ShodhAI: Journal of Artificial Intelligence
July-September 2026 3(2)
https://shodhai.org

Original Article

EXAMINING LEGAL ACCOUNTABILITY FRAMEWORKS FOR AI IN HEALTHCARE

Lakshmi Walia [1*], **Dr. Namrata Yadav** [2]
[1] Research Scholar, Raffles University, Neemrana, Rajasthan, India
[2] Assistant Professor, Raffles University, Neemrana, Rajasthan, India

ABSTRACT

From AI-powered diagnostic tools in hospitals to autonomous drones in military operations, Artificial Intelligence is rapidly shaping decisions that affect millions of lives every day. Imagine a medical AI misdiagnosing a patient due to faulty training data, or a semi-autonomous weapon acting unpredictably on the battlefield — who is held accountable when such high-risk technologies fail?

This paper explores how India, the European Union (EU), and the United States (US) address the legal accountability of AI systems, particularly in public health and military applications, where the consequences of errors or misuse can be catastrophic. Instead of relying on statistics or coding frameworks, this study uses a qualitative comparative approach to examine laws, policy documents, ethical guidelines, and case studies.

Through a detailed review of legislative texts (such as the EU AI Act, US Executive Orders, and India's evolving AI strategy), court decisions, and international instruments like the OECD AI Principles and UNESCO Recommendation on AI Ethics, the paper identifies common legal gaps and unique regulatory approaches. It highlights how the EU has adopted a more structured regulatory path, the US follows a sector-based approach, and India is at a formative stage, relying heavily on ethical guidelines rather than binding rules.

The paper concludes by proposing a layered accountability framework that blends legal responsibility, ethical oversight, and international cooperation to address the fast-evolving challenges of AI in these critical sectors.

Keywords: Artificial Intelligence (AI), AI Accountability, Legal Liability of AI, High-Risk AI Systems, AI Regulation

INTRODUCTION

Historically, physicians deferred surgery and bloodletting to barbers because such procedures were considered undesirable, or "messy"; no one was troubled by the idea of a barber performing surgery. Today, by contrast, there is considerable concern about AI replacing the physician. Why should AI not instead be viewed as a tool to augment doctor-patient interactions? Although limited by the knowledge contained within its database, a conclusion reached by AI may still be used as a basis for developing a course of action for a patient.

The primary distinction between humans and AI lies in consciousness and intuition. While AI is limited to the information it is given, people draw on various sources of information that are never documented or reported, particularly with regard to health. Patients often do not disclose their full medical condition to health care professionals for fear of embarrassment; an experienced physician, however, can rely on intuition built from years of practice to help make sound medical decisions.

*Corresponding Author:
Email address: Lakshmi Walia (lakshmiwalia97@gmail.com)

Received: 15 July 2026; **Accepted:** 18 August 2026; **Published** 03 September 2026
DOI: 10.29121/ShodhAI.v3.i2.2026.106
Page Number: 36-45
Journal Title: ShodhAI: Journal of Artificial Intelligence
Journal Abbreviation: ShodhAI J. Artif. Intell.
Online ISSN: 3048-9245, **Print ISSN:** 3108-1940
Publisher: Granthaalayah Publications and Printers, India
Conflict of Interests: The authors declare that they have no competing interests.
Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Authors' Contributions: Each author made an equal contribution to the conception and design of the study. All authors have reviewed and approved the final version of the manuscript for publication.
Transparency: The authors affirm that this manuscript presents an honest, accurate, and transparent account of the study. All essential aspects have been included, and any deviations from the original study plan have been clearly explained. The writing process strictly adhered to established ethical standards.

Many deaths around the world result from a lack of properly managed health care, since a limited number of experienced physicians cannot be present in every region to provide care. AI, however, has the potential to extend diagnostic and treatment capability to all parts of the world, based solely on the information it is given.

Ultimately, what matters most is what patients themselves want. Those without access to an effective cure are often willing to rely on AI, while others still prefer a doctor they can turn to and hold accountable if something goes wrong in treatment—a form of accountability that AI, as yet, cannot offer.

This debate ultimately comes down to a single question: what guidelines should govern the use of AI in medicine?

