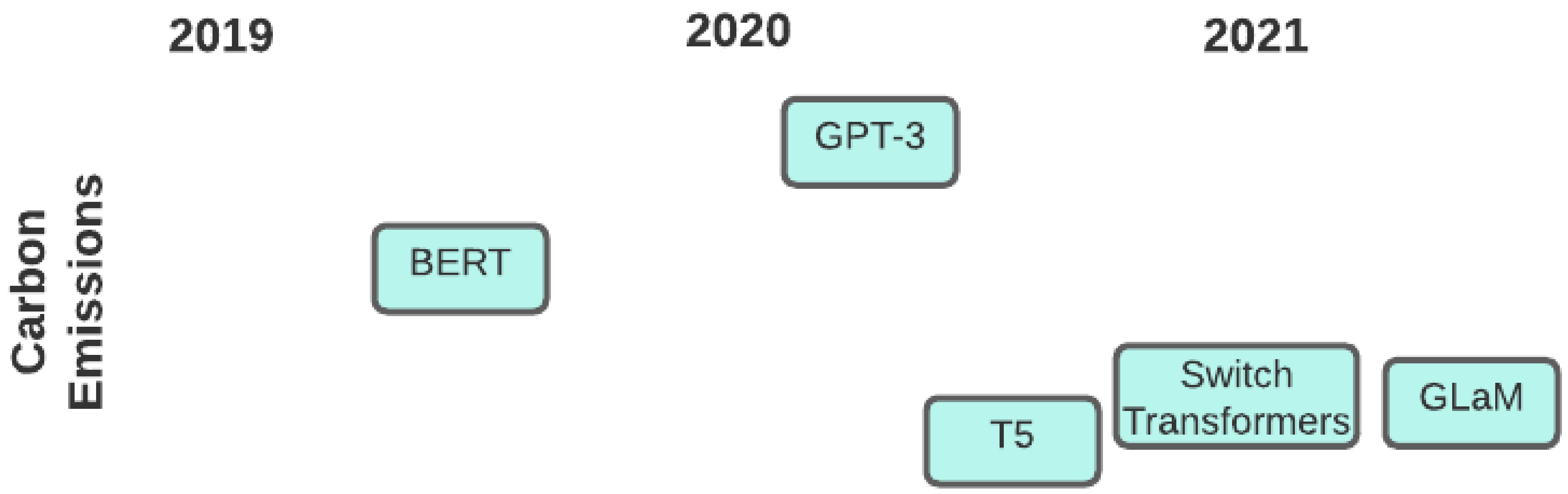
- Choosing small LLMs:
  - Increases privacy (data can be stored on personal computer).
  - Reduced carbon footprint.
  - Sufficient for common use cases.



Recent Transformer models and their carbon footprints
by HUGGING FACE.

Sustainability: Carbon Footprint

Training and deploying machine learning models will emit CO2 due to the energy usage of the computing infrastructures that are used: from GPUs to storage, it all needs energy to function and emits CO2 in the process.



Recent Transformer models and their carbon footprints by HUGGING FACE.

Sustainability: Small Language Models

- Developer decision-making: Cost to train the model
- User decision-making: Cost to use the model

UI question:
Would users want to have control over whether the NLI is using a small model or a larger one?
Would users want a running tally of energy use?

UI Solution:
Enabling the user to choose the model

Choice of model based on user's task (bigger is not always better).

