Towards Contextually Sensitive Informed Consent in the Age of Medical AI

Mahera Sarkar
Newnham College, University of Cambridge

Informed consent is a fundamental aspect of medical ethics, empowering patients to engage in their healthcare decisions. However, the advent of medical AI introduces new challenges, particularly contextual bias, which can undermine informed consent. This paper explores strategies for contextually sensitive informed consent in the UK healthcare system, addressing biases related to gender, ethnicity, and age. It critiques existing informed consent guidelines, highlighting their inadequacy in handling AI's complexities and biases. A novel four-part framework is proposed: enhancing AI literacy among healthcare professionals, implementing dynamic risk communication through "Model Facts" labels, providing patient-centric risk interpretation using electronic health records, and establishing legal and ethical safeguards to support clinicians. This framework aims to ensure that informed consent remains robust and meaningful in the age of medical AI, ultimately promoting equitable and patient-centred care. The paper emphasises immediate improvements to informed consent processes to complement long-term efforts to mitigate contextual bias in AI, contributing to ongoing debates and proposing practical solutions for integrating AI into healthcare ethically and effectively. Future research should focus on refining this framework and exploring its applicability across different healthcare systems and cultural contexts.

Keywords: Contextual Bias, Medical AI, Informed Consent, Healthcare Ethics

Introduction

Informed consent is a cornerstone of medical ethics (Kadam, 2017), acting as a vital mechanism to protect patient safety and ensure the legitimacy of doctors' actions (Wang *et al.,* 2024). Traditionally, it has empowered patients to actively participate in their medical decisions during interactions with physicians (Iserson, 2024). The introduction of medical AI presents new challenges to this concept, extending it to decisions made by algorithms and vast datasets, with varying degrees of success (Mittelstadt, 2021).

Medical AI has the potential to significantly enhance healthcare outcomes (Price, 2019). However, it also poses a substantial threat to informed consent practices due to contextual bias, a phenomenon where AI algorithms demonstrate differing performance or accuracy in diagnoses and treatment recommendations across diverse patient populations (Mittermaier *et al.,* 2023). This bias originates from the use of clinical trials and health studies involving mainly white and predominantly male subjects to train medical AI models (Mittelstadt, 2021). Consequently, patients not or underrepresented by this demographic face unequal healthcare quality and experiences (Cohen, 2020). In the UK, where minority groups and women make up 20% and 51% of the population respectively, such disparities can have severe consequences. While most existing efforts have focused on improving the representativeness of training data to mitigate contextual bias (Cohen, 2020), these are long-term solutions that do not address the immediate safe use of medical AI (Price, 2019). To resolve this problem, this paper explores short to mid-term strategies to facilitate a contextually sensitive approach to informed consent in medical AI. It investigates manifestations of contextual bias concerning gender, ethnicity, and age, using empirical studies to demonstrate their real-world impact. The discussion of these demographic features serves as an example to highlight other forms of contextual bias such as socio-economic status and the specific needs of transgender patients. Additionally, it examines the functions and limitations of informed consent models within medical ethics. As pre-existing discussions have focused on the US healthcare system, this paper seeks to expand these debates by focusing on the UK medical landscape. To achieve this, it draws upon academic articles to discuss the challenges contextual bias poses to informed consent, relevant UK legislation and NHS consent guidelines. By drawing on the

discussion in the preceding sections, the paper proposes a novel framework for contextually sensitive informed consent that can be integrated into everyday medical interactions involving AI. This model enhances traditional practices by incorporating considerations of patients' gender, ethnicity, and age, ensuring that recommendations given by AI systems are tailored to the individual patient. The paper aims to create a practical framework that has the potential to inform deliberation around the use of informed consent in medical AI administered in the UK. This includes enhancing transparency about the mechanisms of AI systems and their potential biases, refining communication methods with patients, and providing clinicians with guidance on how to tailor discussions based on individual patient profiles. In doing so, it advances debate on using informed consent practices to address contextual bias within medical AI, ensuring technological progress does not compromise patient care.

1. Challenges to Informed Consent in Medical AI

Medical AI is an advanced data-driven technology that collects and analyses individuals' health information for the administration of treatment and to support the wider functioning of healthcare services (O'Brien *et al.,* 2022). The complexity of these systems can be a major barrier to patients' understanding of medical procedures (Wang *et al.,* 2024), potentially reducing consent frameworks to administrative formalities rather than meaningful ethical engagements, and exposing patients to new risks (Astromskė, *et al.,* 2021). This section outlines several challenges medical AI poses to traditional informed consent models before focusing specifically on contextual bias, which this paper hopes to address.