New developments in Medical Artificial Intelligence have moved Healthcare from a rule-based to a data-driven decision-making environment. Early clinical software solutions relied primarily on rule-based systems and executed tasks based on pre-defined instructions irrespective of patient variability. Although they produced predictable results, the lack of adaptability in these systems made it impossible to produce valid outcomes across the continuum of care since they could not accommodate for the complexities of real-world medicine. As opposed to using fixed learning paradigms, AI systems today learn directly from their input data and continually adapt the way they produce output as they receive new input data.

Consequently, while the adaptive capacity of these systems provides a significant capability improvement compared to earlier generations of decision support tools, it also poses an entirely different challenge with regard to issues of responsibility in the event of an error. Modern Medical AI systems operate on several different learning paradigms, with each having unique sets of risks associated with them. Alowais et al. (2023) Supervised learning models are typically used for classifying and predicting diagnoses, with their performance directly related to the quality and quantity of the curated training datasets provided by subject matter experts who generate the labelled training data. Consequently, if the training datasets are biased or do not contain sufficient diversity to represent the entire population of patients, an error made within the supervised learning model will categorically misdiagnose a significant number of patients and as a result, lead to the systematic misdiagnosis of individuals who are part of under-represented populations. Unsupervised learning models are designed to identify anomalous events or discover new insights from their input data. However, similar to supervised learning models, the outputs generated by unsupervised learning models may contain clinically incorrect or misleading information and often generate output based upon knowledge that has not been supported by established clinical knowledge. Reinforcement learning systems introduce yet another level of complexity, due to the fact that they are dynamic in nature as they adjust themselves based upon the response of the patient and the absence of any clarity as to why a particular clinical course of action has been chosen. Nagda (2025)

The hybrid methods which include semi-supervised learning and transfer learning and active learning face data challenges but they still create fundamental problems connected to responsibility. The performance of AI systems degrades because medical data undergoes changes over time and AI systems encounter feedback loops which use AI output to direct clinical choice that creates new training data. The existing problems show that AI functions as a medical system which uses probability to deliver outcomes that depend on data quality and data variety and data management. The development of generative AI systems which can create diagnostic summaries and patient notes and treatment recommendations has sparked intense discussions about who should take responsibility for their professional duties. The systems deliver efficiency improvements to both radiology and clinical documentation yet they create urgent problems about who will be held accountable and how people will understand their operations and whether they can be trusted. The challenge arises because AI-based recommendations affect clinical decision-making which creates difficulty in determining who should take responsibility between developers and healthcare organizations and medical practitioners. The construction of effective accountability systems requires thorough comprehension of the technological restrictions. The research considers AI as an assistive tool which requires both legal protection and ethical handling instead of replacing medical professionals with AI. The healthcare regulatory system needs to establish clear responsibility systems which protect people from AI technology but they must not restrict AI technology adoption.

WHAT IS AI

Artificial Intelligence (AI) means that computers can perform tasks which require human intelligence to function. AI executes tasks through algorithms which programmers create using coding languages. The instructions enable the computer system to process information and produce different output results which include answers and recommendations and verdicts.

AI systems gain their strength because they can handle enormous data volumes while maintaining high processing speeds. AI systems develop their capacity to make real world decisions through human goal creation which allows them to choose their outcomes. Some AI systems function independently without requiring significant human assistance.

Machine Learning (ML) belongs to the field of AI. ML enables computers to acquire knowledge through data analysis which employs mathematical and statistical methods instead of requiring complete manual control. The computer identifies data patterns which it uses to forecast future events and determine operational methods.

The three main categories of machine learning include:

- **Supervised learning:** the computer learns using examples where the correct answer is already known (like teaching with answer keys).
- **Unsupervised learning:** the computer analyzes data to detect patterns without receiving any specific guidance on what to search for.
- **Reinforcement learning:** the computer learns through practice, receiving rewards for correct choices and penalties for incorrect decisions (like training a pet).

Deep Learning (DL) represents an advanced category of machine learning. It employs brain-inspired systems which use multilayered neural networks to function. The layers enable the computer to process information through multiple stages which begin with basic details and culminate in sophisticated concepts. Deep learning typically requires extensive data sets to achieve optimal performance. Alhejaily (2024)

AI systems receive classification according to their available capabilities. The classification system measures AI intelligence levels according to human intelligence standards.

