The Ethics of AI in Healthcare:
Balancing Innovation with Privacy and Security

Declaring the extensive presence of AI in the healthcare industry seems to have happened in the blink of an eye. However, AI is at the pinnacle of its hype cycle, with both cynics and die-hard advocates criticizing it. Economic experts predict brisk advancement in AI-driven healthcare in the coming years. Several issues develop as a result of this exponential growth. To avoid accidental, negative aftereffects and risks caused by the use of AI in healthcare, it is vital to investigate the privacy, security, and ethical elements of AI systems. This blog will outline the privacy, security, and ethical concerns faced by AI in healthcare and offer solutions, emphasizing the need of establishing an AI-driven healthcare system that is effective and supports innovation.

1. KEY HUMAN VALUES AND ETHICAL ISSUES IN AI FOR HEALTHCARE:

Human values are frequently ignored in software engineering [1], which is employed for the vast majority of AI applications [2]. Despite the notion that consciously aligning AI with human values can result in several benefits, such as improved cancer care and patient involvement, this is not the case [3]. In contrast, when human values are violated, a slew of ethical difficulties arise, which we discuss in detail in the next subsections. We present a taxonomy of human values from the social sciences and discuss the ethical challenges that arise in AI for healthcare that match these values.

Human value	Ethical principle
Security (safety, harmony, and stability of society, of relationships, and of self)	Non-maleficence (protection from harm, precaution, prevention, non-subversion)
Self-direction (independence in thought and action; creating, exploring, being curious)	Freedom and autonomy (consent, choice, self-determination, liberty, empowerment)
	Dignity
	Privacy (protection of personal or private information)
Benevolence (preserving and enhancing others' welfare, voluntary concern for others' welfare)	Beneficence (benefits, well-being, peace, social good, common good)
	Responsibility (accountability, liability, acting with integrity)
	Trust
	Transparency (explainability, understandability, interpretability, acts of communication and disclosure)
	Solidarity (social security and cohesion)
Universalism (understanding, appreciation, tolerance, and protection for the welfare of all people and for nature)	Justice and fairness (consistency, inclusion, equality, equity, non-discrimination, respect for diversity, plurality, accessibility, redress)
	Sustainability (conserving environment and natural resources)

Table 1

Schwartz's Theory of Basic Human Values describes ten human values that have been validated by empirical investigation in over 70 countries [4]. In an examination of 1,350 recently published software engineering publications, four ideals were found to be the most commonly cited: security, compassion, universalism, and self-direction. A scoping assessment independently performed a thematic analysis of 84 AI ethics guidelines, revealing 11 major ethical principles cited throughout standards. Table 1 lists these human values as well as the ethical standards linked with them. The mapping between human values and ethical standards is arbitrary, and it is merely used to describe and conceptualize the numerous ethical challenges raised by AI in healthcare.

[1] Perera, H., et al.: A study on the prevalence of human values in software engineering publications
[2] Saleh, Z.: Artificial intelligence definition, ethics and standards
[3] Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare.
[4] Schwartz, S.H.: An overview of the Schwartz theory of basic values.

1.1. SECURITY:

1.1.1 NON-MALEFICENCE:

Non-maleficence involves minimizing foreseeable harm in terms of discrimination, violation of privacy, and bodily harm, and is cited considerably more than beneficence in current AI guidelines, suggesting it is a greater priority for AI to avoid harm than to do good [5]. Since AI for healthcare is evolving so rapidly, there is a concern that harms will only be recognised and addressed after they have occurred [6].

Safety is a key priority in AI for healthcare, especially given how few efforts are supported by empirical evidence [7]. Technical failures, such as AI chatbots that stop working properly [8], or AI initiatives that fail during a network outage[9], may cause unforeseen consequences. Furthermore, AI may lack interpersonal or cultural competency, which may impede therapeutic connections and result in unwanted psychological suffering [10].

1.2. SELF-DIRECTION:

Self-direction encapsulates a sense of personal independence, comprising the ethical principles of freedom and autonomy; dignity; and privacy. Freedom and autonomy require informed consent, disclosure of relevant information, comprehension, and voluntary participation. Obtaining consent is challenging for opaque algorithms and large datasets. Dignity involves preserving human decency and rights, and AI in therapeutic relationships may create emotional trauma for users. Privacy is a human right, but AI in healthcare raises concerns over data collection, management, and working with social media data. These issues include poor security practices, inaccurate portrayal of mental state, difficulty in anonymization, and storage of data even after users drop out.