A primary difficulty is the inherent opacity of machine learning algorithms in AI systems (Grote *et al.,* 2020). Often described as "black-boxes", their decision paths can be difficult to decipher even by their developers, complicating the task of clearly explaining and validating their functions (Iserson, 2024). This obscurity weakens clinicians' ability to assure patients about the reliability of medical AI, thereby compromising effective informed consent. Additionally, the level of detail given in explaining how AI converts data into outputs varies with the audience; the same explanation given to a data-scientist will differ from that given to a patient (Grote *et al.,* 2020). This variation can result in information being either overly complex or overly simplified, both of which are detrimental to informed consent. The introduction of medical AI also threatens the collaborative aspect of informed consent. For instance, the Watson for Oncology system, an AI-assisted decision system developed by IBM, prioritises treatments based on maximising patient lifespan (Jie *et al.,* 2021). However, the value set driving these rankings is not specific to individual patients, meaning it may not align with their specific preferences (McDougall, 2019). This discordance creates a risk of medical decision-making reverting to a paternalistic model, where AI recommendations are seen as definitive, potentially overlooking the wishes of the patient.

2. The Challenge of Contextual Bias

Having established various complexities medical AI introduces to the practice of informed consent, this paper will now focus on contextual bias. Notable literature on this issue includes Price's (2019) article '*Medical AI and Contextual Bias*', which highlights the translational disconnects in deploying medical AI across different resource settings and patient demographics, resulting in imprecise treatment recommendations for some population groups. Additionally, Cohen's (2020) *'Informed Consent and Medical Artificial Intelligence: What to Tell the Patient'*, has been particularly inspirational for this paper as it raises the following question: if algorithms deliver suboptimal treatment recommendations for certain patient demographics, should informed consent look different in such cases? Cohen largely dismisses this option, concluding that modifying informed consent is not a viable long-term solution to contextual bias as it fails to address the underlying systemic factors. However, this perspective potentially underestimates the benefits of adapting informed consent processes to temporarily alleviate the challenges of contextual bias. This paper advocates for adjusting informed consent

procedures during this interim period. It uses the UK as a case study to scrutinise the limitations of current consent practices and guidelines in the context of AI-based healthcare. Its goal is to identify meaningful ways to refine these procedures to better handle the challenges posed by contextual bias. By proposing a framework that can supplement existing NHS guidance, the paper aims to foster more responsible medical practices until the broader, structural factors that cause contextual bias have been resolved.

3. Understanding Contextual Bias in Medical AI

Contextual bias in medical AI, as described by Price (2019), refers to the tendency for algorithms to systematically produce unfair or inaccurate outcomes when translated to different contexts. This poses a notable threat to healthcare systems, particularly in their provision of care to diverse patient populations. This section explores three manifestations of contextual bias – gender, ethnicity, and age – using empirical studies to demonstrate their real-world impact. Through these examples, the effects and severity of contextual bias, as well as its potential to undermine the fairness and efficacy of medical practices, including for other demographic features, are illuminated.

Fairness in healthcare is a multidimensional concept that extends beyond resource allocation, encompassing the ethical obligation to provide non-discriminatory care based on the unique characteristics of patients across various demographics (Ueda *et al.,* 2024). This principle, rooted in medical ethics and codified in legislation such as the World Medical Association's Geneva Declaration, risks being eroded by contextually biased AI systems. These algorithms, prone to providing suboptimal diagnoses and treatment recommendations to specific patients (Price, 2019), can worsen pre-existing health inequities and hinder efforts to achieve equitable access to healthcare as stipulated in Article 35 of the EU Charter of Fundamental Rights.

3.1. Gender

The extensive and diverse implications of contextually biased medical AI are initially explored through the lens of gender bias. This bias arises from historic neglect of sex-specific biological differences (Cirillo *et al.* 2020), resulting in discrepancies in research representation and subsequent diagnosis and treatment. For instance, although coronary heart disease is the leading cause of death among women, it is often misdiagnosed due to the predominance of male-centric clinical trials and diagnostic criteria. Additionally, 67% of cardiovascular device testing is conducted on male patients, despite women being the most likely beneficiaries. Moreover, recent findings by the American Heart Association reveal that only 17% of cardiologists correctly identify women as being at a greater risk of heart disease than men (Daugherty *et al.* 2017). Similarly, medications such as zolpidem pose higher risks to women due to differences in drug metabolisation (Cirillo *et al.* 2020), yet dosages are frequently adjusted for patient size without considering sex differences (Norori *et al.* 2021). Medical AI tools intended for disease screening may also perpetuate gender biases due to being trained on datasets that encode false, sexist assumptions. This is evident from a study conducted by UCL, which found that these tools missed 44% of liver disease cases among women compared to 23% among men (Greaves, 2022). As these tools are adopted on a larger scale, their predictive value may limited by the absence or misrepresentation of women (Norori *et al.,* 2021), exacerbating gender inequalities or potentially giving rise to new forms of discrimination (Mittelstadt, 2021).

