Patient-Centric Ethical Frameworks for Privacy, Transparency, and Bias Awareness in Deep Learning-Based Medical Systems

Shivansh Khanna
School of Information Sciences, University of Illinois at Urbana-Champaign
Shraddha Srivastava
School of Information Sciences, University of Illinois at Urbana-Champaign

Abstract
The rapid advancement and deployment of deep learning-enabled medical systems have necessitated the development of robust ethical frameworks to address potential challenges and pitfalls. Based on the foundational principles of medical ethics—non-maleficence, beneficence, respect for patient autonomy, and justice—three ethical frameworks are proposed in this study for the deployment and oversight of deep learning systems in healthcare. This study presents these three distinct yet interconnected ethical frameworks focusing on patient privacy, transparency, and bias mitigation. The patient privacy framework argues for the importance of patient autonomy. It advocates for informed consent, emphasizing the need for patients to be apprised of the system's workings, benefits, potential risks, and alternatives. Consent should be voluntary, devoid of implicit coercion, and patients must retain the right to revoke it without repercussions. The framework also included the principles of transparency, beneficence, privacy, continual consent, accessibility, and accountability. It champions the idea that consent is dynamic, necessitating regular updates, especially when significant system changes occur. Our ethical framework for transparency accentuates the need for full disclosure. Stakeholders should be provided with a general overview of the system's operations, its inputs, and decision-making processes. Performance metrics, including accuracy, sensitivity, and specificity, should be transparently communicated. Openness, through open-source initiatives and third-party audits, is promoted. The principles of accountability, data transparency, continuous improvement, inclusivity, and external validation are also made integral to this framework, ensuring that stakeholders are consistently informed and engaged. The bias minimization framework highlights the imperative of awareness. Stakeholders should be educated about potential biases and their ramifications. The system should be regularly evaluated for inherent biases, both overt and subtle. Representation is crucial; training data must reflect diverse populations, considering various demographic factors. This framework also promotes fairness, ensuring equitable system performance across different patient groups. Transparency in bias reporting, accountability in bias correction, continuous monitoring, inclusivity in stakeholder engagement, and collaboration with interdisciplinary teams are also included and discussed.

***Keywords**: Medical Ethics, Deep Learning, Privacy, Transparency, Bias, Deep Learning Ethics*

Introduction
The current abundant biomedical data is reshaping the healthcare sector, transforming it into a domain that increasingly relies on data-driven decision-making. The collection and storage of varied types of biomedical data such as genomics, patient medical history, electronic health records (EHR), and imaging studies have become more streamlined. The volume and complexity of this data necessitate sophisticated methods of analysis. Machine learning algorithms and artificial intelligence are being employed to examine large datasets to identify patterns, predict patient outcomes, and even suggest treatment pathways. Additionally, the

interoperability of systems and devices in healthcare allows for a more seamless integration of data from different sources, fostering a more holistic approach to patient care.

Within this new data-rich environment, several key factors are influencing healthcare delivery and research. First, personalized medicine is becoming increasingly feasible. The capability to analyze large sets of individual genetic information alongside environmental factors permits more targeted treatments, reducing the dependency on one-size-fits-all solutions. This personalized approach could lead to higher success rates for treatments and possibly decrease healthcare costs by minimizing ineffective interventions. Second, public health initiatives can benefit from the aggregation and analysis of population-level data. This enables healthcare practitioners and policy-makers to identify trends, risk factors, and efficacy of interventions at a macro level, helping to formulate policies or health programs that are evidence-based.

The transition from classical machine learning techniques to Deep Learning (DL) models represents a significant shift, largely due to DL's capability to process vast quantities of data swiftly and extract hidden, useful knowledge. Unlike traditional machine learning methods, which often rely on manually crafted features and simpler algorithms, DL models, particularly neural networks, are capable of automatically learning to identify features from raw data. This ability to learn from data is exponentially beneficial as the volume of available data increases. The architecture of DL models, composed of multiple layers of interconnected nodes, allows for more complex and nuanced understanding of data patterns. This has resulted in accelerated data processing and improved predictive performance, making DL models not only faster but also more accurate. Traditional methods of image analysis often required time-consuming procedures and expert human intervention. However, DL algorithms have been developed to outperform human abilities in tasks such as detecting abnormalities in X-rays, MRI scans, and CT scans. These algorithms can analyze multiple dimensions of data simultaneously and are capable of identifying intricate patterns that may not be easily discernible to the human eye. As a result, there is enhanced diagnostic accuracy, quicker turnaround times for results, and a subsequent increase in the efficiency of healthcare delivery.

This technological shift has introduced challenges related to privacy, transparency, and bias. Privacy concerns arise when medical systems collect, store, and process vast amounts of sensitive personal health information. There is a risk that such data can be compromised through security breaches, unauthorized access, or misuse. Additionally, many deep learning algorithms require access to large datasets to function optimally, which sometimes involves pooling data from multiple institutions or countries. This centralization can exacerbate privacy concerns, as it could potentially lead to the violation of various jurisdictions' data protection regulations.

Many of these algorithms, including neural networks, operate as "black boxes," making it difficult to interpret their decision-making processes. This lack of transparency poses a significant challenge in a medical context, where understanding the rationale behind diagnostic or treatment recommendations is crucial for both healthcare providers and patients. Regulatory agencies such as the Food and Drug Administration in the United States are increasingly demanding more explainable artificial intelligence solutions, particularly in applications with significant ethical and safety implications like healthcare.

Deep learning models are trained on existing datasets, which may include both implicit and explicit biases. For instance, if a dataset disproportionately represents a particular demographic group, the resultant model may perform poorly when applied to individuals from underrepresented groups. This can lead to inaccurate diagnoses and treatment recommendations, exacerbating existing healthcare disparities. Beyond demographic biases, datasets can also contain biases related to the conditions under which data was collected, the medical institutions that contributed the data, or even the specific medical devices used for data collection.

Studies in medical ethics such as [1]–[5] have reported four foundational principles that guide ethical decision-making in healthcare. These principles include non-maleficence, beneficence, respect for patient autonomy, and justice, as shown in table 1.

Table 1. Common ethical principals in healthcare

Principle	Description	Practical Application
Non-maleficence	Obligation for healthcare professionals to do no harm to patients.	In surgical procedures, the surgeon minimizes pain and adverse effects to ensure that the harm does not outweigh the intended benefit of the surgery.
Beneficence	Healthcare providers are compelled to actively promote the well-being of patients.	Recommending preventive measures, providing effective treatment, and considering the long-term health and well-being of the patient.
Respect for Patient Autonomy	Medical practitioners honor the individual patient's right to make informed decisions about their healthcare.	Obtaining informed consent before procedures and respecting patients' choices regarding their treatment, even if those choices contradict medical advice.
Justice	Focuses on the equitable distribution of healthcare resources and services.	Treating all patients equally and allocating medical resources without discrimination based on race, gender, socioeconomic status, or other non-medical factors.

The first principle, non-maleficence, dictates that healthcare professionals have an obligation to do no harm to patients. This principle serves as a foundational tenet for medical ethics and directs medical practitioners to avoid causing unnecessary suffering, pain, or harm to patients during treatment. The concept is deeply ingrained in the Hippocratic Oath and forms the cornerstone of the relationship between medical professionals and those under their care. For example, in surgical procedures, while it is understood that some level of pain or discomfort may be inevitable, the surgeon is ethically bound to minimize such adverse effects and not cause harm that outweighs the intended benefit of the surgery [6].

