UC Center Sacramento
Policy Brief
Volume 07 Issue 03 • February 7, 2024

AI in Healthcare: Towards Explainability, Transparency, and Responsible Governance

Ida Sim, M.D., Ph.D., Professor of Medicine and Computational Precision Health, UCSF (ida.sim@ucsf.edu), Cora Han, J.D., UC Office of the President (cora.han@ucop.edu)

Artificial Intelligence systems are not new to healthcare, but the recent boom in newer AI technologies, such as generative AI, presents new questions and challenges for its uses in healthcare. In this policy brief, Dr. Sim and Ms. Han describe the components of AI trustworthiness in heathcare settings. They also explain the importance of AI governance for upholding these principles in healthcare environments and outline recent and ongoing AI governance efforts to accomplish these crucial objectives.

Trust in Healthcare AI

Trust in AI is earned from a person or community. This requires the AI model to be worthy of trust, which is best achieved by continually demonstrating **robustness** and **reliability.**

In a healthcare context, **algorithmic transparency** implies that individuals, including both patients and clinicians, should understand how their data is being used and how AI systems make decisions. It's ideal for AI systems to be **explainable,** but with the caution that explanations of stochastic 'black box' algorithms may sound plausible but aren't true. Finally, while we traditionally have not routinely inspected the statistical models used in healthcare settings, it is important pursue **algorithmic inspectability** as part of evaluations that can help ensure robust and reliable AI.

While it is important for AI models to be algorithmically transparent, interpretable, and explainable, these criteria alone are insufficient for trustworthy AI in healthcare systems. **AI vigilance** methods are crucial for achieving trustworthy AI.

It is also important to ensure that AI systems, especially those used in healthcare, are subject to continued regulation and oversight – even after they are initially approved. This is because software performance continually changes. These changes may be intentional – such as through planned updates – or unintentional, through cases such as **data drift,** referring to changes in the distribution or relationship of key variables, or **algorithm drift**, where a system's intended use case is no longer aligned with its practical applications. Figure 1 outlines the Department of Health and Human Service's Principles of Trustworthy AI.

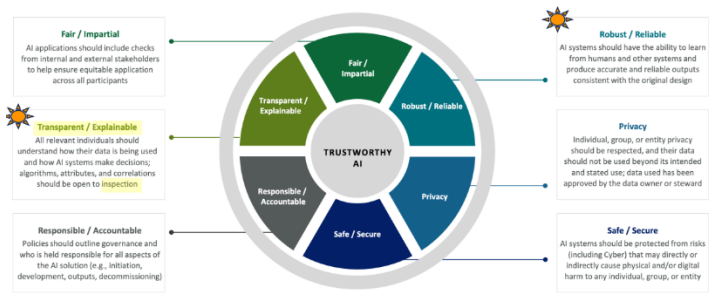

Figure 1: HHS Principles of Trustworthy AI

AI Governance

Guardrails are essential to establish trustworthy AI in healthcare settings. To this end, **AI governance** is valuable because it builds trust with patients, providers, and administrators, enables the efficient vetting and authorization of AI tools in a transparent and replicable manner, and reduces the risk of unexpected harm and reputational damage by ensuring compliance with existing laws and regulations.

President Biden's 2023 **Executive Order on AI** included several directives for the Department of Health and Human Services (HHS). These include **establishing an HHS AI Task Force** charged with developing a strategic plan for the responsible use of AI in the health sector, ensuring compliance with nondiscrimination laws, and developing several new policies and programs related to AI in healthcare, including an AI assurance policy and safety program, as well as preparing a strategy for regulating the use of AI in drug development. There are also many organizations involved in AI governance within healthcare, including the **Coalition for Health AI (CHAI)** and the **National Academy of Medicine's Health Care Artificial Intelligence Code of Conduct.**

Future Directions for AI Governance in Healthcare Settings

While there have been many efforts to develop principles and frameworks for responsible AI in healthcare settings, less has been done to create practical steps to operationalize them. CHAI and other groups have begun to address this by building consensus around ways to measure the trustworthiness of AI systems and developing evaluation criteria for each stage of the AI life cycle, specifically tailored to healthcare contexts.

It remains important to evaluate and address potential risks to trustworthiness over the AI lifecycle. Fulfilling these objectives is especially important when considering generative AI, which presents new risks to AI vigilance and also amplifies existing ones. Given the rapid emergence of new generative AI models and systems, it is especially critical to keep apace with developing laws and regulations governing these new technologies. To this end, developing AI governance – in healthcare settings as well as on the broader scale – will be an ongoing process in the years and decades to come.