Clinical User-Centered Explainable AI for Medical Image Analysis: A Critical Technical Practice

Weina Jin[1][0000−0001−9253−4779]

Simon Fraser University
weinaj@sfu.ca

Abstract. My research centers on the problem of how to develop explainable artificial intelligence (XAI) techniques responsibly to genuinely benefit patients' healthcare and society. I utilize a user-centered approach, conduct a user study with neurosurgeons to understand clinical users' requirements for XAI, and conduct a computational evaluation to assess if the existing XAI algorithms fulfill clinical requirements. These empirical studies identify problems and gaps in the existing XAI algorithms that fail to fulfill clinical requirements. In my user study with neurosurgeons, I identify the fatal problem in XAI evaluation paradigm, and call on the XAI community to stop using the most common XAI evaluation metric of plausibility, which is unscientific, unethical, and can manipulate users. This counterintuitive phenomenon—unethical common practice in the technical AI ethics field of XAI—motivates me to dig into its root causes, and I begin to critically inspect the common assumptions, values, and standard practices that we usually take for granted. I critically analyze the root problems in the ineffective implementation of AI ethics and the vision of AI development, and propose alternative assumptions and values that have been embedded in the technical practices of the two fields of explainable AI and medical image synthesis. My research experience is a vivid example of how deep engagement with users and real-world problems can identify and intervene in the deep issues in technical practice. My dissertation identifies pathways for clinical user-centered XAI, and improves the scientific rigor, ethics, and responsibility of the technical practice in XAI and medical image analysis to genuinely benefit the public and society.

Keywords: Medical image analysis; Interpretable machine learning; Explainable artificial intelligence; Medical AI; Human-centered AI; AI ethics; AI justice; Critical technical practice

1 Motivation and the overarching research question

Medical image analysis (MIA) is the task that uses computational models to analyze and interpret information from medical images. In recent years, data-driven

predictive models (most often referred to now as artificial intelligence/deep learning (AI/DL)) have become the dominant computational models in MIA, mainly due to their capability to perform pattern recognition in high-dimensional data, such as medical images, compared to conventional machine learning techniques that rely more on hand-crafted rules or features. Despite this, the DL models have a fundamental limitation in that they are black-box models, meaning that even though we have access to the model parameters, the prediction process from input to output is still uninterpretable to humans. This is because DL models usually contain millions or more parameters and many layers of non-linear transformations. This greatly impedes their clinical applicability. Being able to explain predictions to clinical users thus becomes an essential prerequisite for applying AI models in clinical settings. The field of interpretable or explainable artificial intelligence (XAI) aims to develop techniques that can explain model predictions in human-understandable ways [1]. Explainability is one of the five AI ethics principles [2]. XAI has the potential to empower users to challenge algorithmic predictions, counter technical and institutional power, hold AI systems accountable [9], and achieve complementary performance in clinical tasks that leverage the strengths of both human and AI models [6].

Despite these great potentials, developing XAI algorithms that are applicable to clinical settings faces significant challenges. My dissertation is thus motivated to explore the following overarching research questions:

1. What are the challenges and obstacles that impede the effective development of clinical user-centered XAI techniques for medical image analysis?
2. How can we develop XAI techniques responsibly to genuinely benefit patients' healthcare and society?

2 Research Approaches and Methodologies

XAI is inherently a human-centered concept. Therefore, my dissertation research utilizes a user-centered approach to answer the research questions by collaborating with end users of XAI.

Specifically, to identify the clinical requirements for user-centered XAI algorithms, I conducted a mixed-method user study with 35 neurosurgeons [3]. I also conducted a computational evaluation of the existing XAI algorithms to understand if they can fulfill the clinical requirements [4].

In addition to the empirical studies, my dissertation also utilizes theoretical analysis to prove theorems [6] and identify limitations in technical practice [7].

3 Main Findings from My Dissertation Research

3.1 Identifying the Symptoms: Existing XAI Paradigm Is Not Clinical User-Centered and Can Mislead Users

Existing XAI algorithms fail to improve human-AI team performance [3]. We conducted a clinical user-centered evaluation with 35 neurosurgeons to assess the utility of AI assistance and its explanation on the glioma grading task.