Beneficence is the second principle and compels healthcare providers to actively promote the well-being of patients. Unlike non-maleficence, which is a more passive principle focused on avoiding harm, beneficence mandates that medical professionals take proactive steps to improve the health outcomes of patients. This may include recommending preventive measures, providing effective treatment for diseases, and considering the long-term health and well-being of the patient in medical decisions. In other words, it calls for an altruistic approach where the primary concern is the patient's welfare, rather than merely avoiding legal repercussions or adhering to minimum standards of care.

The third and fourth principles, respect for patient autonomy and justice, deal with patient rights and social considerations. Respect for patient autonomy requires medical practitioners to honor the individual patient's right to make informed decisions about their own healthcare. Practically, this translates to obtaining informed consent before conducting medical procedures and respecting patients' choices regarding their treatment, even if it contradicts the medical advice given. Justice, on the other hand, focuses on the equitable distribution of healthcare resources and services. It asks healthcare providers to treat all patients equally and to allocate medical resources in a manner that does not discriminate on the basis of race, gender, socioeconomic status, or any other factor unrelated to medical need. This principle is especially important in public health settings where resources are often limited and ethical dilemmas may arise in their allocation.

Trustworthy AI systems require a foundation built on several key principles. The first is transparency, which means that the operations of the AI system are visible to the user. Users should be able to understand how decisions are being made, which fosters trust. The second principle is credibility. The outcomes of the AI system should be acceptable to those who use it or are affected by it. This principle ensures that the AI system makes decisions that users will see as fair and reasonable. Additionally, auditability is crucial, as the efficiency and effectiveness of the system should be measurable. This allows for the assessment and potential improvement of the AI system's functions.

The last two principles focus on the practical aspects of AI systems. Reliability is paramount; the AI systems must perform as intended consistently. Any deviation could lead to unintended consequences that might erode trust or even cause harm. Finally, recoverability is another key principle. If the AI system fails or makes an erroneous decision, manual control should be easily assumed. This serves as a fail-safe, allowing human intervention when necessary to correct or override the system's actions.

Table 2. Ethical principles in healthcare AI systems

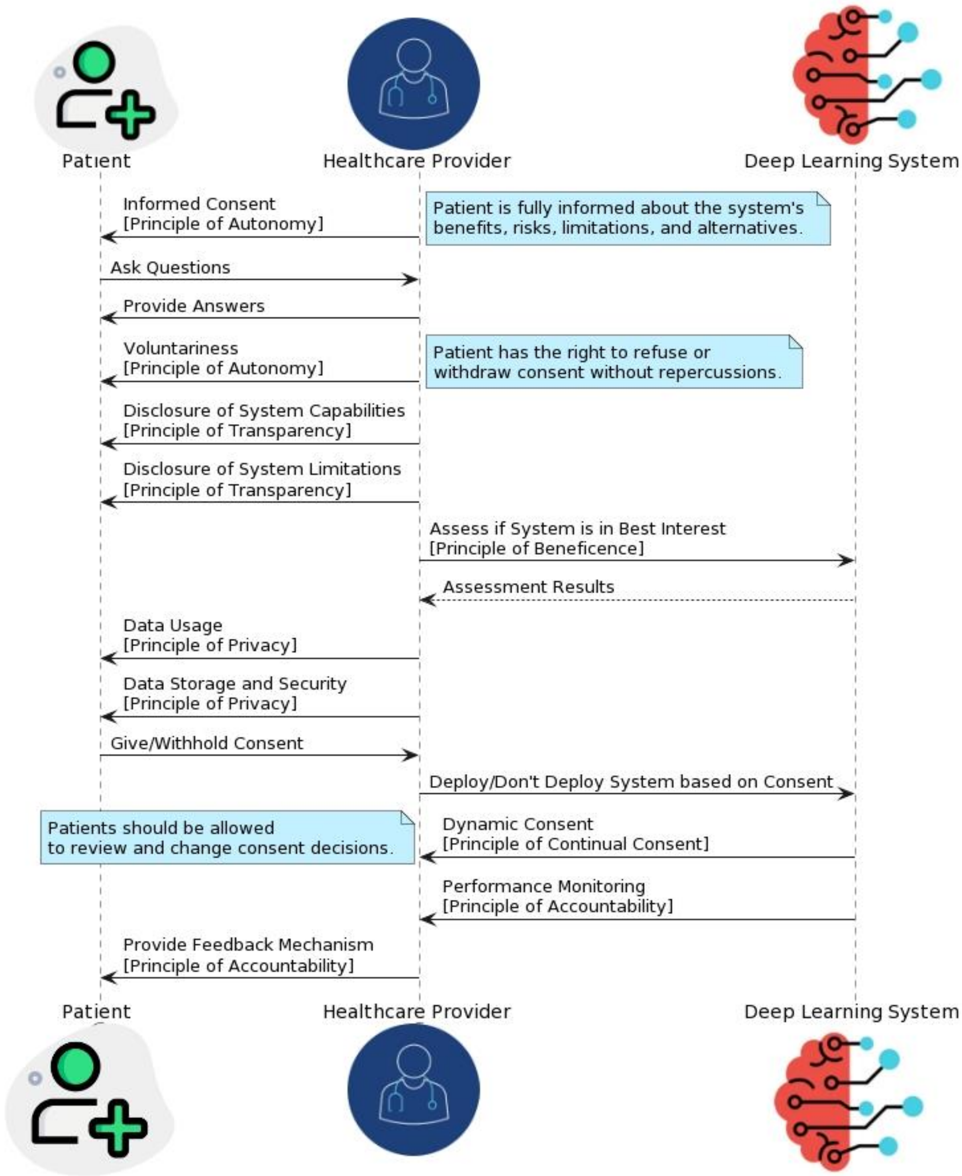
Principle	Description
Transparency	Operations of the AI system are visible and understandable to the user.
Credibility	The outcomes or decisions made by the AI system are acceptable and accurate.
Auditability	The performance and efficiency of the AI system can be easily measured and assessed.
Reliability	The AI system consistently performs as intended without errors.
Recoverability	If the AI system fails or makes an error, manual control can be assumed to correct the situation.

Source: [7] **and** [5]

Ethical Framework for Privacy

Informed consent and voluntariness are foundational concepts in the field of medical ethics, but their implications extend to legal studies, psychology, and philosophy as well. Informed consent is the process by which a healthcare provider communicates relevant information to a patient, allowing them to make an educated decision about their treatment [8]. The crucial elements include disclosure of information, understanding, and the absence of coercion or deception. The patient must be given sufficient details about the procedure, the risks involved, and alternative courses of action.

Figure 3. Flow of information and feedback in the ethical framework for privacy



Voluntariness, on the other hand, refers to the capacity to make a choice free from coercion, manipulation, or undue influence. It is closely related to the concept of autonomy, a principle highly esteemed in Western philosophies. Voluntariness ensures that the individual's actions are a result of their own desires, intentions, and rational deliberation, without external forces unduly influencing the decision-making process.

