# Balancing the Scales: Efficiency Gains versus Practitioner De-skilling in Medical AI

Eshan Chawla
*Department of Computer Engineering, San Jose State University*
eshan.chawla@sjsu.edu

## I. INTRODUCTION

The introduction of Artificial Intelligence (AI) in healthcare represents a significant paradigm shift in the history of medicine. There has been a clear and rapid transition from basic rule-based systems to advanced, data-driven Machine Learning (ML) and specialized Large Language Models (LLM). This progress was possible due to the availability of vast amounts of labeled medical data and the substantial increase in computational power. These innovations promise incremental gains in both diagnostic efficiency and the accuracy of autonomous systems.

While early implementations were limited to assisting doctors with initial evaluations, recent advancements suggest a future where AI "switches seats"—moving from a validation tool to a system capable of autonomous diagnosis. The applications of these systems range from analyzing CT scans to classify pulmonary nodules in Radiology [1] to distinguishing between benign moles and malignant melanomas in Dermatology. These AI innovations have vast potential to benefit the healthcare system by automating repetitive work and making the overall system faster, more efficient, and more accessible.

However, these new technologies also introduce new moral and operational challenges that must be answered before widespread adoption of AI within the healthcare domain. A significant challenge posed by these technologies is the risk of de-skilling medical practitioners in the pursuit of improved efficiency, and the automation of predictive capabilities in these models. As AI systems evolve and become more capable, the temptation to defer the decision to the machine increases, which introduces “Automation Bias” where clinicians blindly accept the machine outputs without in-depth analysis. The lack of explanation in these models, which act as a “black box” [2]-[3], make this problem even more severe as it is not guaranteed the output will be the same each time we provide similar input which becomes a safety hazard for the patient.

This literature review investigates how the healthcare industry can leverage Artificial Intelligence based smart algorithms without making a significant impact on the cognitive ability of the medical professional. This is an important topic for Computer Engineers and Health Informatics professionals as it addresses the importance of Human Computer Interaction (HCI) dynamics that impact the success of deployments of such autonomous systems. To analyse the tension between machine efficiency and human agency, the paper investigates three key areas: the technical evolution of AI and its safe use in the field of healthcare [4]-[5], the evolution of explainability (X-AI) to detect bias and trust in AI [6], and the risks of “automation bias” which causes de-skilling in clinicians.

## II. THE EVOLUTION AND DIAGNOSTIC CAPABILITY OF MEDICAL AI

To understand the impact of AI in the healthcare ecosystem, one must understand the evolution of AI systems and the efficiency gains offered with this technology. The literature review established that we have moved past simple tools used to aid in redundant and simple work to automated systems that rival human perception. Jiang et al. [3] provided the necessary historical framework for this evolution, tracing the lineage of medical AI from the early 1970s to the modern paradigm. Initially, healthcare systems heavily relied on hard-coded conditional logic. While transparent, these systems were brittle and treated all the patients as identical, disregarding the minute personalized details required for a complex diagnosis. Jiang et al. argue that the current paradigm shift is driven by the availability of massive datasets resulting from the digitization of health records, and with this, the AI models can now learn patterns that are too complex for humans. This breakthrough helped specific biomedical AI applications to outperform human efficiency in quantifiable tasks, mostly around visual perception.

To understand why this shift leads to a “black box” problem, one must distinguish between the “feature engineering” of early ML models and the modern Deep Learning approaches. Around 2012, engineers manually selected features like circularity of a nodule or the texture contrast in an X-ray and then fed these attributes to a classifier model. This process was clearly interpretable as the input data was mostly predefined features selected for their clinical relevance. If the model failed and didn’t perform well on the testing data, that meant the features fed to the model are either uncorrelated to the target variable that we need to predict, or we have too much noise in the data that the model cannot comprehend which features to choose.

However, while implementing Deep Learning algorithms as described by Jiang et al. [3] and Pereira et al. [1], the authors abandoned this approach in favour of end-to-end representation learning. In a Convolutional Neural Network (CNN), the system does not look for attributes like “circularity”; instead, it learns from millions of abstract, high-dimensional parameters that represent the data, which cannot be comprehended by the human brain. The initial layers usually detect edges, but as we go deeper, the layers detect abstract shapes and structures of data which correlate with the pathology but have no semantic equivalent in the medical textbooks. This creates a tradeoff, as by removing human intuition from the feature selection process, we achieve a higher accuracy but lose the explainability of the model's output. These frameworks are the root cause of reduced interpretability, where the model is not hiding its reasoning, but it is rather the doctor who cannot understand this mathematical reasoning and prefers clinical explanations.