Each participant read 25 brain MRI scans of patients with gliomas, and gave their judgment on the glioma grading without and with the assistance of AI prediction and explanation. The AI model was trained on the BraTS dataset with 88.0% accuracy. The AI explanation was generated using the explainable AI algorithm of SmoothGrad, which was selected from 16 algorithms based on the criterion of being truthful to the AI decision process. Results showed that compared to the average accuracy of $82.5 \pm 8.7\%$ when physicians performed the task alone, physicians' task performance increased to $87.7 \pm 7.3\%$ with statistical significance (p-value = 0.002) when assisted by AI prediction, and remained at almost the same level of $88.5 \pm 7.0\%$ (p-value = 0.35) with the additional assistance of AI explanation. Based on quantitative and qualitative results, the observed improvement in physicians' task performance assisted by AI prediction was mainly because physicians' decision patterns converged to be similar to AI, as physicians only switched their decisions when they disagreed with AI. The insignificant change in physicians' performance with the additional assistance of AI explanation was because the AI explanations did not provide explicit reasons, contexts, or descriptions of clinical features to help doctors discern potentially incorrect AI predictions. The evaluation showed the clinical utility of AI to assist physicians on the glioma grading task, and identified the limitations and clinical usage gaps of existing explainable AI techniques for future improvement.

Existing XAI algorithms are not truthful and useful for clinical use [4, 5]. Based on the clinical user study [3], we propose the Clinical XAI Guidelines, which consist of five criteria that a clinical XAI needs to be optimized for. The guidelines recommend choosing an explanation form based on Guideline 1 (G1) Understandability and G2 Clinical relevance. For the chosen explanation form, its specific XAI technique should be optimized for G3 Truthfulness, G4 Informative plausibility, and G5 Computational efficiency. Following the guidelines, we conducted a systematic evaluation on a novel problem of multi-modal medical image explanation with two clinical tasks, and proposed new evaluation metrics accordingly. Sixteen commonly-used heatmap XAI techniques were evaluated and found to be insufficient for clinical use due to their failure in G3 and G4. Our evaluation demonstrated the use of Clinical XAI Guidelines to support the design and evaluation of clinically viable XAI.

The most commonly used XAI evaluation criterion can mislead and manipulate users [6]. In my collaboration with neurosurgeons, I noticed that the most commonly used XAI evaluation metric, plausibility, can be problematic, as it encourages XAI algorithms to generate as plausible explanations as possible to increase users' trust, regardless of the underlying decision quality. This is analogous to bribing the auditor to produce as many positive auditing reports as possible. Our examination further reveals the consequences of using plausibility as the XAI criterion, including an increase in misleading explanations that manipulate users, deteriorating users' trust in the AI system, undermining human autonomy, an inability to achieve complementary human-AI task performance,

and the abandonment of other possible approaches to enhancing understandability. Due to the invalidity of measurements and the unethical issues, we argue that the AI community should stop using plausibility as the criterion to evaluate and optimize XAI algorithms. We also delineate new research approaches to improve XAI in trustworthiness, understandability, and utility to users, including complementary human-AI task performance.

By pointing out the fatal problem in XAI evaluation paradigm, my work safeguards the scientific rigor and ethics of the XAI field.

3.2 Identifying the Diagnosis: Power Asymmetry in the Sociotechnical System of AI

Diagnosing the roots in the unjust power structures in AI [9]. From the above phenomena (i.e., the "symptoms") of ineffective implementation of explainable AI and the problematic practice in XAI evaluation, we diagnose the root cause (i.e., the "disease") as the dysfunction and imbalance of power structures in the sociotechnical system of AI. The power structures are dominated by unjust and unchecked power that does not represent the benefits and interests of the public and the most impacted communities, and cannot be countervailed by ethical power. Based on the understanding of power mechanisms, we propose three interventional recommendations to tackle the root cause, including: 1) Making power explicable and checked, 2) Reframing the narratives and assumptions of AI and AI ethics to check unjust power and reflect the values and benefits of the public, and 3) Uniting the efforts of ethical and scientific conduct of AI to encode ethical values as technical standards, norms, and methods, including conducting critical examinations and limitation analyses of AI technical practices.

