Human-Centered Design and Explainable AI: Building Trust in Clinical AI Systems

Kaustubh shaw[1]

[1]*City Montessori School, Gomti Nagar-1, Lucknow, 226010, Uttar Pradesh, India*

Keywords

Artificial Intelligence; Medical Ethics; Human Centred design; Explainable AI; Digital Divide; User Interface;

Abstract

With the rise of AI in healthcare, comes a challenge to trust due to the “black box” problem and algorithmic bias. This paper argues that a human-centered design approach is essential for mitigating these issues by creating transparent and fair systems by exploring how explainable AI can make logic understandable to healthcare professionals while design interventions can ensure equitable outcomes. It also provides an analysis of patient’s perspectives that reveals the key concerns about the loss of empathy, the non-negotiable need for physician oversight, and the imperative for data privacy.

Introduction

Healthcare systems globally are grappling with significant challenges including achieving the ‘quadruple aim’ (improving population health, enhancing patient experience, improving caregiver experience, and reducing costs). The global pandemic has further highlighted shortages in the healthcare workforce and inequities in access to care with many people dying because of lack of beds or oxygen cylinders in hospitals and many not even being able to make it to the hospitals because of the hefty charges. According to PWC, the average cost of healthcare in the United States is $14,570 per person in 2023. High prices for healthcare are a major cause for inequity, preventing these services from reaching the common man. Against this backdrop, the application of AI, offers potential solutions to these issues. By leveraging the abundance of multi-modal data and advancements in technologies like deep learning and cloud computing, AI is ready to make healthcare more accessible, improve

diagnoses and create a new era of personalised medicine. It is envisioned as a tool to further human capabilities, freeing up healthcare professionals to focus on the uniquely human skills of empathy and emotional intelligence, something which AI cannot imitate. However, despite its immense applications, the widespread adoption of AI in clinical practice remains limited due to the barrier of lack of trust from both clinicians and patients. For clinicians, the distrust comes from the opaque, “black box” nature of many AI algorithms, which makes their decision making processes difficult to understand and verify. They also have legitimate concerns regarding safety, accountability, and the practical challenges of integrating AI into complex real world situations. On the other hand for patients, trust is undermined by fears surrounding data privacy, the potential for algorithmic bias, and a lack of empathy in AI interactions. This collective issue is the single greatest hurdle which is preventing AI from fulfilling its potential in medicine. This paper will make the case that fostering trust and guaranteeing the moral application of AI in clinical settings requires the integration of human centred design principles and explainable AI hence demonstrating that we can develop AI systems that are not only strong but also dependable and trustworthy by giving transparency, equity, and user empowerment top priority during the design process.

The Foundations of Mistrust : Opacity, Bias & the ‘black box’ problem.

The rapid proliferation of Artificial Intelligence in various sectors, including healthcare, has no doubt brought forth remarkable opportunities but also significant challenges. One of them being mistrust in artificial intelligence systems, which in critical fields such as healthcare, originates from several key concerns, including the “black box” problem of opaque decision-making and the potential for algorithmic bias which arises from flawed data. [6] The ‘black box’ problem in AI is when a system’s internal workings are a mystery to its users meaning that its algorithm is opaque. In the medical context, users, including patients, doctors and even the designers themselves cannot understand why or how a specific diagnosis or treatment recommendation is produced by the AI. This lack of transparency introduces a tension between accuracy and explainability in turn giving rise to several critical issues that erode trust, [7]the primary concern being the fundamental lack of understanding among patients and doctors about how these predictions are made. [6] For instance, deep neural networks used in image recognition might reliably distinguish between malignant and benign tumours but offer no explanation for their judgements.

Furthermore, in patient-centred medicine, doctors are obligated to provide adequate information to patients for medical decision making. However, the opacity of a black box ai system makes it difficult for the doctor to explain the reasons behind the treatment plan. [6]
Additionally, the unexplainable nature of black box AI makes it difficult to identify and detect medical errors. These systems might make errors that a human might never make, potentially leading to serious harm [6] in a field which is built on the principle of 'do no harm', making the risks of unexplainable algorithmic errors a major concern.

Apart from opacity, AI systems carry a significant risk of amplifying existing disparities and inequalities in healthcare[3].