One of the most significant impacts was first seen in the domain of medical imaging. Pereira et al. [1] served as a pivotal benchmark in this domain, where they demonstrated the power of Convolutional Neural Networks (CNNs) in segmenting brain tumors from MRI images. Manual segmentation done earlier by radiologists was a very tedious task and prone to errors. By automating this process with a higher accuracy than humans, Pereira et al. showcased the first wave of perception-based medical AI. Their findings suggest that AI can significantly improve workflow efficiency, allowing physicians to focus more on interpretation than measurement and classification tasks. This represents the idea of an “assistive” role of AI, where the doctors are free from tedious tasks and can focus on higher-order thinking for diagnosis and research.

However, the scope of AI has expanded beyond medical imaging to complex predictive modeling. Miotto et al. [2] highlight the ability of Deep Learning based techniques to handle data heterogeneity with Electronic Health Records (EHRs). By integrating unstructured data like clinical notes, lab readings, and images into a unified representation of the patient, Miotto et al. argue that AI can identify predictive patterns which were earlier invisible to human practitioners. These new capabilities suggest an evolution to a form of “super-human” efficiency in prediction, capable of forecasting disease probability even before the diagnostic symptoms appear in the patient.

This trajectory of development resulted in the emergence of Generative AI. Singhal et al. [7] demonstrate this in their landmark paper on Med-PaLM which is a Large Language model made specifically for the healthcare domain. This showcased that an LLM can achieve expert level performance on the US Medical Licensing Exam (USMLE) which is more than the median scores among test takers. This proves that AI is no longer limited to narrow tasks limited to segmentation [1]. Instead, AI can now encode and retrieve vast amounts of clinical data, simulating the reasoning process by the doctor but with a more broad and vast knowledgebase than any other human.

Unlike classification models like CNN used for tumor segmentation, where the output is within a fixed set of categories, Generative AI (Gen-AI) models like Med-PaLM, which are based on the transformer architecture, operate on probability to generate the next token or word in the output sentence. This introduces a novel risk of “Hallucination”. In a clinical environment, a hallucination is not just an error and can have severe consequences. This happens because LLMs are optimized for linguistic coherence rather than factual truth. A model can generate a diagnosis that follows the perfect medical syntax and logic, but is factually linked to non-existent data.

This capability for persuasive fabrication of explanation presents a unique danger to the “de-skilled” practitioner. A novice doctor who is heavily reliant on AI may not possess the in-depth knowledge required to distinguish between a hallucination and a valid medical insight. If the AI cites non-existent studies or protocols that sound correct, the cognitive effort required to verify this is significantly higher than just accepting it. Thus, the generative nature of modern AI does not automate tasks; rather, it actively competes with the clinician's authority, requiring a higher level of vigilance from the user.

These advancements [1]-[3], [7] collectively prove that AI is a powerful tool that can simultaneously create the conditions for conflict and analyse them. If an AI can pass medical boards and see the patterns humans miss, the incentive for a human to critically evaluate the AI’s output diminishes, representing the gradual shift in the role of AI in healthcare.

## III. THE IMPERATIVE OF EXPLAINABILITY AND TRUST CALIBRATION

As the reasoning ability of these AI models increases exponentially, so does their opacity, making it even harder to comprehend the reasoning for the diagnosis, which creates a “Black Box” problem, which is a threat to patient safety and well-being. The literature suggests that the efficiency gains described earlier cannot be deployed in the real world unless there is an assurance of safety, which comes with transparency in model reasoning.

Barredo Arrieta et al. [6] defined Explainable AI (XAI) as not just a feature, but as a compulsory requirement for “Responsible AI”. They clearly distinguish between transparent models, which are made interpretable by design, and post-hoc explainability, which represents techniques that are applied on top of existing black box models to get explanations. This distinction is critical because, as Chamola et al. [8] argued, Trustworthy AI (TAI) in healthcare requires reliability, fairness, and accountability. Chamola et al. clarified that without explainability, it is impossible to audit a system for errors, making XAI the necessary bridge between high technical performance and clinical acceptance, as doctors for early adoptions need to understand the reasoning for these AI models before prescribing drugs and alternative treatments.

To solve this problem, there are several developments in the field of XAI that call for a closer examination. The current literature broadly categorizes these into “Model-Agnostic” and “Model-Specific” methods [6]. Model-agnostic methods like Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) [9] have gained popularity in the healthcare domain as they can be applied to any “Black Box” algorithm. SHAP is inspired by game theory to assign a “contribution value” to each feature, like age, blood pressure, sugar, etc, for a specific diagnosis.