3.3 Proposing the Interventions: Encoding Just Assumptions into Technical Development Lifecycle

Improving techniques by first improving our imaginations for techniques [8]. The AI technical community usually focuses on "how" to develop AI techniques, but lacks thorough open discussions on "why" we develop AI. Without critical reflection on a general vision and purpose for AI, the community may be vulnerable to manipulation. In this work, we explore the "why" question of AI. We critically examine the prevailing vision in the AI community, which emphasizes objectives such as replacing humans and improving productivity. Our examination identifies its limitations, harms, and roots in unjust power from the for-profit sector, which impede us from envisioning more diverse alternative imaginations of AI that directly serve public interests. We then call on the technical community to diversify our collective imaginations of AI that aim to directly benefit the public, especially the marginalized and the most impacted communities. To facilitate the community's pursuit of diverse imaginations of AI, we propose the axiomatic scheme to construct alternative imaginations. We demonstrate the scheme in our construction of a new imagination, "AI for just

work," and how the alternative imagination guides the ethical and responsible development of two tasks of explainable AI and medical image synthesis. We hope this work will help the AI community to open critical dialogues with civil society on the visions and purposes of AI, and inspire more technical works and advocacy in pursuit of diverse and ethical imaginations to restore the value of AI for the public good.

Built upon the more ethical imaginations of AI, my last two works propose specific frameworks to guide the technical tasks of medical image synthesis and XAI.

Ethical medical image synthesis [7]. "How can we responsibly use or develop medical image synthesis (MISyn) techniques?"; "What are the limitations and risks of incorporating synthetic images into medical imaging workflows?". Ethical MISyn aims to answer these questions and to ensure that MISyn techniques are researched and developed ethically and responsibly throughout their entire lifecycle, which acts as the "immune system" of medical imaging technologies. To facilitate risk assessment and management when generating or using synthetic images, we provide an ethics analysis to formalize the intrinsic limits of MISyn, including lack of real world grounding and unfaithful representation that inevitably introduces distribution shifts and biases. Our case studies show that existing MISyn can fail to effectively integrate understanding of the above limits into its development, posing significant risks to patients and healthcare professionals. Ethical risks include misinformation arising from substituting synthetic images for medical images without proper disclosure, which can erode trust in the medical imaging dataset environment, and algorithmic discrimination against patients, especially the vulnerable and marginalized patient communities. To account for these intrinsic limits and operationalize our findings, we outline a series of actionable recommendations to use and generate synthetic images ethically. The recommendations can facilitate the use, development, and oversight of synthetic images for healthcare professionals and technical developers.

The Responsible XAI framework: Operationalizing justice assumptions in the technical design and evaluation practices of XAI. We propose the Responsible XAI (RX) framework that aims to improve the scientific rigor, ethics, and responsibility of XAI design and evaluation. The Responsible XAI framework consists of three aspects: the non-technical, technical design, and technical evaluation aspects of XAI. The Responsible XAI framework contributes to the XAI community by supporting and improving scientific, ethical, and responsible practices in XAI technical design and evaluation.

4 Open Challenges and Future Work

Although my research outlines the more viable and ethical pathways to develop clinical user-centered XAI for medical image analysis tasks, it is an open challenge of how the proposed pathways can effectively change the existing technical

practices in XAI and medical image analysis. Future work can explore diverse practical approaches to implement the proposed pathways and the technical framework in XAI for medical image analysis.

Bibliography

[1] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

[2] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. Minds and Machines **28**(4), 689–707 (Nov 2018). https://doi.org/10.1007/s11023-018-9482-5, http://dx.doi.org/10.1007/s11023-018-9482-5

[3] Jin, W., Fatehi, M., Guo, R., Hamarneh, G.: Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. Artificial Intelligence in Medicine **148**, 102751 (2024)

[4] Jin, W., Li, X., Fatehi, M., Hamarneh, G.: Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical Image Analysis **84**, 102684 (2023). https://doi.org/https://doi.org/10.1016/j.media.2022.102684, https://www.sciencedirect.com/science/article/pii/S1361841522003127

[5] Jin, W., Li, X., Hamarneh, G.: Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? Proceedings of the AAAI Conference on Artificial Intelligence **36**(11), 11945–11953 (Jun 2022). https://doi.org/10.1609/aaai.v36i11.21452, https://ojs.aaai.org/index.php/AAAI/article/view/21452

[6] Jin, W., Li, X., Hamarneh, G.: Why is plausibility surprisingly problematic as an XAI criterion? (2025), https://arxiv.org/abs/2303.17707

[7] Jin, W., Sinha, A., Abhishek, K., Hamarneh, G.: Ethical Medical Image Synthesis (2025), https://arxiv.org/abs/2508.09293

[8] Jin, W., Vincent, N., Hamarneh, G.: AI for Just Work: Constructing Diverse Imaginations of AI beyond "Replacing Humans" (2025)

[9] Jin, W., Zheng, E.L., Hamarneh, G.: Making power explicable in ai: Analyzing, understanding, and redirecting power to operationalize ethics in ai technical practice (2025), https://arxiv.org/abs/2510.10588