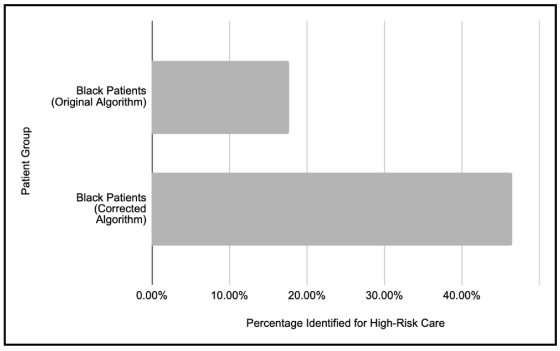

fig 1.1 : data is based on a 2019 **Science** study led by Dr. Ziad Obermeyer, as cited in an
article by **Paubox**

The issue shown in the graph, where Black patients identified for high-risk care management programs increased from 17.7% to 46.5% after bias correction, originates from machine learning models being trained on data from healthcare systems that are unjust and unequal, inherently embedding bias in the data and the resulting recommendations. Now, bias can manifest itself at various stages of the AI system's life cycle, from design and development to deployment and maintenance. Bias during the developmental stages might come from skewed data used to train the model, where it might not accurately represent the target population or important influencing factors that are not included in the data. For example, An ML model trained on X-rays found black patients experienced higher pain with similar osteoarthritis severity. This was due to the model not considering non-radiological factors such as stress, which caused pain in this population.[3] Bias may also be introduced during the human-led labelling of data where the human's prior knowledge, conceptions, can affect the process. This is known as annotation bias and it may present itself as cognitive bias (annotators' experiences influencing labelling decisions), inter-annotator bias (different humans having different interpretation or expertise levels) or confirmation bias [3] Bias may also manifest during the integration and ongoing use of an AI system through data drift, feedback loops, model decay etc. This clearly shows that the consequences of algorithmic bias in healthcare are severe and can lead to misdiagnosis, suboptimal outcomes [9]

and amplified existing inequities. This is where the question lies, how do we use Artificial Intelligence (AI) systems to their maximum potential without running into these issues? By designing systems that can make AI's reasoning clear. In other words developing Explainable AI Systems (XAI).

Designing For Transparency

We now know how important transparency is in AI systems, particularly in healthcare, as it fosters trust, enables understanding, ensures accountability, and helps mitigate biases in critical decisions. While algorithms form the computation backbone of AI, their comprehension, adoption, and trustworthiness hinge significantly on how their outputs and underlying logic are presented and integrated into human workflows through thoughtful design. [11] This perspective highlights a critical shift from a purely technical focus to a human centred approach in AI development. [15] Achieving transparency though, requires more than a strong design intention; it also demands technical approaches that make AI reasoning accessible. This is where the field of Explainable AI (XAI) comes in.

But what is XAI? In academic terms XAI refers to AI systems that are designed to provide explanations for their decisions to its users. It explains the internal processes of a model, detailing its methods, procedures, and outputs in a way that is understandable [12] to users. This transformation from an opaque "black box" to a transparent "white box" is critical for AI systems. XAI aims to increase interpretability, accountability, user trust etc. But, there lies a significant challenge - explanations often remain too technical or abstract for end-users and non technical professionals like doctors [15]. The XAI community frequently evaluates explanations from the perspective of AI or ML experts, rather than the actual users of the AI systems which leads to "lack of high-quality user-centred focus" in XAI research and a failure to assess whether explanations truly fulfill their purpose in an operational context. [15] Thus while XAI provides valuable tools for opening the 'black-box', these explanations remain limited if they are not designed for actual users. To transform these technical outputs into meaningful insights, design plays a critical mediating role.

Effective designs act as a bridge between the complex logic of AI algorithms, translating raw model outputs into actionable and comprehensible insights for human users. [11] Without intuitive interfaces and clear communication of AI's rational clinicians may struggle to understand how a diagnosis was made. [11] In medical imaging, saliency maps or heat maps are employed to visually highlight critical

regions in X-rays and MRIs that influenced an AI's classification. These visual overlays allow doctors to see precisely what parts of an image the AI focused on while making the diagnosis, thereby giving a clear "why" explanation. [12] For decision support tools, methods like SHAP and LIME, generate graphs and force plots that illustrate which patient parameters had the highest impact on a particular diagnosis.