However, relying on these approximations also introduces a secondary layer of risk: the “fidelity and interpretability trade-off”. A simplified explanation, like a heatmap of an MRI, is easier to comprehend for a doctor but might not accurately represent the complex, non-linear correlation the model actually used for the diagnosis. If an XAI tool highlights a region in a scan as “important”, the doctor will assume that the AI saw a tumor here, while in reality, the AI might have correlated another attribute in the region with a disease. Therefore, if the XAI tool simplifies the explanation too much to get the clinician's approval, it may provide a false sense of security, where it convinces the doctor that the AI is reasoning like a human while it is actually working on top of higher-dimensional correlations.

However, just the technical explanation will always be insufficient for a clinical audience. Rong et al. [4] introduced a critical nuance that explainability is subjective. Through a survey of user studies, the authors contend that explainability is a “human-centric” quality rather than a quantitative measure. This explains that an explanation is effective if and only if the reader finds it meaningful. This also aligns with the findings by Markus et al. [10], who surveyed the relationship between explainability and trust, where they found that technical explanations (saliency maps, etc) often fail to establish trust if they do not align with the diagnosis of the clinician, which depends on their existing medical knowledge. Markus et al. argue that for XAI to be effective, it has to communicate in the language of the doctor rather than the computer scientist.

This leads to a crucial concept of “Trust Calibration”, which was introduced by Schmid and Wiesche [11]. They argued that the goal is not to maximize trust, but rather to calibrate it. Over-trust leads to automation bias, where a doctor accepts a suggestion given by AI that is wrong, assuming that the system is generally accurate and knows more about the topic than the doctor themselves, while under-trust reduces the adoption of the technology by the industry. According to Schmid and Wiesche's framework, they suggest that the ethical use of AI requires an interface that encourages doctors to question the AI when it is uncertain. This concept of calibration is the most significant point of the review; without this, the efficiency of the machine will degrade the vigilance of human healthcare workers.

## IV. SYSTEMIC RISKS: ALGORITHMIC BIAS AND PRACTITIONER DE-SKILLING

If trust is not calibrated properly, the healthcare system faces a dual threat of practitioner de-skilling and the amplification of algorithmic bias. This section synthesizes how the reduction in human oversight allows for systematic errors.

Ahmad et al. [5] address the core tension of this paper, where they warn that “uncritical reliance on uninterpretable AI outputs could degrade diagnostic and reasoning skills”. They draw a link from the lack of transparency to de-skilling. If a system acts as an oracle giving out solutions without explanations, the clinician's job transitions from diagnosis to a data entry clerk. Over time, this lack of cognitive practice degrades their ability to function independently with no support, a danger that is acute in high-stress environments where efficiency will always be prioritized over extensive analysis.

This danger of a de-skilled workforce is magnified when the AI tools start to fail. Bolukbasi et al. [12] established the technical roots of this risk in their work on word embeddings, providing that machine learning models often absorb and amplify the societal stereotypes present in the training data, like gender bias. In a medical context, any bias could lead to different standards of care. Hence, if a clinician is de-skilled and suffers from automation bias, they won't notice when the AI acts biased and gives an unfair diagnosis.

This risk has evolved with modern technology. Ravulu et al. [13] investigate how Reinforcement Learning from Human Feedback (RLHF) which is a technique to finetune models like Med-PaLM [7] can introduce new forms of bias based on the subjective feedback of human raters. If an LLM is used by a clinician for decision support, these biases can scale across the industry spreading discrimination. Fletcher et al. [14] provide a global perspective on this problem, where the models trained on western data are inappropriate for other populations. These papers reveal that a "de-skilled" clinician [5], suffering from "over-trust" [8]. Using a "biased" LLM [11], [13], creates a perfect mix for medical errors where the human will not be capable enough for correcting these machines.

## V. MITIGATING DE-SKILLING THROUGH INTERACTION DESIGN

If de-skilling is caused by a reduction of cognitive load on the practitioner, the solution would be "Cognitive Forcing Functions" (CFFs) which are design choices made in a system to deliberately increase the difficulty of a task to force the user to engage in analytical reasoning. In the context of Medical AI, a frictionless interface where the AI simply displays the answer is dangerous.

Instead, the researchers propose "Human in the loop" architectures where the AI withholds its diagnosis until the physician has entered a preliminary hypothesis. Alternatively, the system could provide the relevant data (Highlight important factors for its reasoning) without providing a diagnosis until the doctor diagnoses the same. By introducing AI as an evaluator for the doctor rather than the patient's diagnosis, the system will challenge the doctor's decision-making skills rather than revealing the most probable diagnosis directly. This will make the doctor cognitively active while working and make the Human-AI relationship more collaborative, making sure that the human plays a crucial role in medical diagnosis.