3.2. Ethnicity

A second form of contextual bias involves ethnicity, which is described as a collective identity that draws upon several characteristics, including biological features (Salway *et al.,* 2014). Ethnicity-based biases largely arise from the inaccurate grouping of minority ethnic populations within medical testing, disregarding their diverse health outcomes (O'Brien *et al.,* 2022). This oversight is apparent in melanoma screening algorithms, where predominantly white datasets lead to misdiagnoses among patients with different skin tones. Similarly, AI systems used in the detection of diabetic retinopathy have been found to exhibit a strong divergence in performance, achieving a diagnostic accuracy of 73% for light-skinned patients compared to 60.5% for dark-skinned patients (Ricci *et al.,*

2022). Moreover, the intersectionality of ethnicity with other factors heightens this issue, as highlighted by researchers at MIT, who revealed considerable disparities in AI classification accuracy (Krasniansky, 2019). Their study found that the three most popular AI programmes used by healthcare providers incorrectly classified more than 30% of dark-skinned women as displaying cancerous moles, compared to less than 1% of light-skinned men. As AI systems are increasingly integrated into healthcare processes, it is crucial to collect data from across ethnic groups and to ensure it possesses sufficient breadth to differentiate between demographics (O'Brien *et al.,* 2022).

3.3. Age

The final type of contextual bias explored here concerns age. Ageism represents an implicit bias rooted in age-related prejudice and discriminatory practices against older people (Chu *et al.,* 2023). The concept of digital ageism refers to how AI systems may produce, sustain, or amplify systemic processes of ageism. Chu (2022) identifies a contributing factor to this bias as the tendency to group older adults into broad categories, such as "60+", which starkly contrasts the finer granularity applied to the categorisation of younger age ranges. The pandemic worsened this issue, prompting the UN to note a blatant lack of data on older persons due to inappropriate data collection methods and the exclusion of those over 50 from health surveys (Stypińska *et al.,* 2023). This oversimplification contributes to health professionals' limited understanding of optimal treatment plans for older adults, increasing the risk of missed diagnoses and mortality (van Kolfschooten, 2023). A study by Neal (2022) further illustrates this issue, revealing that 40% of older breast cancer patients receive primary endocrine therapy instead of surgery, the recommended option, due to age-based assumptions made by clinicians.

Addressing contextual bias in medical AI is critical for upholding the NHS' commitment to patient-centred care. As AI begins to assume roles akin to healthcare providers, it is imperative to hold it to comparable standards of ethical conduct. Just as physicians are expected to be attuned to the diverse backgrounds and needs of individual patients (Kempt *et al.,* 2022), AI systems should tailor their advice accordingly. This section has demonstrated the negative consequences contextually biased medical AI can have for patients and the need for effective mitigation strategies. By developing a contextually sensitive model of informed consent, this paper aims to ensure equitable treatment for all patients, combatting the effects of contextual bias until more representative training datasets become available.

4. An Examination of Informed Consent: Functions and Limitations

Insufficient data, technological illiteracy, and inconsistent standards in AI usage within healthcare lead to notable gaps in accurately assessing the risks of misdiagnosis or inappropriate treatment for patients during diagnostic procedures (Astromskė *et al.,* 2021). The modification of informed consent standards represents a tentative solution that could mitigate some of the challenges that arise from contextually biased medical AI. In order to work towards a framework, this section will first discuss the functions and limitations of traditional informed consent models.

It is widely recognised that a thorough practice of informed consent requires flexibility to address multiple objectives (Hall *et al.,* 2012). These include the legal goal of protecting patients' rights, the ethical goal of supporting autonomous decision-making, the administrative goal of providing efficient healthcare and the interpersonal goal of building the trust needed to proceed with therapeutic interventions. At present, the individualisation of informed consent, where physicians tailor their advice and disclosure to specific patients, is required in several areas of medical practice. It largely applies to clinical trials and requires researchers to provide prospective patients with information in an understandable format and to accommodate any additional support needs they may have (GMC, 2013). This paper wishes to extend the personalisation of this process beyond standard medical contexts to encompass medical interventions involving AI and to mitigate the effects of contextual bias. In doing so, it hopes to enable patients to make decisions that align with their unique characteristics and

circumstances, thus enhancing the quality and relevance of care they receive.