Furthermore, presenting confidence scores alongside predictions provides a quantitative measure of the AI's certainty. [15] These interactive elements allow users to explore scenarios by adjusting input parameters or emphasising certain visual aspects, thereby helping them understand the algorithm's sensitivities and refine results to align with their clinical reasoning. [13] Ultimately, these design choices significantly impact not only clarity but also trust. The aesthetic and functional qualities of the interface - colour, layout, and interactivity - contribute to the system's perceived ease of use [11]. An example of these qualities used in real life is in AI-CDSS where visual elements like colours, icons and charts are used to convey urgency and facilitate rapid interpretation of information. Research has also shown a significant positive correlation between perceived visual features and level of trust in digital agents. Hence, by carefully designing how AI explanations are presented, designers can empower clinicians to critically engage with AI outputs, fostering appropriate trust and confidence in the diagnostic process.

In conclusion, transparency in AI is co-created by technical explanations and design interventions. While algorithms generate reasons, it is design that determines whether those reasons are visible, meaningful and trustworthy.

Designing for Fairness & Ethics

While design has a crucial role in increasing transparency, it can also help with fairness and ethics in artificial intelligence which is crucial for mitigating bias. Bias which can develop at any stage of the AI lifecycle can be tackled by, Human-Centered AI (HCAI) approach, which is a primary strategy for recognizing and mitigating these biases. This multidisciplinary collaboration between diverse stakeholders - human centred design (HCD) specialists, lawyers, healthcare workers, patients etc - ensures that AI systems are not only effective but also fair, ethical and aligned with human values. [14] HCD specialists, in particular, design and evaluate AI-based systems to be easy to use and understand, ultimately developing systems for universal access and accessibility for people with disabilities. [14] Interfaces can be designed to allow users to interact with and even modify the AI's processing. For

example, one system designed for diagnostic support in cancer allowed pathologists to refine image search results using refine by region, refine by example, refine by concept or manual input of contextual data. Additionally, Providing users with the opportunity to iteratively tune an imperfect system with their feedback can significantly improve the system's performance and user acceptance in practice. This continuous feedback loop, facilitated by design, allows for dynamic bias detection and mitigation post-deployment. In conclusion, design serves as the essential bridge that translates raw algorithmic logic into human-understandable explanations and interactive tools. By committing to diverse datasets and creating interfaces that enable transparency, parameter adjustment, and continuous feedback, designers can actively mitigate biases and promote the development of fair, ethical, and trustworthy AI systems.

The patient's perspective

The successful integration of AI into healthcare requires a great understanding of patient attitudes, which can be far more complex than a simple measure of technological acceptance. The findings from three distinct studies provide a comprehensive view of patient perceptions.

The cross-sectional survey by **Fritsch et al. (2022)** in a German hospital investigated the influence of sociodemographic factors on patients' AI perceptions. In the U.S., **Esmaeilzadeh et al. (2021)** used an experimental design to measure how individuals perceived the risks and benefits of AI for both acute and chronic conditions. Finally, **Witkowski et al. (2024)** employed a mixed-method approach to explore patient comfort with various AI-driven tasks in Florida.

Collectively, these studies reveal that patient trust is not a given but is critically dependent on addressing deeply human concerns about empathy, data, and the enduring role of the physician. The most profound source of patient mistrust originates from a concern over AI's perceived inability to provide the "human touch." As noted in Fritsch et al.'s survey, patients complained about the "missing empathy of the system," similar to the Witkowski et al. study where a significant portion of respondents expressed a "fear of losing the 'human touch' associated with doctors." This finding directly challenges the design of autonomous AI systems, suggesting that any tool meant to replace a human must first address this critical gap in perceived humanity.