1.3. BENEVOLENCE:

Benevolence encompasses enhancing and maintaining 'good' for oneself and others through ethical principles like beneficence, responsibility, trust, transparency, and solidarity.

1.3.1.BENEFICENCE:

Limitations of AI in contributing to individual and societal well-being raise ethical issues, particularly its impact on clinicians' decision-making, where it could miss high-risk cases and potentially become a self-fulfilling prophecy.

1.3.2. RESPONSIBILITY AND TRUST:

Accountability, transparency, and integrity are needed for building trust in AI initiatives, especially those with unclear lines of responsibility. Trust can be lost due to false results, incompetence, or misuse of public data.

---

[5] Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines.

[6] Wiens, J., et al.: Do no harm: a roadmap for responsible machine learning for health care.

1.3.3. TRANSPARENCY:

Interpretability and explainability of AI-based decisions are crucial for transparency, but 'black box' algorithms make this difficult. Disclosing AI's shortcomings, including false results and bias, is also a concern.

1.3.4. SOLIDARITY:

Vulnerable populations and those of low socioeconomic status may be excluded from AI healthcare, and its implementation may cause harm or be used for ulterior motives, such as raising insurance premiums for high-risk individuals.

1.4. UNIVERSALISM:

Universalism encapsulates a sense of appreciation towards people and the planet, comprising the ethical principles of justice and fairness, and sustainability.

1.4.1. JUSTICE AND FAIRNESS:

Justice and fairness in AI require representing diverse society, preventing discrimination against vulnerable groups, and allowing challenges to AI-based decisions. Excluding certain socio demographic groups from AI research and development could overlook their needs, perpetuate disparities, and exacerbate issues with biases in training data. Implementers must question questionable outputs, especially in high-risk scenarios like suicide prediction algorithms, as AI decisions are not infallible. The delegation of decision-making solely to machines has been criticized.

1.4.2. SUSTAINABILITY:

Sustainability involves considering the environment and minimizing the ecological footprint of AI initiatives. This is of importance for all AI initiatives, with relevant factors specific to healthcare not yet identified.

2. A FRAMEWORK FOR IMPLEMENTING ETHICAL PRINCIPLES IN AI WITHIN THE HEALTHCARE INDUSTRY:

Although there has been significant work in mapping the ethical principles in AI for healthcare, this understanding alone does not translate readily into actionable practices. To address this gap, here we present our framework for operationalising ethics in AI for healthcare across the development pipeline, as illustrated in Fig. 1. Firstly, ethical AI is a product of numerous layers of influence: the professional practices of developers and users; the governance of these individuals at an organizational level; and the regulation of individuals and organizations at a legal level. Secondly, the AI life cycle involves three major stages, which must progressively be fulfilled before the subsequent stage is initiated. These stages are as follows:

---

[7] Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines.

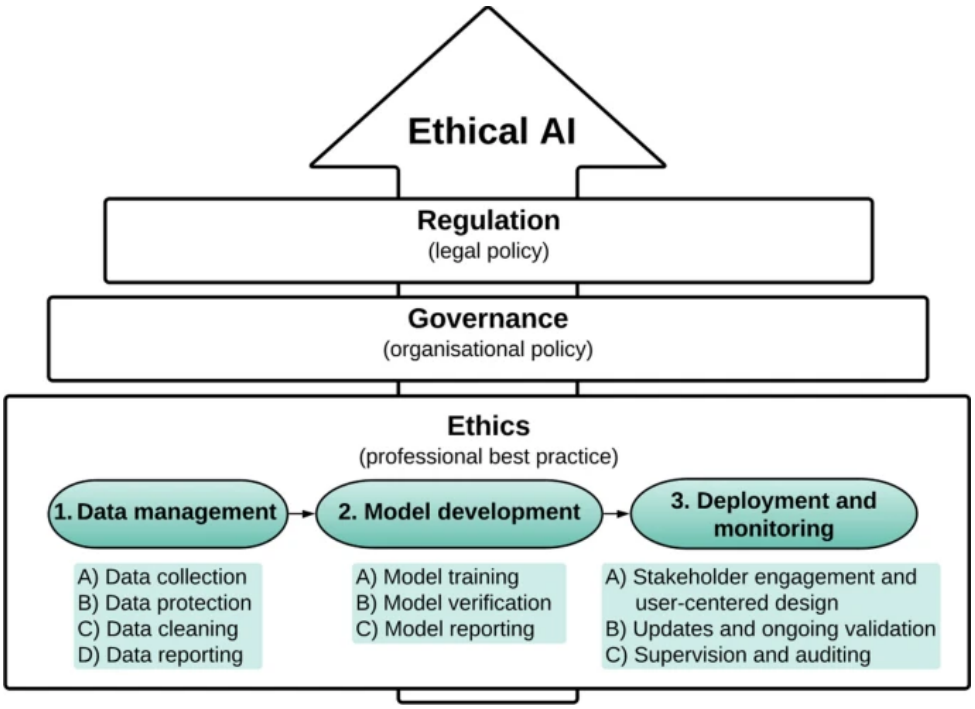
[8] Luxton, D.D., Anderson, S.L., Anderson, M.: Ethical issues and artificial intelligence technologies in behavioral and mental health