Sustainability: Carbon Footprint

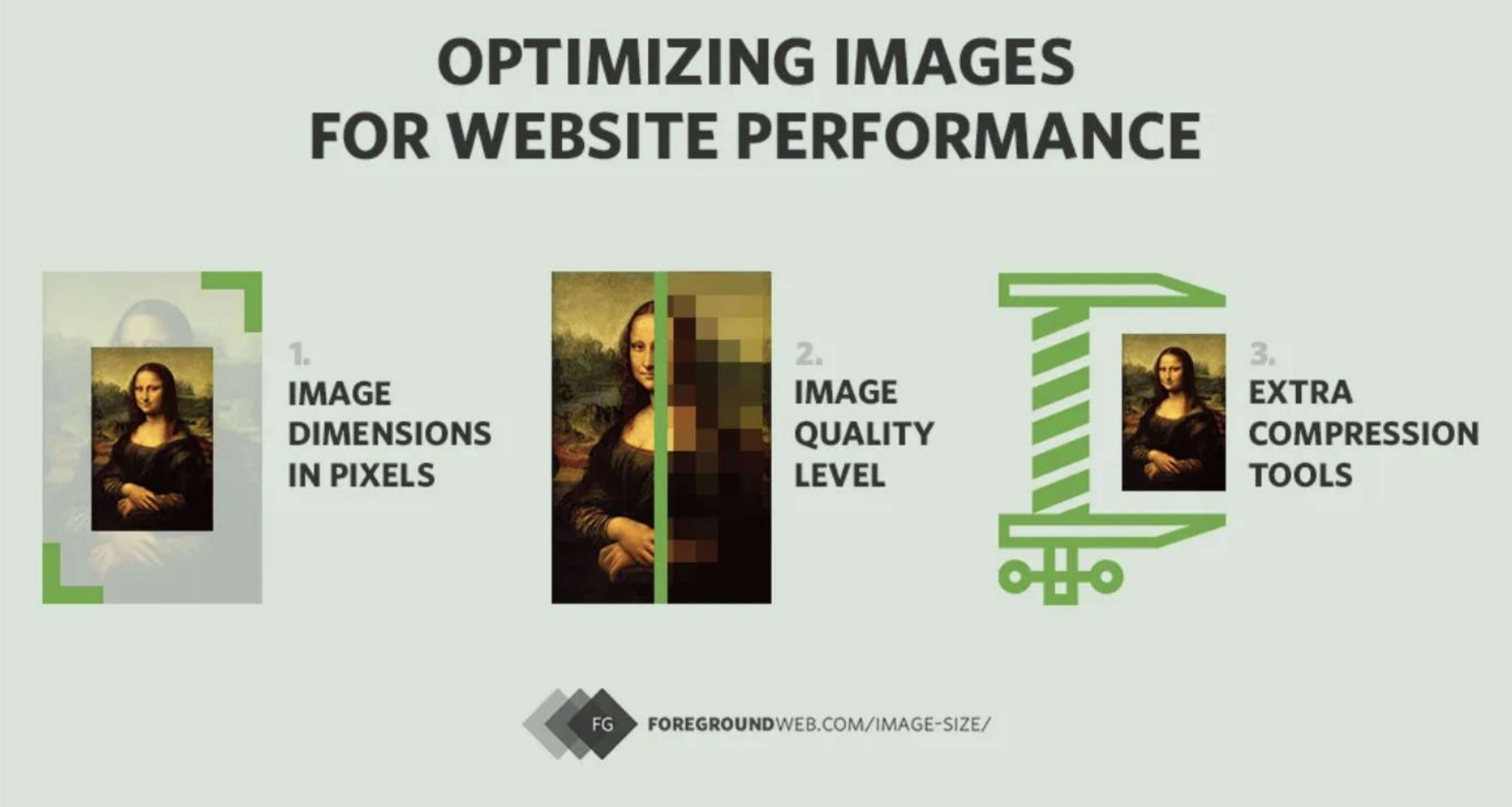
Brining LLMs to Consumer's hardware:
"The key to unlocking this potential lies in *quantization*, a technique that allows reducing the size of these increasingly large models to run on everyday devices with minimal performance degradation."

Quantization: Use smaller-precision numbers in the math of the model (eg. 32-bit instead of 64-bit.) That's a good thing, because even though the number of nodes in the model hasn't changed, the *footprint* - the size the model takes up in memory and how much compute power it takes - is smaller.

Quantization shrinks LLMs to consume less memory, require less storage space, and make them more energy-efficient

Learn more: https://www.datacamp.com/tutorial/quantization-for-large-language-models

Sustainability: Carbon Footprint



Example: Image Compression (Resolution)