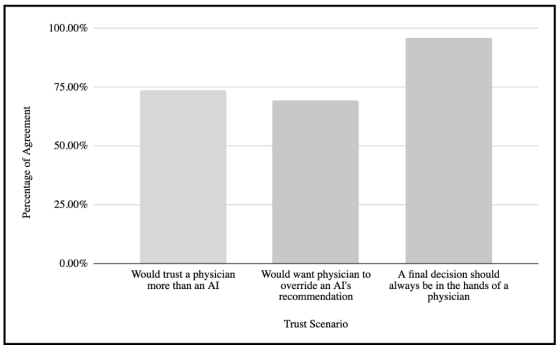

fig 1.2: Patient Preference for Physician Oversight in AI-Driven Decisions.
Fritsch et al. (2022)

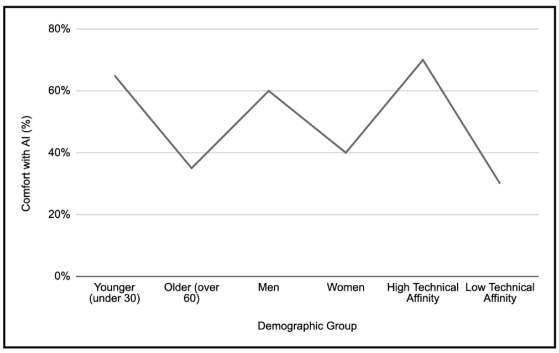

Fig 1.3 : The Digital Divide in AI Comfort and Trust.
Witkowski et al. (2024) and Fritsch et al. (2022)

The collective findings of all three studies reveal that patients are not opposed to AI but are unwilling to give control to it entirely. This sentiment is best understood through the clear demand for physician oversight. As illustrated in **Figure 1.2: Patient Trust in AI vs. Human Physicians**, based on data from Fritsch et al., patients overwhelmingly desire that their physician remains the ultimate authority in the diagnostic and treatment process. This data signals a design requirement for AI to function as an **augmenting** rather than a **substituting** technology. Patients view AI through the lens of their relationship with their doctor, not as a standalone technological product. Their willingness to accept an AI diagnosis is highly dependent on whether it is endorsed by their trusted physician. As Esmaeilzadeh et al. found, patients are more receptive to AI when it operates as a recommendation system vetted by a physician, suggesting that the human-AI partnership is the only acceptable model for patient care.

Additionally, patient comfort with AI is not uniform; it is shaped by demographic factors. As highlighted in **Figure 1.3: The Digital Divide in AI Comfort**, research consistently reveals a "digital divide" in attitudes toward AI in healthcare. Based on findings from Witkowski et al. and Fritsch et al., this graph illustrates that older patients, women, and individuals with lower educational levels or technical affinity consistently report lower comfort and a more cautious stance on AI. This divide highlights the need for a human-centered design approach that prioritizes accessibility and clear communication, ensuring that AI is built for everyone, not just for digitally savvy individuals.

In conclusion, the patient perspective demands that AI systems be more than just accurate, they must be empathetic, accountable, and secure. A truly human-centered approach to AI design must move beyond the clinical workflow to address these fundamental human concerns, ensuring that trust is not assumed but

is proactively and transparently built into the core of the system.

Conclusion

In conclusion, the successful integration of AI into healthcare is dependent on fostering trust, which is currently undermined by the "black box" problem, algorithmic bias etc. This paper advocates for a human-centered design (HCD) approach and Explainable AI (XAI) as strategies that can help us overcome these issues. While XAI provides the technical foundation for transparency, it is design that bridges the gap, translating complex AI logic into human-understandable explanations for clinicians. Patient perspectives strongly reinforce the need for AI systems that are empathetic, ensure data privacy, and maintain physician oversight as non-negotiable. Patients clearly prefer AI as an assisting tool, not as a substitute for human care.
Lastly, building dependable and trustworthy AI systems in clinical settings demands that transparency, equity, and user empowerment are prioritized throughout the design process, ensuring AI is developed for universal access and aligns with core human values.

References

[1] Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare. 2020:25–60. doi: 10.1016/B978-0-12-818438-7.00002-2 . Epub 2020 Jun 26. PMCID: PMC7325854.

[2] Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021 Jul;8(2):e188-e194. doi: 10.7861/fhj.2021-0095. PMID: 34286183; PMCID: PMC8285156.

