THE MERL Tech INITIATIVE

Common AI Definitions & Risks for Development & Humanitarian Actors

This list of definitions and risks associated with AI was developed by MTI based on our collective experience in the development and humanitarian sectors, and our work in the field of AI, ethical AI, and responsible digital practice, including via the Natural Language Processing Community of Practice. We used ChatGPT 4.o to help us simplify some of the definitions once we had written a first version.

Even amongst technical experts, there can be disagreement about these terms, and, with the rate of AI developing exponentially, some definitions may soon be outdated. We believe that it is a good primer at the time of writing and can be used to help teams working together have a shared understanding.

Definitions of key AI terms

Artificial Intelligence / AI

Artificial Intelligence is an overarching term that describes the use of computer programs to perform tasks that would typically require human intelligence - from learning, to problem-solving, and language understanding. AI is used as a 'catch-all' term, but can encompass any of the following scenarios:

- *When you ask ChatGPT to explain the root causes of gender digital divide in a particular country.*
- *When Google analyzes your past search habits to recommend results during landscape research.*
- *When you use spell check, Grammarly or DeepL to assist in writing a funding proposal.*
- *When you use Google maps and follow the traffic-optimized route during a research trip.*
- *When a WhatsApp chatbot uses AI to provide up to date weather forecasts for a agricultural chatbot.*

Natural Language Processing / NLP

Natural Language Processing is a subfield of AI focused on language related tasks, and describes the ability of computer programs to understand, interpret and more recently, generate human language in a way that's meaningful and useful.

Suggested citation: Amazon-Brown, I, Raftree, L, (2024) Common AI Definitions and Risks for Development & Humanitarian Actors, The MERL Tech Initiative

Definitions of key AI terms (cont'd)

The types of tasks supported by NLP and relevant to SBC chatbots include using NLP to categorise user questions and signpost them to relevant content, and using NLP to conduct sentiment analysis of self-reported impact data for Monitoring & Evaluation purposes.

Predictive AI

Increasingly (and not accurately) referred to as 'old AI' or 'traditional AI', predictive AI is the type of AI most of us were using before the GenAI boom in November 2022. It's a form of AI that analyzes data (words or numbers), and learns to make predictions, usually in service of a specific task. For example, a predictive model trained on anonymised data can learn to recognise sentences that indicate a safeguarding disclosure. A system can then be programmed to raise a flag to online moderators or redirect a survivor to appropriate resources.

Although predictive AI is not getting as much coverage because of the excitement over GenAI AI, one key advantage over GenAI is that the inputs (the data it is trained on) and outputs (the action it performs based on its prediction) are almost entirely controllable by its maker.

Increasingly, AI powered tools use a blend of 'predictive' and Generative models depending on the function required, as each model has strengths and weaknesses.

Large Language Models / LLMs

LLMs are an advanced type of NLP model built using multiple layers and billions of data points and variables. They have the ability to learn and improve from experience, even without being explicitly programmed to do so. LLMs do need human intervention, for example during the preparation of training data, or when correcting errors. Because they learn from such large data sets, LLM outputs often sound very humanlike. ChatGPT, the most widely known generative AI tool, is powered by an LLM.

LLMs can be adapted, or 'fine-tuned', using processes like Retrieval Augmented Generation (RAG) to make them more reliable or more relevant for the intended audience.

Generative AI / GenAI

GenAI is powered by advances in Natural Language Processing, specifically, Large Language Models. GenAI works by using vast quantities of training data to make predictions about the most likely next word in a sentence, in a split second. When given an instruction or asked a question, GenAI tools can generate entirely new-seeming data, whether text, images or video, that look like they could have been produced by a human.

Generative AI is completely dependent on the information it can scrape from the internet and other data sources during a particular window in time. It also means that the answers provided by GenAI reflect the biases present in the training data, reinforcing issues such as language, cultural, and gender bias.

Definitions of key AI terms (cont'd)

Prompts/ prompt-engineering

Prompts are instructions given to an AI powered tool to generate an output, for example, a response to a question. They can be written both by those designing the tool, and those using the tool.

Prompts are written by those designing AI tools to provide the model with more or less strict 'guardrails' in order to improve the quality and safety of its outputs (these are called System Prompts). For example, Anthropic, the organisation behind the GenAI model Claude, makes its system prompt available publicly to provide more transparency on the instructions behind the model.

Additional prompts are then written by the user of the tool, to elicit the specific information they are looking for. Good prompts should be clear, specific, and involve a degree of iteration (playing around with the prompts a few times to get the best results).

Prompt-engineering is the process of writing and refining prompts. One blind spot in the deployment of GenAI tools for SBC, is that we are not providing users with prompt-engineering skills to increase the efficiency of their use of these tools.

Chatbot

The term chatbot has two meanings depending on the context:

- Until recently, chatbots referred almost exclusively to a digital service, usually available via a chat interface on a web browser or instant messaging app like WhatsApp, that enables users to have a human-like conversation via text or voice. Many chatbots are not AI powered, and instead use a predetermined decision tree architecture that allows users to browse a menu of options (though they may still seem 'chatty'). These chatbots often incorporate a blend of mechanisms, for example, predetermined elements, GenAI, and predictive AI.

- Since the advent of GenAI, the term chatbot has been increasingly used to describe GenAI powered virtual assistants such as ChatGPT, Claude or Gemini.

Human in the Loop / HitL

This refers to human oversight or intervention in a mostly automated chatbot service. For example, a user's question could be handed over to a human agent if the question can't be answered with confidence by the chatbot. Handling a user back and forth between human and chatbot should be handled sensitively to manage users' expectations.