The UK Supreme Court's decision in *Montgomery v Lanarkshire Health Board* (2015) established that clinicians must inform patients of material risks and reasonable alternatives during medical procedures (Burr *et al.,* 2023). However, the ruling does not compel doctors to tailor this information to individual patient risk factors. As a result, the informed consent process often fails to meet the specific informational needs of patients and appears more focused on protecting doctors from legal action than on genuinely empowering patients (Astromskė *et al.,* 2021). This concern becomes more pronounced with the integration of medical AI systems in healthcare. As previously discussed, deficiencies in the representativeness of training data may result in poor performance for certain patient populations and give rise to contextually biased AI (Cohen, 2020). Given that doctors are merely the end-users of this technology, they may not always have a detailed understanding of its operating mechanisms or its propensity for bias (Wang *et al.,* 2024). This creates a risk of them providing patients with inaccurate information about proposed medical interventions. Considering that one of the purposes of informed consent is to ensure treatments reflect the ends desired and chosen by patients (Hall *et al.,* 2012), such misinformation threatens to erode the legitimacy of the consent given.

This section has highlighted how the integration of medical AI in healthcare necessitates a revaluation of informed consent practices. Traditional models, while effective for governing interpersonal relationships, fall short in addressing the unique challenges posed by contextually biased AI decision-making. This paper advocates for a tailored approach to informed consent that focuses not only on legal compliance and physician protection but also on empowering patients through bespoke risk communication.

5. A Critical Analysis of Existing Informed Consent Guidelines

Having discussed the ethical considerations of informed consent, this section will now critically analyse two existing guidelines used by the NHS, the Department of Health's *'Reference Guide to Consent for Examination or Treatment'* and the British Medical Association's *'Ethics Toolkit for Consent and Refusal by Adults with Decision-Making Capacity'*. These guidelines are fundamental to informed consent practices within the UK healthcare system, setting standards that are routinely applied in a variety of medical settings. This examination highlights how these guidelines do not offer sufficient protection to patients from the risks of contextually biased medical AI before proposing amendments in the subsequent section that can be integrated into new guidance specifically tailored to medical AI.

A notable weakness in both these frameworks is their failure to specifically mention AI. Whilst their contents have been successfully applied to other medical technologies, AI introduces complexities that are fundamentally different from such tools (Davenport *et al.,* 2019). The British Medical Association's (2024) guidance emphasises that doctors should share information about the purpose of the investigation or treatment, details and uncertainties of the diagnosis, and the probabilities of success amongst other points. However, this does not account for technological complexities introduced by AI such as contextual bias or the lack of interpretability of algorithmic decision-making (Celi *et al.,* 2022). Additionally, these guidelines do not explicitly address how informed consent should consider variations in demographic features such as gender, ethnicity, and age, which are critical given that these factors can significantly influence the accuracy and reliability of medical AI. When these frameworks are applied to AI, "uncertainties of diagnosis" can assume vastly different meanings, and often involve probabilistic outcomes that may not be transparent or easily understandable for either physicians or patients (Krishnan *et al.,* 2023). This risk is exacerbated by the fact that AI systems are prone to contextual bias, potentially leading to differential treatment outcomes across diverse groups (Mittermaier *et al.,* 2023). Such disparities are particularly problematic because they may not be evident at the individual patient level. A physician treating one patient at a time may not realise that the AI system's diagnosis or

treatment recommendation is influenced by biases inherent in its training data (Liu *et al.,* 2023). This issue is compounded by the reality that the scope of datasets used to train AI systems are not always viewable or known to the healthcare providers using these technologies (Daneshjou *et al.,* 2021).

Another weakness in the consent frameworks outlined by the Department of Health and British Medical Association is their treatment of material risks and the requisite knowledge healthcare professionals must possess in the context of AI. The Department of Health (2009) guidelines state that for consent to be valid, a health practitioner must inform the patient of any material risks, defined by the British Medical Association (2024) as physical risks that a reasonable person in the patient's position would be likely to attach significance to, or that a doctor reasonably believes that the particular patient would find significant. Although these definitions are comprehensive for traditional procedures, they are inappropriate for shielding patients from potential harm caused by contextually biased AI. Within medical interactions involving AI, determining what constitutes a material risk requires understanding not just the immediate risks of a procedure but also the broader implications of algorithmic decisions (O'Brien *et al.,* 2022). Contextual bias, which has the potential to compromise the reliability and fairness of medical decisions, certainly qualifies as a material risk for patients. However, the current guidelines lack specificity in guiding clinicians on how to identify and communicate these risks, particularly the subtleties of contextual biases, to patients. This omission is critical as the legitimacy of patient consent hinges on their understanding of these risks (Astromskė *et al.,* 2021). When patients are unaware that recommendations from an AI system may be skewed due to biases in its training data, their consent is not fully informed. This calls into question the validity of consent obtained as well as the adequacy of these existing frameworks in safeguarding patients against the potential harms of contextually biased AI.