[3] Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-Centered Design to Address Biases in Artificial Intelligence. J Med Internet Res. 2023 Mar 24;25:e43251. doi: 10.2196/43251. PMID: 36961506; PMCID: PMC10132017.(jmir-2023-1)

[4] Cecilia Panigutti, Andrea Beretta, Daniele Fadda, Fosca Giannotti, Dino Pedreschi, Alan Perotti, and Salvatore Rinzivillo. 2023. Co-design of Human-centered, Explainable AI for Clinical Decision Support. ACM Trans. Interact. Intell. Syst. 13, 4, Article 21 (December 2023), 35 pages. https://doi.org/10.1145/3587271 (3587271)

[5] van Leersum, Catharina & Maathuis, Clara. (2025). Human Centred Explainable AI Decision-Making in Healthcare. Journal of Responsible Technology. 21. 100108. 10.1016/j.jrt.2025.100108. (van leersum mathius)

[6]Hanhui Xu, Kyle Michael James Shuttleworth,
Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm”,
Intelligent Medicine,(s2.0-s266)
Volume 4, Issue 1,
2024,
Pages 52-57,
ISSN 2667-1026,
https://doi.org/10.1016/j.imed.2023.08.001.

[7] Director, S. “Does Black Box AI In Medicine Compromise Informed Consent?”. Philos. Technol. 38, 62 (2025). https://doi.org/10.1007/s13347-025-00860-1 (s13347-025)

[8] Tjeerd A.J. Schoonderwoerd, Wiard Jorritsma, Mark A. Neerincx, Karel van den Bosch,
Human-centered XAI: Developing design patterns for explanations of clinical decision support systems,
International Journal of Human-Computer Studies,
Volume 154,
2021,
102684,
ISSN 1071-5819,
https://doi.org/10.1016/j.ijhcs.2021.102684. (s1071)

[9]Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns (N Y). 2021 Oct 8;2(10):100347. doi: 10.1016/j.patter.2021.100347. PMID: 34693373; PMCID: PMC8515002. (main 1)

[10] Verganti, R., Vendraminelli, L., & Iansiti, M. (2020). Innovation and Design in the Age of Artificial Intelligence. Journal of Product Innovation Management, 37(3), 212-227. (20-091)

[11] Iris Glassberg, Yael Brender Ilan, Moti Zwilling,
The key role of design and transparency in enhancing trust in AI-powered digital agents,
Journal of Innovation & Knowledge,
Volume 10, Issue 5,
2025,
100770,
ISSN 2444-569X,
https://doi.org/10.1016/j.jik.2025.100770. (11)

[12] Saranya A., Subhashini R.,
A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends,
Decision Analytics Journal,
Volume 7,
2023,
100230,
ISSN 2772-6622,
https://doi.org/10.1016/j.dajour.2023.100230. (12)

[13] AUTHOR=Wang Liuping , Zhang Zhan , Wang Dakuo , Cao Weidan , Zhou Xiaomu , Zhang Ping , Liu Jianxing , Fan Xiangmin , Tian Feng
TITLE=Human-centered design and evaluation of AI-empowered clinical decision support systems: a systematic review
JOURNAL=Frontiers in Computer Science
VOLUME=Volume 5 - 2023
YEAR=2023
URL=https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2023.1187299
DOI=10.3389/fcomp.2023.1187299
(fcomp)

[14] Chen Y, Clayton E, Novak L, Anders S, Malin B
Human-Centered Design to Address Biases in Artificial Intelligence
J Med Internet Res 2023;25:e43251
URL:
https://www.jmir.org/2023/1/e43251
DOI: 10.2196/43251

[15] Tove Helldin, Christian Norrie,
Designing for human-centered AI—Lessons learned from a case study in the clinical domain,
International Journal of Human-Computer Studies,
Volume 205,
2025,

103623,
ISSN 1071-5819,
https://doi.org/10.1016/j.ijhcs.2025.103623.

[16]Fritsch SJ, Blankenheim A, Wahl A, Hetfeld P, Maassen O, Deffge S, Kunze J, Rossaint R, Riedel M, Marx G, Bickenbach J. Attitudes and perception of artificial intelligence in healthcare: A cross-sectional survey among patients. Digit Health. 2022 Aug 8;8:20552076221116772. doi: 10.1177/20552076221116772. PMID: 35983102; PMCID: PMC9380417.(10.1177)

[17] Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' Perceptions Toward Human-Artificial Intelligence Interaction in Health Care: Experimental Study. J Med Internet Res. 2021 Nov 25;23(11):e25856. doi: 10.2196/25856. PMID: 34842535; PMCID: PMC8663518. (jmir-2021)

[18] Witkowski, K., Dougherty, R.B. & Neely, S.R. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 25, 74 (2024). https://doi.org/10.1186/s12910-024-01066-4 (s12910