In a wider sense, HitL refers to the process of human-led monitoring and re-training that can go into the ongoing maintenance of AI models.

Definitions of key AI terms (cont'd)

Algorithm

A set of instructions or rules designed to perform a specific task or solve a problem using a computer. Algorithms are used to train the models which power AI.

Model

A representation of the inter-relations between data, created by applying an algorithm to data. Models capture patterns, structures and relationships which can then be used for predictions (as in predictive AI) or to generate entirely new data (as in GenAI). Models are the 'engine' or 'brain' driving AI tools.

Small Language Models / SLMs

SLMs are a form of GenAI that can be used to generate new text or images, but they use less data and are less complex. Crucially, they can be ideal in resource-limited contexts, such as on mobile devices. They can still perform a variety of tasks, like text classification or question answering, but are better suited to simpler tasks, for example, producing summaries in a specific format from raw data.

Retrieval Augmented Generation / RAG

LLMs can be adapted, or 'fine-tuned', using processes like Retrieval Augmented Generation (RAG) to make them more reliable or more relevant for the intended audience. To conduct RAG a database or collection of documents is uploaded. The Large Language Model (LLM)l is then programmed to fetch relevant information from this dataset to generate more accurate and context-aware responses.

Risks associated with AI

Note: this list has been presented in alphabetical order and is not a reflection of the relative importance of these risks. Different risks will be of different importance to each of us, and risk-tolerance for each should be assessed using a risk assessment exercise as part of the decision to use AI or not.

Bias

The data used to train AI models is inherently biased as it reflects the demographic composition of global internet users and AI developers (white, English-speaking, male, Western) - its responses or decisions often therefore reflect biases related to race, gender, age, sexuality, dis/ability, etc.

Data privacy

Tech companies are using our data in order to build and maintain LLMs, enriching themselves and exerting political and market power in the process. The amount of data gathered, how exactly it is used, and by whom, is often opaque, and sometimes unknown even by those who control the LLMs.

Digital divide

AI powered tools require more electricity and more airtime (data) to run on people's digital devices, and require digital skills in order to use effectively. This could contribute to a widening of the digital divide, which itself contributes to gender and other inequities.

Environmental harms

AI models require significant energy and natural resources to train, run and maintain. This includes the electricity and water required to run (and cool) data centers, and hardware leading to electronic waste. There is also evidence that the stress and displacement caused by climate events leads to increased incidences of VAW/VAC.

Inclusion & participation

Related to relevance and bias. The models used by GenAI powered tools have, for the most part, been developed without input from those they are interacting with (for example, Gender Based Violence experts, or survivors), leading to the models' responses or decisions not reflecting specific needs and realities.

Inequity

Emerging evidence suggests that the use of AI as a work tool can either make mediocre outputs appear outstanding, or widen the gap between outstanding and poor performances.

There is a fear that using AI is anti-meritocratic, and will enhance disparities already present globally.

Risks associated with AI (cont'd)

Power & patriarchy

AI is mostly developed and controlled by white men in the global north; as such its success can only perpetuate existing gender and postcolonial power imbalances.

Reliability

AI tools are programmed to take a confident tone. This means they may provide answers or analysis that seem convincing, but are actually wrong. These are referred to as "hallucinations."

Whilst developers can work to minimise and detect inaccurate responses, ultimately GenAI is a 'wild horse' whose answers can't 100% be controlled by developers. But - the same is true of humans!

Similarly, even where responses are still developed by humans, AI which simply signposts users in a specific direction can also be misleading.

Relevance

Related to bias. AI tools will provide answers based on their training data, and if this training data does not reflect the reality of the person using it (including language), the response provided or action taken will be if not incorrect, then less applicable and less appealing to the user.

Transparency & replicability

Proprietary models lack information for open and reproducible research. Once a language model is retired, a researcher cannot replicate their outcome. Even for the same model and same prompt, it is common for outcomes to change over time and researchers don't know why this happens.

Workers rights

AI models require human labour as part of the data preparation and evaluation phases. The work is often menial, repetitive, poorly paid, isolating, and unreliable. Tasks are often conducted in isolation from the wider development process, with limited opportunities for education or upskilling.

References

Abdulai, A.-F. (2025), Is Generative AI Increasing the Risk for Technology-Mediated Trauma Among Vulnerable Populations? Nursing Inquiry

Barnett A, Savic M, Pienaar K, Carter A, Warren N, Sandral E, Manning V, Lubman DI, (2021) Enacting 'more-than-human' care: Clients' and counsellors' views on the multiple affordances of chatbots in alcohol and other drug counselling. Int J Drug Policy

Crawford, K, (2021) Atlas of AI:Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Dalal, Sm Mackenzie-Hall, S, Johnson, N, (2024(Provocation: Who benefits from Inclusion in Generative AI?, ARXIV.org

Hosseini, M, Gao, P, Vivas-Valencia. C, (2024), A social-environmental impact perspective of generative artificial intelligence, Environmental Science & Ecotechnology

Lamensch, M, (2024), Generative AI Tools are Perpetuating Harmful Gender Stereotypes. Centre for International Governance Innovation

Nunez Duffourc, M, Gerke, S, & Kollnig, K, (2024) Privacy of Personal Data in the Generative AI Lifecycle, NYU Journal of Intellectual Property & Entertainment Law

Artificial intelligence Tools Offer Harmful Advice on Eating Disorders (2023), Harvard School of Public Health

Taxonomy of Human Rights Risks Connected to Generative AI, (2023), UN B-Tech