Finally, while this paper is mainly concerned with protecting patients from incorrect treatment recommendations, these guidelines are also unable to suitably shield physicians from the legal and ethical complexities that arise from contextually biased AI-based tools. The legal standard of care, applied to the physician's professional duties in the process of informed consent, requires a full understanding of the medical treatment. Consequently, the Department of Health (2009) framework states that if healthcare professionals fail to obtain proper consent and the patient subsequently suffers harm as a result of treatment, this may be a factor in a negligence claim against them. For physicians, explaining how contextual bias may influence the AI system's recommendation is a complex task (Mittelstadt, 2021), which is not suitably supported by the current consent guidelines. Without explicit instructions on what to disclose and how to navigate these potential harms, physicians are at risk of inadvertently failing to provide complete information, leading to future legal ramifications (Terranova *et al.,* 2024). This places an undue burden on individual doctors to interpret and communicate complex biases without a standardised framework or support (Wang *et al.,* 2024), further calling into question the adequacy of existing frameworks.

The analysis of NHS informed consent guidelines reveals several shortcomings in addressing the challenges posed by medical AI, particularly contextual bias. Although current frameworks are suitable for traditional medical practice, they fail to account for the complexities introduced by AI, putting both patients and healthcare professionals at risk. Consequently, there is a pressing need to revise and expand these guidelines to ensure comprehensive protection for patients and adequate support for physicians in managing AI-driven medical decisions, thereby fully upholding the principles of informed consent.

6. Enhancing Informed Consent for Medical AI: A Context-Sensitive Approach

This section introduces an original framework consisting of four distinct components, each of which are designed to address a particular aspect of the informed consent process in an AI-integrated healthcare environment. By establishing a new framework of informed consent that is contextually sensitive, this paper

envisions elevating consent procedures to a robust tool for patient empowerment instead of a mere contractual mechanism.

6.1. Comprehensive AI Literacy

The first part of the framework focuses on enhancing healthcare professionals' understanding of medical AI by embedding AI education into the medical curriculum. It involves providing foundational knowledge on the technical, ethical, and practical aspects of medical AI (Krive *et al.,* 2023), crucial for addressing the opaque nature of AI systems (Ng *et al.,* 2023). This knowledge will empower healthcare professionals to communicate more effectively with patients about AI, enhancing the informed consent process. By improving AI literacy, clinicians will be able to critically evaluate AI tools, understand their limitations, and identify potential biases, especially those pertaining to gender, ethnicity, and age. This approach seeks to equip medical professionals not to become AI developers but competent users able to interpret AI tools in clinical settings (Mangalji *et al.,* 2019), and is in line with recommendations made by the Royal College of Physicians (Kimiafar *et al.,* 2023). With a robust understanding of AI, healthcare professionals can better navigate the risks of exacerbating healthcare inequalities due to contextually biased systems (Wood *et al.*, 2021). Implementing a comprehensive AI education programme faces challenges, including a lack of faculty with AI expertise and logistical barriers within existing curricula (Krive *et al.,* 2023). To address these, medical schools could look towards developing core curricula that define AI competencies essential for healthcare professionals. This would held in identifying and training educators who possess adequate knowledge and skills in AI applications relevant to clinical practice, ensuring effective and relevant AI education in medical training (Ng *et al.,* 2023).

6.2. Dynamic Risk Communication

The next step in the framework addresses the challenge of keeping healthcare professionals and patients updated on the risks associated with AI-driven medical decision-making. This approach involves creating an adaptable communication process, ensuring that all parties are aware of any changes in the risks or performance of AI models over time. Specifically, this would help to address the limitations in current informed guidelines that do not account for the evolving nature of AI technologies and the associated risks. A key aspect of implementing dynamic risk communication is the development of Model Facts labels, a concept currently employed in the US (Sendak *et al.,* 2020). These labels are akin to nutritional labels on food products (Licholai, 2023), providing essential information about an AI model's performance, including the demographic representation of training and evaluation data, and guidelines for their appropriate use in clinical settings (Sendak *et al.,* 2020). They serve to communicate critical information about AI models in a concise and understandable format, enabling physicians and patients to make collaboratively informed decisions on how and when to incorporate AI insights into clinical care, thereby mitigating contextual bias. To implement this, healthcare organisations need to establish a system for regularly updating and disseminating these labels (Alharbi *et al.,* 2023). This process would ideally require a central authority continuously monitoring AI models, evaluating their performance in real-world settings, and updating the labels as new data becomes available or as the model evolves. The labels must include information on model performance within the local population, highlight variability of the quality of medical predictions between different demographic groups, any changes in the model functioning, and the specific context in which the model is validated to work. This approach, like the previous component, seeks to enhance the transparency and understanding among healthcare professionals and helps mitigate the effects of contextually biased AI by making clinicians aware of the limitations of models they use. This leads to better-informed clinical decisions and in turn bolsters informed consent processes. However, barriers to implementation include potential information overload for healthcare professionals, the need for ongoing training to understand and interpret the Model Facts labels, and the logistics of regularly updating and disseminating these labels. Overcoming these requires collaboration between healthcare providers, AI developers, and regulatory bodies

to ensure that the information provided is relevant, accurate and actionable. The creation of a designated NHS Model Facts Assessment Unit would further alleviate this. By integrating dynamic risk communication into actual informed consent practices, it provides a mechanism for healthcare professionals to stay informed about the AI tools they use, thus empowering them to communicate risks more effectively to patients and make better-informed medical decisions.