The principle of autonomy is foundational in medical ethics and underpins the concept of informed consent. Autonomy implies the right of individuals to make decisions about their own bodies and lives without undue external influence. This principle gains heightened importance in the context of healthcare, where patients often find themselves in vulnerable situations. When deep learning systems are introduced into healthcare settings, the respect for autonomy must remain paramount. Patients should be fully informed about how these systems function, what benefits they offer, the limitations inherent to their design, potential risks involved, and any available alternatives. This information equips patients with the necessary tools to make informed decisions. The provision of an opportunity for the patient to ask questions and receive clear, understandable answers is integral to upholding this autonomy. Without comprehensive information, patients cannot exercise their autonomy effectively, which undermines the ethical foundations of healthcare.

Informed consent is an extension of the principle of autonomy, serving as a procedural safeguard to ensure that patients are actively involved in decisions about their healthcare. The process of informed consent is not merely a formality or a legal requirement but a moral obligation that healthcare providers owe to their patients. It becomes even more significant in the context of deploying deep learning systems in healthcare because these systems may involve complex algorithms that are not easily understandable by the general population. The aim should be to distill this complex information into understandable terms without diluting its essential elements, thus facilitating a genuinely informed decision-making process. Ensuring that patients understand not just the benefits but also the limitations and potential risks of a deep learning system promotes ethical integrity by aligning the deployment of technological advancements with the principle of patient autonomy [9].

The concept of voluntariness complements and strengthens the principles of autonomy and informed consent. Voluntariness ensures that the consent given is free from any form of coercion, manipulation, or undue influence. In a healthcare setting, this means that patients should feel empowered to either give or withdraw consent at any stage without fear of repercussions. This right is especially crucial when deep learning systems are involved, as patients might have varying levels of comfort and trust with technology. Some may harbor reservations based on ethical, cultural, or personal grounds. The decision to participate should stem from the individual's own reasoned choices and not from external pressures.

The disclosure of system capabilities is a critical component in the deployment of deep learning systems in healthcare settings. It aligns closely with the ethical imperative of informed consent, which is rooted in the principle of autonomy. Patients have the right to understand the tools and methods that will be employed in their healthcare, including the intricacies of deep learning algorithms. For example, understanding the system's accuracy rates, its potential for misdiagnosis, and any known biases it may have are vital pieces of information. Providing these details does not merely fulfill a legal requirement but serves an ethical role. This transparency allows patients to gauge the reliability of the system and make a balanced decision regarding whether they wish to proceed with healthcare interventions that involve deep learning technologies. Clear, straightforward communication about system capabilities thus serves to enhance the individual's autonomy, empowering them to make informed choices regarding their healthcare.

The ethical obligation to disclose does not end with the capabilities of the system but extends to its limitations as well. Informing patients about the potential for errors, the necessity for human oversight, and areas where the system might not be as effective is crucial for maintaining ethical integrity. Deep learning systems, while powerful, are not infallible and can yield incorrect or misleading results. Furthermore, they often operate in tandem with human professionals, meaning that their output is part of a larger decision-making framework that may include additional tests or professional judgment [10]. Patients should be made aware of these limitations so that they can factor them into their decisions. By doing so, healthcare providers honor the patients' right to be fully informed, facilitating more nuanced and individualized decision-making processes.

The principle of privacy is a cornerstone in healthcare ethics and becomes increasingly salient with the introduction of deep learning and other advanced technologies that rely on data analytics. The usage of patient data by these systems must be communicated clearly to patients to ensure that their privacy rights are respected and maintained. It is essential to disclose how the data will be utilized, whether for diagnostic, treatment, or research purposes. An integral part of this disclosure involves ensuring that the data is anonymized whenever possible and cannot be traced back to individual patients unless it is necessary for their treatment. This process serves to protect patient identity and fosters trust between patients and healthcare providers. In adhering to these guidelines, the healthcare industry aligns itself with the principle of privacy, thereby reinforcing the importance of patient autonomy in data-related decisions.

Similarly, the storage and security of patient data are crucial factors that warrant serious ethical consideration. Patients should be assured that robust security measures are in place to prevent unauthorized access, data breaches, or misuse of their information. These measures can range from encrypted storage solutions to strict access protocols that limit the number of personnel who can access the data. Clear and transparent communication about these security measures is essential to engender patient trust. The secure storage of patient data not only adheres to regulatory guidelines but also embodies the ethical commitment to respect patient privacy. By diligently safeguarding data, healthcare providers and institutions validate the patient's autonomy over their personal information, fulfilling an ethical obligation that extends beyond mere compliance with laws and regulations.

The principle of continual consent elevates the ethical standards for deploying deep learning systems in healthcare by acknowledging that consent is not a static, one-time event but a dynamic process. This perspective integrates the concept of dynamic consent, which allows patients to revisit and revise their initial decisions as new information emerges or as their personal circumstances change. For instance, if initial consent was provided for the use of a deep learning system for diagnosis, but the patient later learns more about potential biases or limitations in the system, they should have the freedom to modify or withdraw their consent. Dynamic consent provides the flexibility to adapt to changing conditions or understandings, thus preserving the patient's autonomy over an extended period. It allows the patient to be an active participant in their healthcare journey, continuously evaluating the risks and benefits as they evolve [11].

Furthermore, the principle of continual consent also addresses scenarios where the deep learning system itself undergoes significant modifications. Technology is ever-evolving, and updates to the system could impact its performance, reliability, or the way it uses and stores patient data. If such changes occur, it is ethically imperative that patients be re-informed about these updates and be given an opportunity to renew or withdraw their consent. This provision not only adheres to the principles of transparency and autonomy but also recognizes the ethical complexities that arise from the evolving nature of technology. Not informing patients of significant system changes would violate their right to make informed decisions, thereby undermining the core ethical principle of autonomy.

The principle of accessibility is pivotal in ethical healthcare frameworks, particularly when introducing technology such as deep learning systems. One of its key tenets is equal access, ensuring that all patients, irrespective of their socioeconomic background, ethnicity, or other distinguishing factors, have the same opportunity to receive information about the system. Equally important is the opportunity to give or withhold consent based on that information. The absence of equal access not only contravenes ethical principles but also poses a risk of exacerbating healthcare disparities. For example, marginalized communities may find themselves further disadvantaged if they are not afforded the same opportunities to understand and consent to the deployment of advanced technologies in their healthcare. By providing equal access, healthcare providers and institutions uphold the principle of justice, which aligns closely with individual autonomy, ensuring that all patients can make informed decisions.

Tailored communication serves as a complementary aspect of the principle of accessibility. While equal access ensures that all patients have the opportunity to receive information, tailored communication guarantees that this information is presented in a way that is comprehensible to each individual. This involves taking into account factors like literacy levels, language proficiency, and cultural considerations. For instance, medical jargon should be avoided or thoroughly explained, and translators should be made available for those who are not proficient in the primary language used by healthcare providers. Additionally, cultural sensitivities should be acknowledged to ensure that the information is not just technically accurate but also socially and culturally appropriate.