1. Data management involves: (A) data being collected; (B) data being appropriately protected with best-security practices; (C) data being cleaned (including pre-processing and augmentation where appropriate); and (D) data being reported.

2. Model development involves: (A) an AI model being trained on a dataset; (B) an AI model being verified in its performance on test dataset/s; and (C) an AI model being reported.

3. Deployment and monitoring involves: (A) a model being deployed in a real-life setting with stakeholder engagement and user-centered design; (B) updates and ongoing validation; and (C) supervision and auditing.

Fig .1



ETHICAL AI FRAMEWORK

2.1. DATA MANAGEMENT:

2.1.1. DATA COLLECTION:

Collect diverse and inclusive data, avoid biased datasets, and unnecessary data types. Respect privacy and consider deidentification. Obtain data from underserved populations and pre-process data for improved quality. Use ethical and diverse teams to address biases and raise awareness of ethical principles.

[9] Ashmore, R., Calinescu, R., Paterson, C.: Assuring the machine learning lifecycle: desiderata, methods, and challenges

[10] Amershi, S., et al.: Software engineering for machine learning: a case study.

2.1.2. DATA PROTECTION:

Protecting health data systems from attacks requires using best security practices such as de-identification and anonymization, as well as regularly updating data privacy and security measures. Differential Privacy can be used, and data sharing practices should be transparent and require explicit consent. Synthetic databases created with strong privacy features can be an alternative to sharing original data, and access to data should be clearly defined and logged.

2.1.3. DATA CLEANING:

To ensure fairness in AI, unbiased data for vulnerable groups is crucial. Biased or incomplete datasets should be avoided, and techniques such as MLClean, HoloClean, Activeclean, and Universal Cleanser can be used to clean them. Discrimination during data cleaning can be prevented with minimally intrusive data modifications and selective instances relabelling. Synthetic dataset creation and resampling can address inadequate data from vulnerable groups and tackle data imbalance, including removing poisoned samples.

2.1.4. DATA REPORTING:

Provide thorough documentation for datasets to increase transparency and help ML engineers detect issues with training data. Documentation should include motivation, composition, collection process, pre-processing, uses, distribution, maintenance, impact, and challenges for ML. For NLP, documentation should cover curation rationale, language variety, speaker and annotator demographics, speech situation, text characteristics, and recording quality. AI service FactSheets can be used to build trust in AI systems by providing information on purpose, performance, safety, and security that can be audited by AI service consumers and regulators. A template is available to describe the contexts in which AI systems are developed and deployed.

3.1. MODEL DEVELOPMENT:

3.1.1. MODEL TRAINING:

To develop responsible machine learning models, a pre-analysis plan should clearly outline goals, technical approaches, research questions, and outcomes of interest, as well as ethical and value-based constraints. Models should be trained on ethical datasets prioritizing actionable predictor variables and clinically relevant, unbiased outcomes for marginalized groups. Interpretable models are preferred for transparency and trust. Techniques like multiple classifiers, smoothing decision boundaries, or data sanitization can minimize adversarial attacks or poisoned data. Zero-sum game theory can be used to plan against attacks.