6.3. Patient-Centric Risk Interpretation

This component is arguably the most direct response to the threat contextually biased medical AI poses to patients. It builds on established principles of personalised risk communication, advocating for providing patients with individualised information about the specific risks and benefits of AI-assisted recommendations (Han *et al.,* 2013). This addresses variations in AI performance that correlate with a patient's ethnicity, gender, age, and other factors often neglected in standard risk communication (Noul, 2024). The implementation of this step involves using Electronic Health Records, real-time patient-centred records that function as digital versions of patients' paper charts, to inform patients about how an AI system's output may be influenced by their unique health and demographic profile, predicting and explaining potential biases (Sokhack, 2023). This method goes beyond general explanations about AI functionalities and focuses on how its decision-making might exhibit biases when applied to their specific case. The effect of this is to create a more transparent informed consent process that is tailored to each patient's circumstances. While the previous component, dynamic risk communication, focuses on keeping healthcare professionals and patients informed about general updates in AI model risks and performance, this step concentrates on individualised communication. It requires healthcare professionals to convey personalised risk information in a manner that is understandable to the patient, potentially using proven tools such as customised printed materials, visual aids, or interactive media (Green, 2011). This approach upholds the true nature of informed consent as an instrument that enables patients to make their own health-related decisions (Astromskė *et al.,* 2021). Challenges in implementing this include the abstract nature of risk information and the time constraints of clinical practice (Han *et al.,* 2013). Despite these barriers, this remains a worthwhile initiative that represents a significant step towards countering the one-size-fits-all approach often seen in healthcare, particularly in the deployment of medical technologies (Noul, 2024).

6.4. Legal and Ethical Safeguards

The final part of the framework aims to protect physicians by establishing clear standards and guidelines for obtaining valid informed consent for use of medical AI. The traditional legal standard of care necessitates that physicians have a full understanding of all medical treatment and care options to effectively inform patients (Astromskė *et al.,* 2021). However, the complexity of medical AI introduces a higher level of difficulty in understanding and explaining these systems and their potential biases, often placing an unreasonable burden on healthcare professionals (Wang *et al.,* 2024). Instead, this paper recommends defining new standards that detail the necessary level of AI understanding for different roles within healthcare. These include *consumers*, clinicians who use AI tools in patient care, *translators*, who act as intermediaries between AI developers and clinical practitioners, and *developers*, who are responsible for the technical development of AI tools (Ng *et al.,* 2023). Consumers, forming the majority of the clinical workforce, must understand how to select and apply tools effectively and be equipped to discuss AI usage with patients within the informed consent process. Translators must ensure that AI tools are properly validated and integrated into clinical settings, making certain they are practical and safe for patient care. Developers, often with a background in both medicine and computer science, must ensure the efficiency of medical AI and work to reduce biases within them. By differentiating between these tiers, the exact duties of clinicians become clear and should be codified by regulatory bodies such as the Department of Health to give rise to corresponding legal and ethical responsibilities. The maintenance of material or physical risk comparisons is another critical aspect of these safeguards (BMA, 2024). These should be

assessed in relation to AI Model Labels, electronic health records, and the patient's values, ensuring that treatment recommendations uphold principles of autonomy and that patient preferences drive decision-making (McDougall, 2019).

This section has outlined a four-part framework to enhance informed consent in AI-integrated healthcare, addressing the specific challenge of contextual bias. The first component, AI Literacy, equips healthcare professionals with core knowledge to understand and communicate the intricacies of AI to patients. Second, Dynamic Risk Communication, which seeks to introduce Model Fact labels for medical AI, ensures healthcare interactions allow for AI's evolving nature, maintaining informed consent as a continuous process. Third, Patient-Centric Risk Interpretation, directly addresses contextual bias by customising risk information to the individual patient's background, ensuring informed consent is not only comprehensive but also personalised. Finally, the framework incorporates Legal and Ethical Safeguards, which offer a structured approach to protect both patients and physicians. Collectively, these components move towards a more robust medical environment that remains patient-focused in the face of technological advancement.