- **Artificial Narrow AI (Narrow AI):** This is the only type of AI that exists today. The system operates according to precise instructions which enable it to execute tasks with greater speed and efficiency than human operators yet its programming limits it to that particular function. The system provides voice assistance through Siri and Alexa while ChatGPT only supports text-based communication.
- **Artificial General Intelligence (AGI):** Theoretical AGI system will learn to execute any human intellectual task by applying previously acquired knowledge to solve new problems across various situations without needing human assistance.
- **Super AI:** This is a theoretical future AI that would possess cognitive abilities surpassing those of human beings. The system would possess reasoning abilities and decision-making skills which enable it to create personal beliefs and desires and emotional states.

AI systems receive classification based on their different operational capabilities. The system displays its processing capabilities through data handling methods and its environmental interaction methods.

- **Reactive Machine AI:** These systems process current information to execute designated functions because they lack memory capabilities. A chess-playing computer demonstrates this concept by analyzing the board during play yet it lacks the ability to remember previous moves.
- **Limited Memory AI:** This AI can recall past events and outcomes for a short period to decide on the best course of action. The majority of contemporary technology products such as self-driving vehicles and generative AI systems depend on this specific operational feature.
- **Theory of Mind AI:** An unrealized functional class where AI would understand human thoughts and emotions, allowing it to simulate human-like relationships. "Emotion AI" is a current area of research aimed at recognizing and responding to human feelings.

Self-Aware AI represents a further, purely theoretical stage in which AI would comprehend human emotions while also possessing its own internal mental states, characteristics, and self-awareness.

The Robotics field uses Narrow AI to create robots which support surgeons during medical operations by monitoring vital signs and detecting potential issues. The systems train on medical data to develop expert systems which mimic human decision-making. The systems analyze extensive data sets to identify patterns which assist healthcare providers in forecasting future health outcomes and determining the reasons behind previous health events. Computer vision technology enables machines to comprehend visual data through object detection and classification capabilities that extend beyond basic image processing. Any AI system that needs to work with the real world or process visual medical data depends on this capability. IBM Data and AI Team (n.d.)

RISKS POSED BY LLMS

The implementation of Large Language Models (LLMs) in medical environments creates multiple hidden yet significant threats to healthcare operations. The systems demonstrate unpredictable behavior because their inherent instability manifests in ways that humans cannot naturally perceive according to the sources. The specific dangers about LLMs include the following elements. Savchuk (2024)

1) The first element of the list describes hallucinations which LLMs create when they generate unrealistic results that appear to be genuine. The models fail to comprehend prompts because they create misunderstanding between human users who present clinical queries that need to be answered. Documented cases of LLMs that performed dangerous drug-dosage miscalculations show how these systems create direct threats to patient safety. The dangers which emerge from AI systems extend beyond the technology itself because they arise from the machine's operational dynamics with medical staff and the clinical environment. The sources establish expert-led supervision as the essential requirement for clinical LLM

management because experts possess the necessary skills to detect potential failures and create protection mechanisms through output limitations and required human authentication before any system actions take place.

RACIAL DISCRIMINATION

Data gathered by artificial intelligence systems can embed racial bias, producing results that perpetuate racial discrimination through inherited stereotypes. This reflects the broader nature and legacy of race correction in medicine, whose origins trace back to the 1800s. In the mid-19th century, for example, the spirometer was introduced and used to support claims of racial superiority by suggesting that Black individuals possessed inferior lung capacity. These theories assumed that genetic makeup determined human traits, while ignoring environmental factors such as malnutrition and substandard living conditions. Shaham et al. (2024)

- **Clinical Impact:** Such changes have downstream ramifications, including critical decisions related to diagnoses, treatment, and referrals. Consider, for example, how race-based changes in kidney function (eGFR) led to artificially inflated scores in Black populations, causing delays in diagnosis and transplants.
- **Scientific Critique:** This process of race correction has been criticized because it seems to lack a scientific basis since race in reality is a sociopolitical construct.