The principle of accountability addresses the ethical obligations of healthcare providers and institutions to maintain rigorous oversight and monitoring of deep learning systems. Oversight mechanisms serve multiple purposes: they ensure that the system is functioning as intended, identify any potential errors or biases, and help in evaluating whether the technology is causing unintended harm. Monitoring should be ongoing and dynamic, adapting to new data, updated algorithms, and other changes that could impact the system's performance. If issues arise, there is an ethical imperative to promptly correct them to prevent harm to patients. Furthermore, transparent communication about the oversight process can enhance patient trust and participation, thereby supporting the principle of autonomy. By actively monitoring and addressing issues, healthcare providers fulfill their ethical duty to protect the well-being of their patients while also bolstering the integrity of the technology they employ.

A feedback loop from patients serves as a complementary accountability measure, reinforcing the ethical principles of autonomy and informed consent. Patients are among the most critical stakeholders in any healthcare system, and their experiences and insights can offer valuable perspectives that may not be readily apparent to healthcare providers or system developers. Establishing a structured system for collecting patient feedback not only helps in identifying unforeseen problems but also creates an environment where patients feel heard and respected. This feedback should be taken seriously, analyzed systematically, and used to make necessary adjustments to both the deep learning system and the informed consent process. By incorporating patient feedback, healthcare providers and institutions demonstrate a commitment to continuous improvement and ethical responsibility.

Table 3. Ethical framework for privacy in deep learning-based AI systems

Principle	Sub-Principle	Description
Principle of Autonomy	Informed Consent	Before deploying any deep learning system, patients should be fully informed about how the system works, its benefits, limitations, potential risks, and alternatives. They should have the opportunity to ask questions and receive clear answers.
	Voluntariness	Consent should be given voluntarily, without any form of coercion or undue influence. Patients should have the right to refuse or withdraw their consent at any point without facing any repercussions.
Principle of Transparency	Disclosure of System Capabilities	Clearly communicate the capabilities of the deep learning system, including its accuracy, potential for misdiagnosis, and any known biases.
	Disclosure of System Limitations	Patients should be made aware of the limitations of the system, such as potential errors, the need for human oversight, and any areas where the system might not be as effective.
Principle of Privacy	Data Usage	Clearly communicate how patient data will be used by the system, ensuring that data is anonymized and cannot be traced back to individual patients unless necessary for treatment.
	Data Storage and Security	Ensure that patient data is stored securely, with robust measures to prevent unauthorized access, breaches, or misuse.
Principle of Continual Consent	Dynamic Consent	Recognize that consent is not a one-time event. Patients should be given the opportunity to review and change their consent decisions as more information becomes available or as their personal circumstances change.
	Updates and System Changes	If the deep learning system undergoes significant updates or changes that might affect its performance or the way it uses patient data, patients should be re-informed and given the opportunity to renew or withdraw their consent.
Principle of Accessibility	Equal Access	Ensure that all patients, regardless of their background, have equal access to information about the system and the opportunity to give or withhold consent.
	Tailored Communication	Information about the system should be presented in a manner that's accessible and understandable to all patients, taking into account factors like literacy levels, language proficiency, and cultural considerations.
Principle of Accountability	Oversight and Monitoring	There should be mechanisms in place to monitor the performance of the deep learning system, ensuring that it's working as intended and not causing harm. Any issues or errors should be promptly addressed.
	Feedback Loop	Establish a system where patients can provide feedback or raise concerns about the deep learning system. This feedback should be taken seriously and used to improve the system and the consent process.

Ethical Framework for Transparency

The principle of full disclosure mandates a transparent approach to the deployment of deep learning systems in healthcare, aiming to provide stakeholders with comprehensive information. One key element of this principle is the algorithmic explanation. While the inner workings of deep learning models are often complex and may not be fully understood even by experts, providing a general overview of the system's functioning is crucial. Stakeholders should be informed about how data is collected, processed, and utilized by the model for decision-making. This aspect of disclosure fosters an environment where stakeholders, particularly patients, can have a better understanding of what to expect from the technology, thereby allowing them to make more informed decisions [12]. Transparency about the algorithm bolsters trust and aligns with the ethical principle of autonomy, as it enables stakeholders to have a foundational understanding of the technology that impacts their healthcare.

Another critical aspect of full disclosure is the communication of performance metrics. Information about the system's accuracy, sensitivity, specificity, and other relevant performance indicators should be clearly communicated to all stakeholders. This includes not just the successes but also the limitations and failures of the system. For instance, if the system has a particular rate of false positives or false negatives, that information should be disclosed. Such transparency serves multiple purposes: it aids stakeholders in evaluating the system's reliability, informs healthcare providers about when to rely on the system and when to be cautious, and ultimately helps patients make more informed decisions. This commitment to full disclosure goes beyond mere statistical transparency; it is an ethical obligation to provide stakeholders with a balanced and accurate understanding of the system's capabilities and limitations.

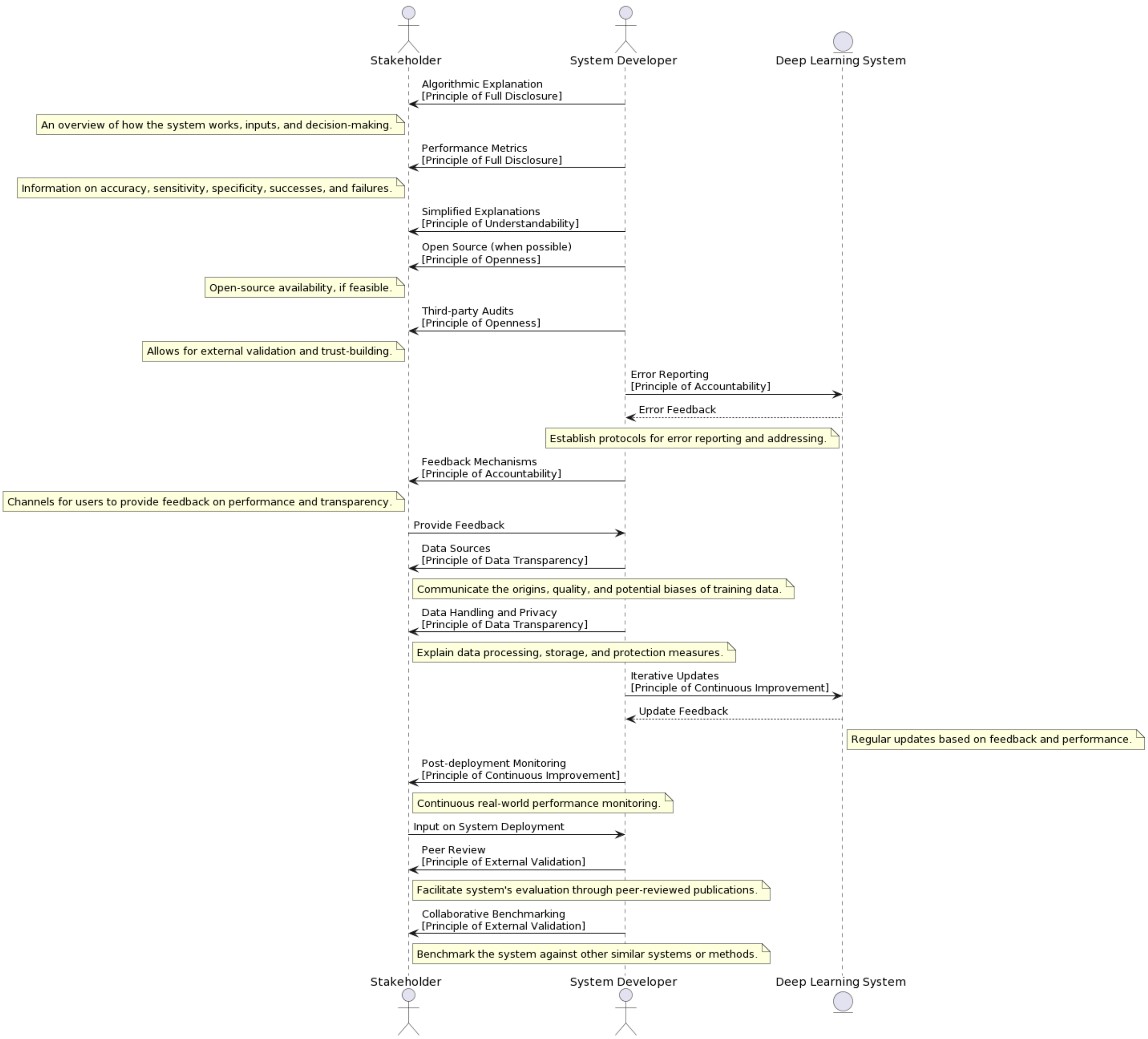
The principle of understandability recognizes the complexity inherent in deep learning systems and seeks to bridge the gap between technological sophistication and public comprehension. One approach to achieving this is through simplified explanations. Offering explanations in layman's terms ensures that the system's workings can be understood by a broad audience, including patients and medical professionals who may not have a technical background. When people understand the underlying mechanics of a system that impacts their health, they are better equipped to exercise their autonomy effectively, thus upholding the principle of informed consent [13].

