# Good Artificial Intelligence is Trustworthy & Responsible

Nabil Georges Badr
*Associate Professor & Researcher*
*Higher Institute of Public Health - USJ*

I was asked to describe what good Artificial Intelligence (AI) is, and here is what I answered:

In its second draft of the Risk Management Framework, the National Institute of Standards and Technology, NIST, dubs Trustworthy AI as Good AI.

Advances in AI capabilities have given rise to a plethora of ideas that could improve almost every facet of our economy and society, from cybersecurity and transportation to healthcare and commerce.

Several applications of AI in healthcare are focused on enabling caregivers to better care for patients in order to improve the quality of care[i]. Healthcare providers may now increase operational efficiency, precision, and the fundamental building blocks of decision-making processes thanks to the combination of AI algorithms, machine learning paradigms, and deep learning methodologies. AI is assisting medical professionals and life sciences organizations in improving early illness detection and intervention. AI technology (with machine learning capabilities) can identify disparities (e.g., housing conditions, food insecurity, transportation issues) that negatively impact the ability to find the right patients for the right trials and assist them in participating successfully by sifting through unstructured data and narrative notes. By using AI technologies, healthcare professionals and academics may more effectively assess unfair inequities and assist communities and providers in creating solutions that improve health equality by connecting them to community resources, treatment alternatives, and access to care.

## AI Risk Management Framework (AI RMF)

AI technologies are frequently applied to inform, advice, or streamline tasks in order to have a positive effect. AI systems do, however, come with a number of hazards that call for particular strategies and attention. AI systems have the potential to magnify, maintain, or worsen unfair results. Artificial intelligence (AI) systems might show emergent characteristics or have unforeseen effects on people and communities. If the data relationships that drive the behavior of the AI system cannot be adequately represented mathematically, then the current approaches of quantifying risks and negotiating the risk-benefit tradeoff are not sufficient. AI dangers might come from the data used to train the system, the system itself, how it is used, or how people interact with it. One of the biggest obstacles to implementation is cited as being a lack of transparency, since doctors need to feel comfortable that the AI system is reliable.

The AI Risk Management Framework (AI RMF) is intended to the understanding of how the contexts in which AI systems develop and implement may interact with and impact people, communities, and groups. Although opinions on what constitutes a trustworthy AI system vary, trustworthy systems share a few essential traits. Valid and trustworthy AI is reliable, secure and resilient, accountable and transparent, fair and bias is managed, it is also explainable and interpretable, and privacy-enhanced[ii].


**Figure 4:** AI trustworthy characteristics.

Source: AI Risk Management Framework, Second Draft, NIST 2022[iii]

## Explainable AI (XAI) and Trustworthy AI

Explainable AI has the potential to overcome this issue and can be a step towards trustworthy AI[iv]. Building trustworthy and explainable AI (XAI) in healthcare systems is still in its early stages[v]. Where explainability is the process by which the AI model derives its output and can be presented so that users can understand it[vi]. Explainable AI (XAI) for example is a set of tools and frameworks to help the user understand and interpret predictions made by machine learning models. Explainable AI (XAI) is a growing field that aims to make AI models more understandable.

In Healthcare, XAI aims to make AI system decision-making processes more transparent, allowing users to trust, understand, and manage AI. AI-driven models used in diagnosing diseases or suggesting treatment options often leverage XAI to help physicians understand the basis of their recommendations. Hospitals can use explainable AI for cancer detection and treatment, where algorithms show the reasoning. This is done by providing clear explanations of how AI models make decisions or predictions. Best use cases are relative to (1) Doing something more efficiently (reviewing research and data sources for clinical summarization) – and - (2) Doing something humans cannot do (early detection; Faster genomic sequencing; biomarkers; etc.) Relieve humans from tiring repetitive tasks.

## Closing thoughts

Artificial intelligence (AI) systems are socio-technical in nature, which means that their design, development, and use are the result of numerous organizational, technical, and human variables. Numerous attributes of trustworthy AI, including privacy, interpretability, bias, and fairness, are closely linked to human behavior and society dynamics. Risks and benefits associated with artificial intelligence (AI) can arise from the interaction between technical and socio-technical elements pertaining to the use, operation, and social environment of a system, as well as its interactions with other AI systems. The trustworthy nature of AI systems is complemented by their responsible use and application. AI systems are not intrinsically dangerous or harmful; rather, their potential for harm depends largely on the context in which they are used.

[i] Badr, N. G. (2022, October). Learning Healthcare Ecosystems for Equity in Health Service Provisioning and Delivery: Smart Cities and the Quintuple Aim. In The Proceedings of the International Conference on Smart City Applications (pp. 237-251). Cham: Springer International Publishing.
[ii] AI Risk Management Framework, Second Draft, NIST 2022
[iii] https://www.nist.gov/document/ai-risk-management-framework-2nd-draft
[iv] Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of biomedical informatics, 113, 103655.
[v] Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ... & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion.
[vi] Jaspers, M. W., Smeulers, M., Vermeulen, H., & Peute, L. W. (2011). Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. Journal of the American Medical Informatics Association, 18(3), 327-334.