3.1.2. MODEL VERIFICATION:

Choose appropriate performance metrics for AI algorithms based on their purpose, such as prioritizing few false positives or few false negatives. Evaluate the algorithm's performance on marginalized groups, identifying poor performance using subset methods. Post-hoc classified

auditing can improve fairness, but trade-offs between fairness and accuracy must be considered and ethical justification is necessary. Other techniques like 'machine unlearning' can remove the effects of poisoned data. Test AI algorithms across multiple healthcare systems, socioeconomic groups, or age ranges with scrutiny for large-scale applications.

3.1.3. MODEL REPORTING:

Clear reporting of the model's data source, participants, predictors, and outcome variables, as well as its features and verification environment, is crucial for transparency and building trust in the technology. Documentation, including "model cards," can summarize the model's attributes, intended use cases, conditions, and application domains, and aid in explainability for stakeholders. Evaluations of the model across different cultural, demographic, or phenotypic groups should also be documented to encourage reuse. Ad-hoc and external post-hoc techniques can make models more interpretable and promote transparency.

2.3. DEPLOYMENT AND MONITORING:

2.3.1. STAKEHOLDER ENGAGEMENT AND USER-CENTERED DESIGN:

Stakeholders should be involved throughout the AI lifecycle, including during deployment. This includes knowledge experts, decision-makers, and users, with efforts to feature sociocultural diversity. There should be a clearly identified clinical problem to be addressed, and users should be appropriately trained. When integrating AI into clinical practice, the system's capabilities and limitations should be specified, and information should be provided in a context-specific manner. Clinicians should be able to disagree with and override AI recommendations. The final AI model should be accessible, and patients should be informed about the use of AI and its limitations. The language of the AI system should not reinforce unfair stereotypes, and patient preferences should be considered. The AI system should alert a healthcare professional if a patient exhibits acute high-risk behaviors.

2.3.2. UPDATES AND ONGOING VALIDATION:

To ensure the AI system stays relevant, it should have a method to be updated over time and learn from user interactions. User feedback should be taken into consideration, and a mechanism for redress should be available for unjust mistakes. Continuous evaluation of performance metrics is necessary, and a prospective clinical trial design may be considered for formal evaluation of real-life performance across diverse population groups.

2.3.3. SUPERVISION AND AUDITING:

AI systems in healthcare pose new and existing risks due to their ability to operate automatically and at scale. Ethical principles must be implemented through active governance, which includes policies, procedures, and standards. However, a "principles-based" approach alone is insufficient. An active governance approach is preferred and emphasizes regulatory influence, human-centered governance, and possibly going beyond the Westernized human rights-based approach to ethics. Ethics-based auditing can support governance through internal and external audits, algorithmic-use risk analyses, and structured processes to assess adherence to ethical

[11] The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care | Health Affairs

principles. An institutional approach is needed to restore trust in AI systems, including clarity about organizational values, red teaming exercises, third-party auditing, and communicating known or potential incidents related to the system.

CONCLUSION:

The exponential development of AI in healthcare is promising, but a naïve application of AI to healthcare may lead to a wide array of ethical issues, resulting in avoidable risks and harms. As such, there is a growing need for ensuring AI for healthcare is developed and implemented in an ethical manner. In this paper, we recognise and go beyond solutions that offer principle-based guidance (e.g. adherence to 'fairness'), adopting a solution-based framework that AI developers can use to operationalise ethics in AI for healthcare across all stages of the AI lifecycle—data management, model development, and deployment and monitoring. We emphasize solutions that are actionable (whether technical or non-technical), and therefore utilizable by AI developers. Finally, we acknowledge the growing need for 'ethical AI checklists' co-designed with health AI practitioners, which could further operationalize existing solutions into the AI lifecycle.

Written by : Bouchra AL HALMUNI

Entrepreneur, Full-stack developer, Designer, Speaker and Mentor.

Bouchra is a web developer with a vast array of knowledge in many different front end and back end languages, responsive frameworks, databases, and best code practices. She is the founder of Boxra, a creative digital agency in communication and marketing, digital specialist, website & application creation, social media, and E-commerce.

She's also the co-founder of T.A.W a pioneer African social enterprise focused on equipping women with digital, business & leadership training to thrive!

Being an avid learner herself, Bouchra wanted to share her learning journey and her knowledge that could help people specially youth by being a Mentor with many international and national organizations like Codbar, WeTech, Youth International Conclave, Garage 48, Geekle.us, JCI, DigiGirlz, Academy for Women Entrepreneurs…

She is also certified #IamRemarkable Facilitator and Women Techmakers Ambassador | Google Developers.