Complementing simplified explanations, visual aids such as diagrams or infographics can be utilized to further elucidate the system's processes and decisions. The integration of visual tools can often convey complex ideas more succinctly and intuitively than text alone. The use of these aids contributes to a more robust understanding and facilitates informed decision-making. Like simplified explanations, the use of visual aids serves the broader ethical goal of making complex technologies accessible to the public, thereby enhancing individual autonomy and informed consent [14].

The principle of openness advocates for transparency through avenues like open-source software. When feasible, and without compromising proprietary or sensitive information, making the algorithm or its components open source can allow for external validation and scrutiny. This act of openness contributes to building trust among stakeholders and provides an opportunity for collaborative improvement of the system. By opening up the algorithm for public scrutiny, healthcare providers and developers invite a broader evaluation of the system's merits and shortcomings, thereby enhancing accountability.

Error reporting requires establishing clear protocols for recording, investigating, and rectifying errors or misdiagnoses made by the deep learning system. Such protocols should outline the steps for internal review, immediate corrective action, and future prevention. The objective is to ensure that any faults in the system are not only corrected promptly but also analyzed thoroughly to prevent recurrence. Instituting robust error reporting mechanisms signals a commitment to maintaining high standards of care and enhances the trust that stakeholders, particularly patients, have in the healthcare system.

Figure 2. Flow of information and feedback in the ethical framework for privacy



Feedback mechanisms form another crucial component of accountability. Channels should be established to allow users—both medical professionals and patients—to provide feedback regarding the system's performance and its efforts in maintaining transparency. Medical professionals could offer insights into how well the system integrates into existing workflows and its efficacy in diagnosis or treatment, while patients could comment on their personal experiences and whether they felt sufficiently informed to give consent. This feedback should not be relegated to a cursory review; rather, it should be methodically analyzed and used to inform future iterations of the system and the overarching informed consent process.

The principle of accountability in the context of healthcare technology is designed to ensure that errors, misdiagnoses, and other issues are promptly identified, reported, and rectified. One fundamental aspect is the establishment of clear protocols for error reporting. These protocols should include guidelines on how errors are to be documented, the processes for internal and, if necessary, external review, and the steps for corrective action. Such measures are crucial for the timely resolution of problems and to mitigate the risk of harm to patients. The presence of clear and transparent error-reporting protocols enhances trust among all stakeholders and demonstrates an organization's commitment to maintaining a high standard of care. This approach aligns with the ethical imperatives of non-maleficence and beneficence, ensuring that the technology does no harm and operates in the best interests of the patients.

Feedback mechanisms are another essential facet of accountability. These are channels designed to enable both medical professionals and patients to offer insights on the performance of the deep learning system, as well as on the transparency and adequacy of the informed consent process. Collecting feedback from a diverse set of users provides a more comprehensive view of the system's performance and impact. It also serves as a tool for continual improvement, offering real-world data that can be analyzed to enhance both the technology and the policies surrounding its use. These feedback channels should not merely exist but should be actively monitored, and the feedback collected should be acted upon to improve the system and its governance mechanisms.

The principle of data transparency necessitates that stakeholders have a thorough understanding of how data is sourced, processed, and safeguarded, especially when deep learning systems are involved in healthcare settings. One vital aspect of this is the clear communication of data sources. The origin, quality, and potential biases in the training data should be made transparent to all stakeholders. Providing this information is essential for several reasons, including validating the reliability and integrity of the data and consequently, the system's results. For example, if a system was trained primarily on data from a specific demographic group, stakeholders need to be aware of this limitation. Clear communication about data sources aligns with the ethical principle of informed consent, as stakeholders can only make truly informed decisions when they are fully aware of the context in which the data was gathered and how it might influence the system's behavior. Data handling and privacy are equally critical in ensuring data transparency. This includes details about how the data is processed, stored, and protected. Measures for data anonymization and security protocols should be explicitly communicated to stakeholders. The processing of sensitive patient data carries with it a high ethical responsibility to safeguard individuals against breaches of privacy or data misuse [15]. Information about encryption methods, access controls, and data storage practices should be disclosed. This kind of transparency is not merely a technological requirement but an ethical obligation, aimed at ensuring that stakeholders can trust the system with their sensitive information.

The principle of continuous improvement emphasizes the importance of iterative updates and post-deployment monitoring to ensure that deep learning systems in healthcare remain effective, safe, and aligned with ethical standards. Iterative updates are a natural aspect of any evolving technology, and in the context of healthcare, such updates can have significant implications for patient outcomes. Therefore, it is crucial to keep stakeholders informed about any significant changes to the system, especially those that could affect its performance or decision-making processes. This aligns with the principle of informed consent, as well-informed stakeholders are better equipped to understand and navigate the complexities of an evolving healthcare technology. Transparency about updates serves to maintain trust and allows stakeholders to make informed decisions about continued engagement with the technology.

Post-deployment monitoring is another pillar of the principle of continuous improvement. The performance of the system in real-world settings should be continuously tracked, and this information should be made readily available to stakeholders. This practice not only serves as a mechanism for identifying areas for improvement but also as a way to maintain public trust in the technology. Monitoring could include tracking error rates, patient outcomes, and any unintended consequences of the system's deployment. By making this information accessible, stakeholders can gauge the system's ongoing efficacy and reliability, ensuring that it meets the ethical standards expected in a healthcare setting.

The principle of external validation provides further checks and balances in the form of peer review and collaborative benchmarking. Peer review, either through publication in academic journals or presentations at professional conferences, lends an additional layer of credibility and scrutiny to the system. It allows for constructive criticism and validation from experts in the field, contributing to system improvements and reinforcing stakeholder confidence. Collaborative benchmarking, on the other hand, involves comparing the system's performance against other similar systems or traditional diagnostic methods. Such comparisons offer valuable insights into how the system measures up to existing standards and can illuminate areas for improvement.