Conclusion

The central research aim of this paper has been to explore how modifications to informed consent can address the challenges posed by contextual bias in medical AI, specifically focusing on the UK healthcare system but with implications for global practices. Unlike the perspectives offered by Cohen (2020), who advocates for long-term solutions such as reducing dataset biases, and Price (2019), who discusses the systemic nature of bias in AI deployment, this paper emphasises practical enhancements to informed consent procedures to mitigate contextual bias in the short to mid-term. In doing so, it extends and refines the debates initiated by these scholars, suggesting that immediate changes to informed consent practices can substantially complement long-term strategies. Moreover, by creating a four-part framework, this paper contributes a structured approach that actively engages with the complexities posed by medical AI. While this framework cannot solve the structural problems that give rise to contextual bias, it serves as both a response to the identified deficiencies in current practices, and an example for future adaptations in diverse healthcare settings worldwide. Future research should refine and explore implementation strategies for this framework as well as its applicability and adaptability in different national contexts and healthcare systems, which each have their own guidelines and cultural norms. Such initiatives mark a crucial step towards a future where medical AI not only advances healthcare outcomes but does so in a manner that is just, empathetic, and patient-centred.

References

Alharbi, A. *et al.* (2023) 'Factors Influencing the Implementation of Medicine Risk Communications by Healthcare Professionals in Clinical Practice', *Research in Social and Administrative Pharmacy*, 19(1), p. 50.

Article 35 - Healthcare (2015) *European Union Agency for Fundamental Rights*.

Astromskė, K. *et al.* (2021) 'Ethical and Legal Challenges of Informed Consent Applying Artificial Intelligence in Medical Diagnostic Consultations', *AI & Society*, 36(2), pp. 511, 512, 517.

British Medical Association (2024) *Guidance for Doctors on Patient Consent*, p. 8.

Burr, N.E. *et al.* (2023) 'Individualised Consent for Endoscopy: Update on the 2016 BSG Guidelines', *Frontline Gastroenterology*, 14(4), pp. 273-274.

Celi, L.A. *et al.* (2022) 'Sources of Bias in Artificial Intelligence that Perpetuate Healthcare Disparities — A Global Review', *PLOS Digital Health*, 1(3), p. 12.

Chu, C. *et al.* (2022) 'Digital Ageism: Challenges and Opportunities in Artificial Intelligence for Older Adults', *The Gerontologist*. Edited by S. Meeks, 62(7), p. 950.

Chu, C. *et al.* (2023) 'Age-Related Bias and Artificial Intelligence: A Scoping Review', *Humanities and Social Sciences Communications*, 10(1), p. 2.

Cirillo, D. *et al.* (2020) 'Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare', *Digital Medicine*, 3(1), pp. 2, 4, 7.

Cohen, I. (2020) 'Informed Consent and Medical Artificial Intelligence: What to Tell the Patient?', *The Georgetown Law Journal*, 108(6), pp. 1464, 1468.

Davenport, T. *et al.* (2019) 'The Potential for Artificial Intelligence in Healthcare', *Future Healthcare Journal*, 6(2), p. 95.

Daneshjou, R. *et al.* (2021) 'Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms', *JAMA Dermatology*, 157(11), p. 1363.

Daugherty, S. *et al.* (2017) 'Implicit Gender Bias and the Use of Cardiovascular Tests Among Cardiologists', *Journal of the American Heart Association*, 6(12).

Department of Health (2009) *Reference Guide to Consent for Examination or Treatment*, pp. 5, 13.

General Medical Council (2013) *Good Practice in Research and Consent to Research*.

Green, N. (2011) 'Artificial Intelligence and Risk Communication', *AI and Health Communication*, (Papers from the AAAI 2011 Spring Symposium).

Greaves, M. (2022) Gender Bias Revealed in AI Tools Screening for Liver Disease, *UCL News*.

Grote, T. *et al.* (2020) 'On the Ethics of Algorithmic Decision-Making in Healthcare', *Journal of Medical Ethics*, 46(3), p. 208.

Hall, D. *et al.* (2012) 'Informed Consent for Clinical Treatment', *Canadian Medical Association Journal*, 184(5), pp. 534, 536.

Han, P. *et al*. (2013) 'The Value of Personalised Risk Information: A Qualitative Study of the Perceptions of Patients with Prostate Cancer', *BMJ Open*, 3(9), pp. 1-2.

Iserson, K. (2024) 'Informed Consent for Artificial Intelligence in Emergency Medicine: A Practical Guide', *The American Journal of Emergency Medicine*, 76, pp. 225-228.

Jie, Z. *et al.* (2021) 'A Meta-Analysis of Watson for Oncology in Clinical Application', *Scientific Reports*, 11(5792), p. 1.

Kadam, R.A. (2017) 'Informed Consent Process: A Step Further Towards Making it Meaningful!', *Perspectives in Clinical Research*, 8(3), p. 107.

Kempt, H. *et al.* (2022) 'Relative Explainability and Double Standards in Medical Decision-Making', *Ethics and Information Technology*, 24(2), p. 20.