Ethics in Interface Design

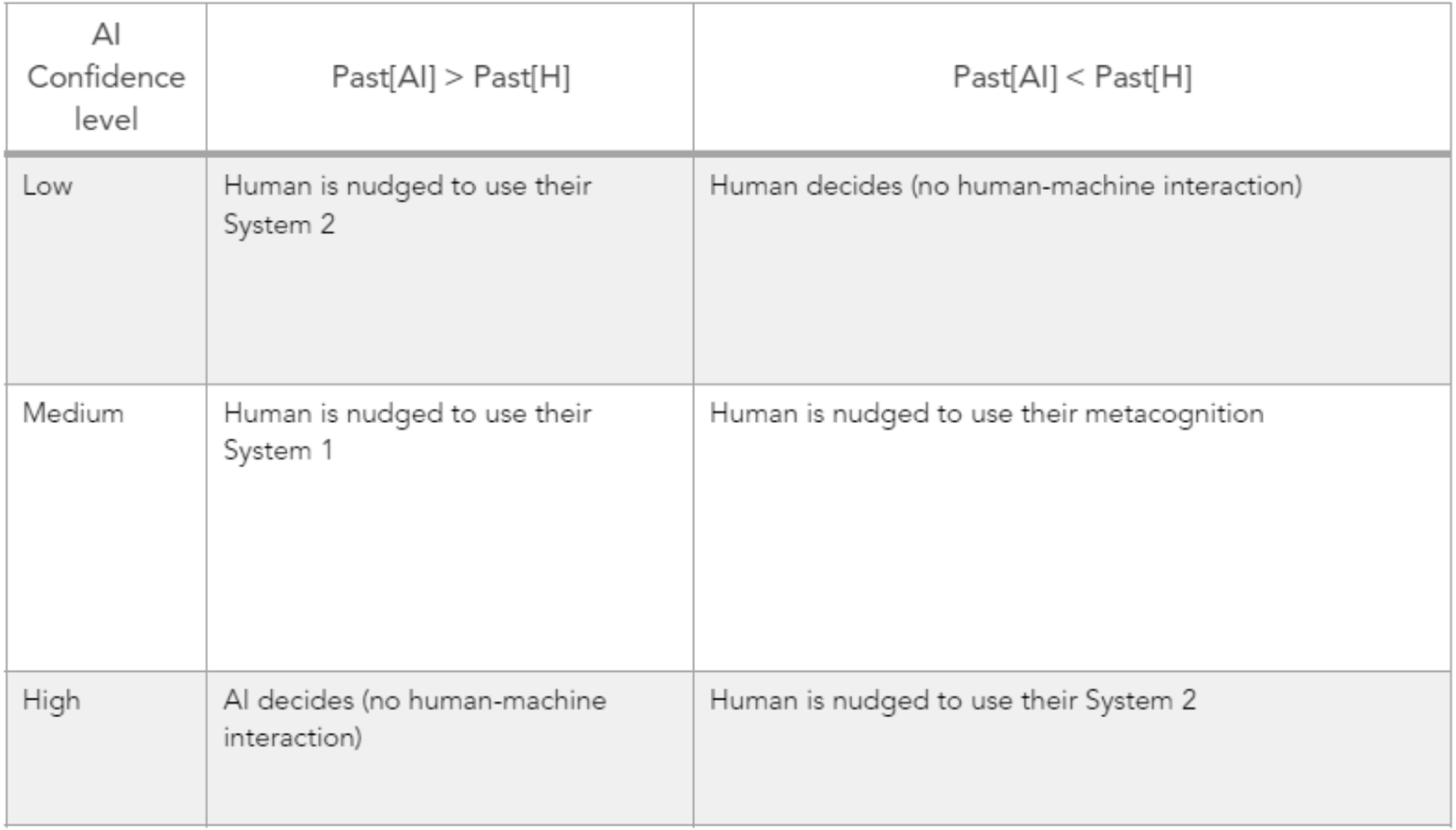
UX: UX is about how people interact with digital products like websites or apps. It’s like designing a comfortable and easy-to-use chair.

Nudge Theory: Nudge theory is a concept from psychology that says small, subtle changes in how choices are presented can influence people's decisions. Imagine arranging snacks in a way that makes the healthier options more noticeable, encouraging people to choose them.

- System 1 nudges take advantage of our quick, instinctive thinking
- System 2 nudges encourage us to engage our slow, reflective thinking.

Interface Prompts: Nudging

AI = AI decision
H = Human decision



AI Confidence level	Past[AI] > Past[H]	Past[AI] < Past[H]
Low	Human is nudged to use their System 2	Human decides (no human-machine interaction)
Medium	Human is nudged to use their System 1	Human is nudged to use their metacognition
High	AI decides (no human-machine interaction)	Human is nudged to use their System 2

Reference: https://montrealethics.ai/value-based-fast-and-slow-ai-nudging/

AI Ethics and Governance Frameworks

The frameworks encourage a multidisciplinary approach, involving diverse stakeholders in the AI system's lifecycle, from **design** to **deployment** and **monitoring**.

EU AI Act

- Regulation that categorizes AI systems into four distinct risk levels:
  - Unacceptable
  - High
  - Limited
  - Minimal.

- This system allows for a tailored approach to AI governance, reflecting the varying degrees of impact AI systems can have on society and individuals.

- More immediate, practical guideline.

The U.S. NIST AI RMF

(National Institute of Standards and Technology, AI Risk Management Framework),

- No specific risk categories
offers a structured methodology for managing AI risks.

- It focuses on the creation of trustworthy AI systems, prioritizing attributes such as
  - Reliability, Safety, Security, Resilience, Accountability, Transparency...

- Broader set of principles and a flexible process-oriented approach.

Designing Natural Language Interfaces

> “The ethical responsibility in the AI ecosystem is a collective effort that requires the active participation of all stakeholders.”

IBM