Table 4. Ethical framework for transparency in deep learning-based AI systems

Principle	Sub-Principle	Description
Principle of Full Disclosure	Algorithmic Explanation	While the intricate details of deep learning models can be complex, a general overview of how the system works, its inputs, and its decision-making process should be provided to stakeholders.
	Performance Metrics	Clearly communicate the system's accuracy, sensitivity, specificity, and other relevant performance metrics. This includes both its successes and its failures.
Principle of Understandability	Simplified Explanations	Offer explanations in layman's terms to ensure that non-experts, including patients and medical professionals without a tech background, can understand the system's workings.
	Visual Aids	Use visual aids, diagrams, or other tools to help explain the system's processes and decisions.
Principle of Openness	Open Source (when possible)	If feasible and without compromising proprietary or sensitive information, consider making the algorithm or parts of it open source. This allows for external validation and trust-building.
	Third-party Audits	Allow for third-party experts to review and audit the system to ensure its efficacy and safety.
Principle of Accountability	Error Reporting	Establish clear protocols for reporting and addressing errors or misdiagnoses made by the system.
	Feedback Mechanisms	Creating channels for users, including medical professionals and patients, to provide feedback on the system's performance and transparency efforts.

Principle of Data Transparency	Data Sources	Clearly communicate where the training data comes from, ensuring that stakeholders are aware of the data's origins, quality, and potential biases.
	Data Handling and Privacy	Explaining how patient data is processed, stored, and protected. This includes measures taken to anonymize data and ensure its security.
Principle of Continuous Improvement	Iterative Updates	As the system evolves and improves, stakeholders should be informed of significant updates, especially those that might affect the system's performance or decision-making process.
	Post-deployment Monitoring	Continuously monitor the system's performance in real-world settings and make this information available to stakeholders.
Principle of External Validation	Peer Review	Encourage and facilitate the system's evaluation through peer-reviewed publications or presentations at professional conferences.
	Collaborative Benchmarking	Participate in collaborative efforts to benchmark the system against other similar systems or traditional diagnostic methods.

Ethical Framework for Bias Mitigation

The Principle of Awareness mandates that all stakeholders involved in the deployment of deep learning systems in healthcare settings are well-informed about the possible existence of biases. These stakeholders range from developers and data scientists who build the algorithms to healthcare professionals and end-users who interact with or rely on these systems for medical decisions. Education on bias is critical, as even a slight bias can have significant ethical and clinical implications. A biased system may produce outcomes that unfairly favor one group over another, whether it be based on ethnicity, gender, or other sociocultural factors. This could potentially lead to misdiagnoses, inappropriate treatment plans, or unequal access to healthcare resources. Hence, awareness and education programs need to be implemented, focusing on the different types of biases that can manifest, such as selection bias, confirmation bias, and algorithmic bias, among others [16].

Bias Detection is the next critical step in ensuring the fairness and equity of deep learning systems in healthcare. It is not sufficient to merely educate stakeholders about the potential for bias; proactive measures must be taken to identify and rectify biases within the system. This involves a combination of frequent testing, evaluation, and auditing of the algorithms. Techniques such as sensitivity analysis, fairness-aware modeling, and disparate impact analysis can be applied to detect both glaring and subtle biases that might compromise the system's performance across various patient demographics. The detection should also encompass the analysis of the data sets used to train the models, as biased data will inevitably lead to biased outcomes. Furthermore, this process must be ongoing and updated as new data becomes available or as the system is scaled to different settings to maintain its accuracy and fairness.

The Principle of Representation emphasizes the importance of using diverse training data when creating deep learning systems for healthcare applications. The foundation for a machine learning model's performance lies in the quality and diversity of its training data. If the dataset used to train these models is not adequately representative of the population it serves, the system will inevitably produce skewed or incorrect results, possibly leading to medical errors and injustices. Factors such as age, gender, ethnicity, socioeconomic status, and other relevant variables should be carefully considered during data collection and subsequent model training.

A well-curated, diverse dataset not only improves the generalizability of the model but also ensures that it caters to the unique healthcare needs of different demographic groups.

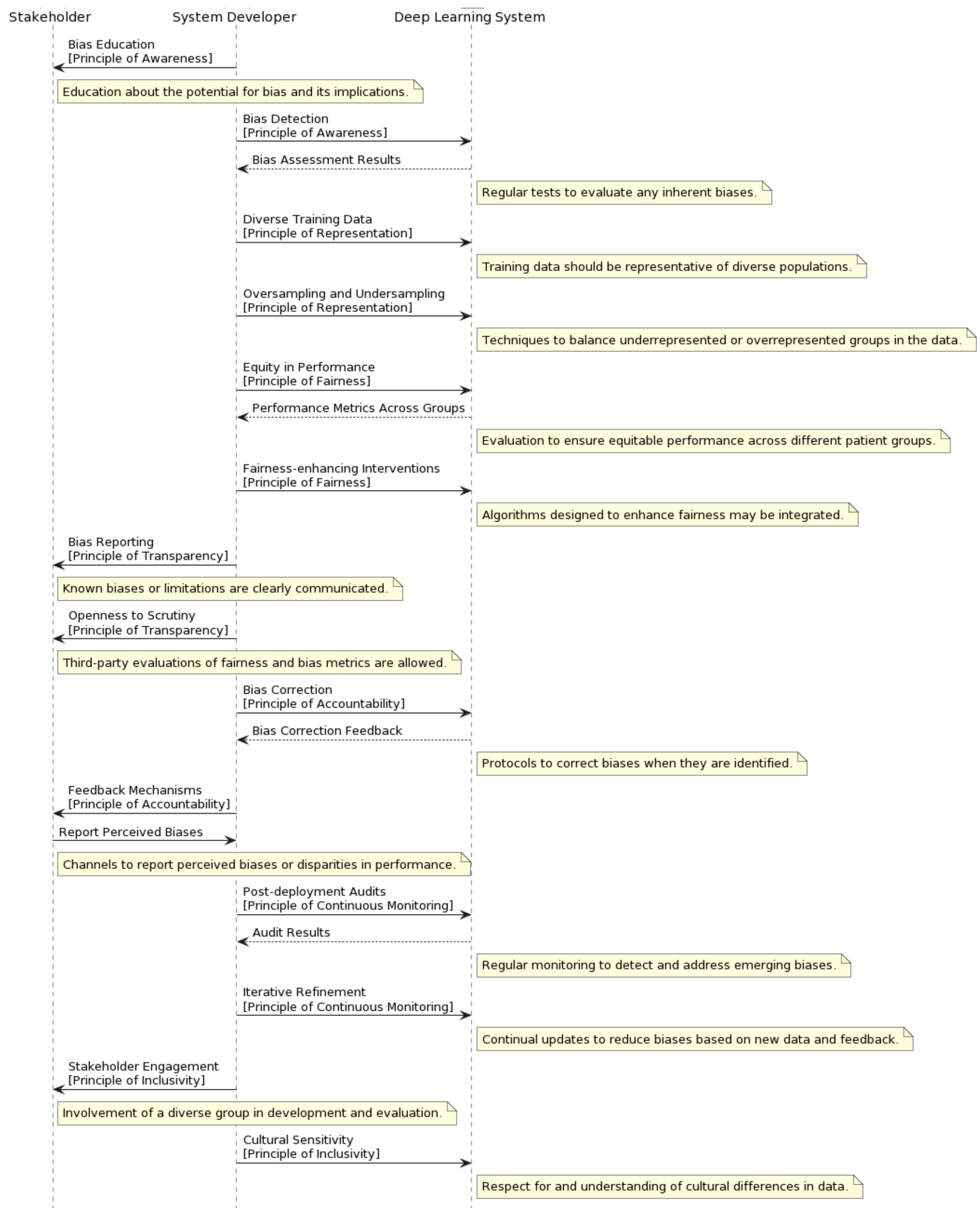
Oversampling and undersampling are techniques that can be employed to address data imbalances in training datasets. If certain demographic groups are underrepresented, their instances can be oversampled, meaning that these instances are replicated or given higher weights during the model training process. On the other hand, if some groups are overrepresented, their instances can be undersampled, reducing their influence on the model. These techniques aim to create a more balanced dataset, and subsequently, a more equitable and reliable machine learning model. However, care must be taken to not distort the data's natural distribution, as that may result in a model that is artificially biased or performs poorly when faced with real-world, unbalanced data [17].