Kimiafar, K. *et al.* (2023) 'Artificial Intelligence Literacy Among Healthcare Professionals and Students: A Systematic Review', *Frontiers in Health Informatics*, 12(168), p. 2.

Krasniansky, A. (2019) *Understanding Racial Bias in Medical AI Training Data - Bill of Health*.

Krishnan, G. *et al.* (2023) 'Artificial Intelligence in Clinical Medicine: Catalysing a Sustainable Global Healthcare Paradigm', *Frontiers in Artificial Intelligence*, 6, p. 5.

Krive, J. *et al.* (2023) 'Grounded in Reality: Artificial Intelligence in Medical Education', *JAMIA Open*, 6(2), pp. 6, 7.

Licholai, G. (2023) *It's Time For 'Nutrition Labels' In Artificial Intelligence*, *Forbes*.

Liu, M. *et al.* (2023) 'A Translational Perspective Towards Clinical AI Fairness', *Digital Medicine*, 6, p. 3.

Mangalji, A. *et al.* (2019) 'Preparing for the Future of Medicine: Considering the Need for Data-Literate Physicians', *British Columbia Medical Journal*, 61(8), p. 326.

McDougall, R. (2019) 'Computer Knows Best? The Need for Value-Flexibility in Medical AI', *Journal of Medical Ethics*, 45(3), pp. 157-158.

Mittelstadt, B. (2021) *The Impact of Artificial Intelligence on the Doctor-Patient Relationship*. Council of Europe, pp. 46, 48, 49, 51.

Mittermaier, Mirja *et al.* (2023) 'Bias in AI-Based Models for Medical Applications: Challenges and Mitigation Strategies', *Digital Medicine*, 6(1), p. 1.

Neal, D. *et al.* (2022) 'Is There Evidence of Age Bias in Breast Cancer Healthcare Professionals' Treatment of Older Patients?', *European Journal of Surgical Oncology*, 48(12), p. 2401.

Ng, F.Y.C. *et al.* (2023) 'Artificial Intelligence Education: An Evidence-Based Medicine Approach for Consumers, Translators, and Developers', *Cell Reports Medicine*, 4(10), pp. 2, 8.

Norori, N. *et al.* (2021) 'Addressing Bias in Big Data and AI for Healthcare: A Call for Open Science', *Patterns*, 2(10), pp. 2, 3.

Noul, D. (2024) *Personalised Healthcare: Fusing Biomedical Engineering and AI for Customised Patient Care*.

O'Brien, N. *et al.* (2022) Ad*dressing Racial and Ethnic Inequities in Data-driven Health Technologies*. Institute of Global Health Innovation, Imperial College London, pp. 7, 10, 11, 18.

Population of England and Wales (2022).

Price, N. (2019) 'Medical AI and Contextual Bias', *Harvard Journal of Law & Technology*, 33(1), pp. 66, 67, 68, 100.

Ricci Lara, M.A. *et al.* (2022) 'Addressing Fairness in Artificial Intelligence for Medical Imaging', *Nature Communications*, 13(1), p. 1.

Salway, S. *et al.* (2014) 'Race Equality and Health Inequalities: Towards More Integrated Policy and Practice', *Race Equality Foundation*.

Sendak, M. *et al.* (2020) 'Presenting machine learning model information to clinical end users with model facts labels', *Digital Medicine*, 3(1), p. 2.

Sokhach, D. (2023) *The Role of AI in Personalised Healthcare, Institute of Entrepreneurship Development*.

Stypińska, J. *et al.* (2023) 'AI Revolution in Healthcare and Medicine and the (Re)emergence of Inequalities and Disadvantages for Ageing Population', *Frontiers in Sociology*, 7, p. 4.

Terranova, C. *et al.* (2024) 'AI and Professional Liability Assessment in Healthcare. A Revolution in Legal Medicine?', *Frontiers in Medicine*, 10, p. 3.

The NHS Constitution for England (2023).

Ueda, D. *et al.* (2024) 'Fairness of Artificial Intelligence in Healthcare: Review and Recommendations', *Japanese Journal of Radiology*, 42(1), p. 4.

Van Kolfschooten, H. (2023) 'The AI Cycle of Health Inequity and Digital Ageism: Mitigating Biases Through the EU Regulatory Framework on Medical Devices', *Journal of Law and the Biosciences*, 10(2), p. 5.

Wang, Y. & Ma, Z. (2024) 'Ethical and Legal Challenges of Medical AI on Informed Consent: China as an Example', *Developing World Bioethics*, pp. 3, 4, 5.

What is an Electronic Health Record (EHR)? | *HealthIT.Gov* (2019).

Wood, E. *et al.* (2021) 'Are We Ready to Integrate Artificial Intelligence Literacy into Medical School Curriculum: Students and Faculty Survey', *Journal of Medical Education and Curricular Development*, 8, p. 1