## VI. SUMMARY AND CONCLUSION

This review examined the tension between efficiency and the de-skilling of clinicians in the healthcare industry due to medical AI via four interconnected themes. First, the review established that the massive leaps in AI systems in healthcare are driven by smarter algorithms and frameworks like Deep Learning [2], CNNs [1], and increasingly capable LLMs powered by the vast amounts of gathered biomedical datasets. Second, it highlighted that these gains are unsafe without transparency and explainability, where XAI [6] and human-centric design [4] are required to perform "trust calibration" [11] and prevent automation bias. Third, the review exposed the risks involved with a reduced cognitive engagement by clinicians and the risks of de-skilling [5], making them vulnerable to inheriting biases [13], [14] of the systems they use. Finally, the review proposed "Cognitive Forcing Functions" (CFFs) as an essential design choice while developing applications for the healthcare industry, positioning AI as a collaborative evaluator rather than an oracle.

A significant gap is identified in this review where there is a lack of studies on the cognitive retention of physicians using AI tools. While current research warns of de-skilling theoretically, there is little empirical data measuring how a doctor's diagnostic accuracy changes without AI assistance after years of dependency. Future work should focus on CFFs [15] interaction designs that force clinicians to pause and think before accepting the AI recommendations, as by keeping them engaged, we can achieve the benefits of efficiency of AI without losing the expertise of the practitioner.

## REFERENCES


[1] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," in *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240-1251, May 2016, doi: 10.1109/TMI.2016.2538465, [Online]. Available: https://ieeexplore.ieee.org/document/7426413

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," in *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236-1246, Nov. 2018, doi: 10.1093/bib/bbx044, [Online]. Available: https://academic.oup.com/bib/article/19/6/1236/3800524

[3] F. Jiang et al., "Artificial intelligence in healthcare: past, present and future," in *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230-243, Dec. 2017, doi: 10.1136/svn-2017-000101, [Online]. Available: https://svn.bmj.com/content/2/4/230

[4] Y. Rong *et al*., "Towards human-centered explainable AI: A survey of user studies for model explanations," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2104-2122, Apr. 2024, doi: 10.1109/TPAMI.2023.3331846, [Online]. Available: https://www.computer.org/csdl/journal/tp/2024/04/10316181/1S2UjfiwnDO

[5] N. Ahmad, A. Julaihi, and P. Rajalingam, "From data to diagnosis: Rethinking clinical decision support with explainable AI," in *Computer*, vol. 58, no. 8, pp. 130-135, Aug. 2025, doi: 10.1109/MC.2025.3572653, [Online]. Available: https://ieeexplore.ieee.org/document/11104189

[6] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," in *Inf. Fusion*, vol. 58, pp. 82-115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103

[7] K. Singhal et al., "Large language models encode clinical knowledge," in *Nature*, vol. 620, no. 7972, pp. 172-180, Aug. 2023, doi: 10.1038/s41586-023-06291-2, [Online]. Available: https://www.nature.com/articles/s41586-023-06291-2

[8] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," in *IEEE Access*, vol. 11, pp. 78994-79015, 2023, doi: 10.1109/ACCESS.2023.3294569, [Online]. Available: https://www.researchgate.net/publication/372536351_A_Review_of_Trustworthy_and_Explainable_Artificial_Intelligence_XAI

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774, [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[10] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," in *J. Biomed. Inform.*, vol. 113, p. 103655, Jan. 2021, doi: 10.1016/j.jbi.2020.103655, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420302835

[11] A. Schmid and M. Wiesche, "The importance of an ethical framework for trust calibration in AI," in *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 27-34, Nov.-Dec. 2023, doi: 10.1109/MIS.2023.3320443, [Online]. Available: https://ieeexplore.ieee.org/document/10268318

[12] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.* (NIPS), Barcelona, Spain, 2016, pp. 4349–4357, [Online]. Available: https://arxiv.org/pdf/1607.06520

[13] C. Ravulu, R. Sarabu, M. Suryadevara, V. Gummadi, and M. D. Kidiyur, "Mitigating bias in reinforcement learning from human feedback for large language models," in *Proc. Int. Conf. AI x Data Knowl. Eng. (AIxDKE)*, Tokyo, Japan, 2024, pp. 70-73, doi: 10.1109/AIxDKE63520.2024.00019, [Online]. Available: https://ieeexplore.ieee.org/document/10990073

[14] R. R. Fletcher et al., "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," in *Front. Artif. Intell.*, vol. 3, p. 561802, 2021, doi: 10.3389/frai.2020.561802, [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2020.561802/full

[15] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making," in *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, Art. no. 188, pp. 1–21, Apr. 2021, doi: 10.1145/3449287, [Online]. Available: https://dl.acm.org/doi/10.1145/3449287