The Principle of Fairness stipulates that deep learning systems in healthcare should be evaluated for equitable performance across diverse patient groups. Equity in performance means that the machine learning algorithms should not favor one group over another based on variables such as ethnicity, gender, age, or socioeconomic status. Any disparities in performance can have far-reaching consequences, potentially leading to misdiagnoses or ineffective treatment plans for certain demographic groups. To assess the system's equity, rigorous evaluation metrics and methodologies must be employed. This often involves cross-validation techniques, stratified sampling based on various demographic attributes, and detailed performance analyses for each subgroup. If any disparities in performance metrics are identified, they should be comprehensively addressed and rectified before the system is deployed in a healthcare setting.

In addition to routine performance evaluations, Fairness-enhancing Interventions can be incorporated into the system to proactively ensure fairness. These are algorithms and techniques specifically designed to minimize or eliminate bias in machine learning models. Methods such as reweighting the training data, employing adversarial training, and using fairness-aware regularization are some of the techniques that can be implemented. These interventions are designed to modify either the training data or the learning algorithm itself to produce a model that meets pre-defined fairness criteria. The choice of fairness-enhancing methods should be made in consultation with domain experts, ethicists, and data scientists to ensure they align with the specific use-case and ethical requirements of the healthcare application in question.

The Principle of Transparency advocates for clear communication about any known biases or limitations inherent in the healthcare-based deep learning system. Transparency is pivotal for the end-users, which may include healthcare providers, administrators, and patients, to make well-informed decisions. This requires not only clear documentation of the system's capabilities and limitations but also explicit declarations if known biases exist within the system. Detailed reports, metadata, or even interface messages can be used to relay this vital information. Providing this level of transparency allows end-users to interpret the system's suggestions or findings more effectively and contextualize them within the broader healthcare scenario. Additionally, openness to third-party scrutiny is crucial to uphold the system's credibility. Allowing experts from diverse fields like data science, ethics, and healthcare to evaluate the system's fairness and bias metrics independently contributes to an extra layer of accountability and quality assurance.

Figure 3. Flow of information and feedback in the ethical framework for bias mitigation



The Principle of Accountability takes the concepts of fairness and transparency a step further by establishing frameworks to correct biases and receive feedback. Specific protocols must be set up to address and correct identified biases, both before the system is deployed and after it becomes operational. This could involve algorithmic adjustments, data re-sampling, or even system re-calibration. A formal procedure must be in place to ensure swift actions to rectify biases, and these actions should be well-documented for future reference and continuous improvement. Furthermore, channels should be created for users to report any perceived biases or disparities in the system's performance. These feedback mechanisms can take the form of digital interfaces, hotlines, or regular audits, and they serve as valuable sources of information for ongoing system refinement.

The Principle of Continuous Monitoring focuses on the importance of post-deployment audits and iterative refinement of the deep learning system's performance in healthcare settings. Once a system is deployed, it must not be considered a finished product; rather, it should be subject to ongoing scrutiny to ensure that it continues to operate fairly and effectively. Regular monitoring of the system in real-world settings allows for the early detection of emerging biases or performance issues that might not have been apparent during the testing phase. This monitoring can take the form of periodic audits, real-time tracking of performance metrics, and review of user feedback. Iterative refinement is another critical aspect of continuous monitoring. Based on the insights gained from post-deployment audits and user feedback, the model should be continuously updated and refined to mitigate biases and improve performance. The introduction of new data and methodologies may also necessitate modifications to the system.

The Principle of Inclusivity stresses the engagement of a diverse group of stakeholders in the system's development and evaluation process. Given the sensitive and complex nature of healthcare, it is imperative to include perspectives from a broad range of individuals. This group should comprise patients, clinicians, ethicists, community members, and particularly representatives from groups that might be at risk of bias. Such diverse engagement ensures that the system is examined through multiple lenses, providing a more thorough and balanced evaluation. Cultural sensitivity is another crucial facet of inclusivity. Medical imaging algorithms, for instance, should be trained to recognize and appropriately interpret cultural variations, such as tattoos or body modifications, which might appear in images. Understanding these variations requires the incorporation of anthropological and sociocultural perspectives in the system's development phase.

Table 5. Ethical framework for bias mitigation in deep learning-based AI systems

Principle	Sub-Principle	Description
Principle of Awareness	Bias Education	Ensure that all stakeholders, from developers to end-users, are educated about the potential for bias in deep learning systems and its implications.
	Bias Detection	Regularly test and evaluate the system to detect any inherent biases. This includes both obvious and subtle biases that might affect performance across different patient groups.
Principle of Representation	Diverse Training Data	Ensure that the training data is representative of diverse populations, considering factors like age, gender, ethnicity, socioeconomic status, and other relevant variables.
	Oversampling and Undersampling	In cases where certain groups are underrepresented in the training data, consider techniques like oversampling (increasing the weight of underrepresented groups) or undersampling (decreasing the weight of overrepresented groups) to balance the dataset.
Principle of Fairness	Equity in Performance	The system should be evaluated to ensure it performs equitably across different patient groups. If disparities are found, they should be addressed before deployment.
	Fairness-enhancing Interventions	Consider integrating algorithms and techniques specifically designed to enhance fairness in machine learning models.

Principle of Transparency	Bias Reporting	Clearly communicate any known biases or limitations of the system to end-users, so they can make informed decisions.
	Openness to Scrutiny	Allow for third-party evaluations of the system's fairness and bias metrics.
Principle of Accountability	Bias Correction	Establish protocols to address and correct biases when they are identified, both pre-deployment and post-deployment.
	Feedback Mechanisms	Create channels for users to report perceived biases or disparities in the system's performance.
Principle of Continuous Monitoring	Post-deployment Audits	Regularly monitor the system's performance in real-world settings to detect and address any emerging biases.
	Iterative Refinement	Continuously refine and update the model to reduce biases, using new data and feedback from real-world deployments.
Principle of Inclusivity	Stakeholder Engagement	Engage a diverse group of stakeholders in the system's development and evaluation process. This includes patients, clinicians, ethicists, community members, and representatives from groups that might be at risk of bias.
	Cultural Sensitivity	Ensure that the system respects and understands cultural differences, which can be crucial in medical imaging (e.g., understanding cultural variations in tattoos or body modifications that might appear in images).

Conclusion

Based on the foundational principles of medical ethics—non-maleficence, beneficence, respect for patient autonomy, and justice—three ethical frameworks are proposed for the deployment and oversight of deep learning systems in healthcare. The first framework focuses on the Ethical Framework for Privacy. Aligned with the principle of autonomy, this framework emphasizes the importance of informed consent, which requires that patients should be fully apprised of how the deep learning system functions, its merits, limitations, and potential risks. Furthermore, the principle of voluntariness ensures that consent is not coerced but freely given, allowing patients the right to withdraw their consent at any time without adverse consequences. Data usage and storage security are given due importance, consistent with the ethical imperative of respecting patient privacy. The framework also incorporates the concept of continual consent and the principles of accessibility and accountability to safeguard patient interests and ensure equitable treatment.

The second framework targets Ethical Framework for Transparency. This framework aligns with the principles of beneficence and non-maleficence by ensuring that stakeholders, including both healthcare providers and patients, are well-informed about the deep learning system. Key sub-principles include full disclosure of algorithmic explanations and performance metrics, understandable simplifications for non-experts, and provisions for third-party audits. Moreover, it lays emphasis on data transparency, accountability, and continuous improvement. The framework aspires to provide all relevant information, making it easier for stakeholders to make informed decisions while also setting up mechanisms for feedback and continuous improvement.
The third framework, Ethical Framework for Bias Mitigation, concentrates on the principle of justice by aiming to make deep learning systems as impartial and equitable as possible. The

framework initiates with the principle of awareness, advocating for educational programs that inform all stakeholders about the potential biases in machine learning systems. To address representational fairness, the framework emphasizes the use of diverse training data and suggests oversampling and under-sampling techniques where needed. The principles of fairness and transparency are invoked to ensure equitable performance across different patient groups and to report any limitations or biases transparently.

The absence of specific legal rules for artificial intelligence (AI) is a major issue . Without set laws, developers have to rely on their own understanding of ethics, which can differ greatly among individuals. For example, what one developer thinks is "fair" in AI could be different from another's view. This lack of standards can lead to inconsistent practices and even allow ethically dubious actions. While laws are important for setting basic ethical standards, they can't cover every ethical problem that comes up. This is similar to healthcare ethics, where laws offer some guidance but can not answer all moral questions.

Additionally, even when technology companies want to tackle ethical issues in AI, they often lack the expertise to do so effectively. Many AI developers have backgrounds in fields like computer science or engineering, where ethics is usually not a major part of their education. This lack of training in ethics makes it more likely that ethical issues will be overlooked. It also makes it difficult to turn broad ethical ideas into specific rules for AI, as that is a complex task requiring specialized knowledge that many developers lack.

The proposed frameworks have some limitations. The first limitation concerning the Ethical Framework for Privacy arises from the dynamic nature of technology and patient conditions, which could make the concept of informed consent a moving target. While the framework emphasizes the need for continual consent, implementing such a process in a practical, efficient manner remains a challenge. Patients may find it overwhelming to keep up with constant updates or changes to the deep learning system and re-evaluate their consent accordingly. Additionally, the issue of data storage and security poses a significant limitation. Despite best efforts to safeguard data, the risk of unauthorized access or breaches remains, and the framework does not provide a detailed roadmap for immediate action in such eventualities.

The second framework, focusing on Transparency, also has limitations. Although it advocates for full disclosure and understandability, the complexity inherent in deep learning models may make it difficult to translate technical aspects into layman's terms without losing critical information. The call for making algorithms or parts of them open source for external validation might also conflict with proprietary interests, making it less feasible in many commercial healthcare settings. Another limitation is the framework's dependency on continuous feedback and iterative updates. While feedback mechanisms are indeed beneficial for system improvement, they may lead to an overemphasis on quantitative metrics, potentially overshadowing qualitative aspects like patient satisfaction or nuanced ethical considerations that are harder to measure.

One primary challenge in our ethical framework for bias mitigation is the difficulty in achieving truly representative training data. Despite calls for diverse data sets, collecting and incorporating such data can be logistically difficult and resource-intensive. Additionally, even if diverse data sets are used, the issue of "hidden biases" or "latent variables," which are not immediately obvious, can still impact the system's decision-making process. The framework also suggests third-party evaluations for fairness and bias metrics. Identifying impartial third-party evaluators who have the requisite expertise in both medical ethics and machine learning can be a challenging task. Furthermore, the principles of cultural sensitivity and stakeholder engagement, while ideal, are difficult to standardize and implement universally, given the wide range of cultural norms and stakeholder interests that may exist in different healthcare settings.

References

[1] G. B. J. Andersson *et al.*, "Do No Harm: The Balance of 'Beneficence' and 'Non-Maleficence,'" *Spine* , vol. 35, no. 9S, p. S2, Apr. 2010.
[2] R. Gillon, "Where respect for autonomy is not the answer," *Br. Med. J.* , vol. 292, no. 6512, pp. 48–49, Jan. 1986.
[3] M. Mawere, "Critical reflections on the principle of beneficence in biomedicine," *Pan Afr Med Jrnl*, vol. 11, no. 1, 2012.
[4] G. M. Trujillo Jr, "Principlism in Biomedical Ethics: Respect for Autonomy, Non-Maleficence, Beneficence, and Justice," *1000wordphilosophy.com,* 1000.
[5] F. Rossi, "Artificial intelligence: Potential benefits and ethical considerations," Oct. 2016.
[6] D. F. Tsai, "Ancient Chinese medical ethics and the four principles of biomedical ethics," *J. Med. Ethics*, vol. 25, no. 4, pp. 315–321, Aug. 1999.
[7] K. H. Keskinbora, "Medical ethics considerations on artificial intelligence," *J. Clin. Neurosci.*, vol. 64, pp. 277–282, Jun. 2019.
[8] C. A. Tschider, "The healthcare privacy-artificial intelligence impasse," *Santa Clara High Tech. LJ*, 2019.
[9] M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AIMag*, vol. 28, no. 4, pp. 15–15, Dec. 2007.
[10] I. Bartoletti, "AI in Healthcare: Ethical and Privacy Challenges," in *Artificial Intelligence in Medicine*, 2019, pp. 7–10.
[11] P.-P. Verbeek, "Materializing Morality: Design Ethics and Technological Mediation," *Sci. Technol. Human Values*, vol. 31, no. 3, pp. 361–380, May 2006.
[12] N. Sendi, N. Abchiche-Mimouni, and F. Zehraoui, "A new Transparent Ensemble Method based on Deep learning," *Procedia Comput. Sci.*, vol. 159, pp. 271–280, Jan. 2019.
[13] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, Nov. 2019.
[14] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019.
[15] J. Torresen, "A review of future and ethical perspectives of robotics and AI," *Frontiers in Robotics and AI*, 2018.
[16] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Artificial intelligence and the 'good society': The US, EU, and UK approach," *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 505–528, Apr. 2018.
[17] N. Bostrom and E. Yudkowsky, "The ethics of artificial intelligence," *Artificial intelligence safety and security*, 2018.