SOLUTION

- In order for racial and other biases in healthcare AI to be overcome, an approach from the entire Total Product Lifecycle (TPLC) must be considered. In this regard, all the sources have suggested that tackling the issue of bias from a technical standpoint only is not enough; rather, the approach has to be combined with representation, data science, and experts.
- **Inclusive Design and Diverse Teams:** The keys to ensuring an unbiased AI begin with teams consisting not just of computer scientists and physicians, but also social scientists, patients, and representatives from marginalized groups themselves.
- **Stakeholder Engagement:** Working within diverse communities during the conception phase of the algorithm allows for the discovery of disease-specific inequities, which can address the divergent challenges of diverse, marginalized populations.
- **Co-production:** Involving patients as active participants in the production of health data can provide the lived experience and contextual knowledge that existing medical data may not account for.
- **Education:** Educating and training the next generation of research leaders to think of race as a sociopolitical construct rather than a biological one is vital in addressing issues of biological essentialism.
- **Improving Data Quality and Variable Selection:** Bias may arise because clinical data have inherited historical underdiagnosis or structural racism.
- **Representative Datasets:** The developer ought to guarantee that the training data reflects the total variety of the source population, including enough members of minority groups.
- **Instead of Race, Causal Drivers:** Rather than using race as a surrogate for biology, algorithmic fairness can be improved by using social determinants of health (SDoH) such as income, housing, and education, or specific biomarkers such as cystatin C.
- **Data Linkage:** Linking patients' electronic health records to external environmental information, educational information, and social information helps to obtain a more comprehensive and unbiased view of health risk.

Technical and Statistical Mitigation: When a bias is detected, various statistical techniques may be employed at different stages of model building:

- **Preprocessing:** This refers to the process where there is rebalancing of the data used before the model starts training, which involves oversampling the data and the Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic data samples.
- **In-Processing:** Techniques learned during training include adversarial debiasing, where the model is penalized if its output can predict certain features such as race.
- **Post-processing:** After building a model, the predictions can be refined or post-processed using techniques like recalibrating the model or setting thresholds for groups to standardize error rates across various population groups.
- **Bias Detection Metrics:** Analysts should also strive to segment their performance metrics, e.g., the C-statistic or AUC, also by racial and ethnicity groups, to better understand where these models may be performing less accurately.
- **Expert Oversight and Risk Management:** The fact that AI systems pose inherent instability means that management is likely to be expert-led.

- **Risk Analysis and Control:** Experts must determine the failure mode of a corresponding technology, such as large language models hallucinating, and provide control by adding safeguards such as those involving validation and constraints.
- **Human-in-the-Loop:** Mandatory human verification, as a guardrail, before an action is performed as suggested by an AI model, forms a critical safety mechanism.
- **Transparent Reporting:** The developers of a CR system must make its underlying codes, datasets, and performance measures available so that all racial groups are supported and public trust can be achieved.
- **Continuous Monitoring:** Once deployed, AI tools must be continually monitored for shifts in data or usage patterns to ensure they remain safe and effective for all patient populations over time.

Beyond racial bias, the integration of artificial intelligence in healthcare introduces a wide array of technical, operational, and ethical risks—ranging from direct patient safety threats to systemic issues concerning privacy and accountability. Hoffman et al. (2016)

TECHNICAL SAFETY AND STABILITY RISKS

AI systems, especially those employing deep learning, tend to be inherently unstable, but this is not always obvious. Bajwa et al. (2021)

- **Hallucinations and Misinterpretation:** Large Language Models (LLMs) can hallucinate results and/or misinterpret the prompt used in the model, thereby generating factually incorrect information. Even though the information in the prompt may have been clearly provided as a clinical scenario, the LLM can misinterpret the information.
- **Critical Calculation Errors:** There are well-reported cases of AI systems critically under-calculating drug doses, which pose an acute physical threat to patients.
- Overfitting: In addition, it is important to know that there exists a phenomenon called overfitting whereby algorithms tend to draw irrelevant information regarding patient characteristics and outcomes. As a result, they will not accurately forecast results when put to use in other situations.
- **Unstable Dependencies:** AI behavior can become unpredictable due to feedback loops, correction cascades, and unstable data dependencies. UNESCO (n.d.)

SOLUTION

- **Expert-Led Risk Management:** For the proper management of risks related to AI, experts are required instead of generalists.
- **Technical Safeguards:** To avoid hallucination and computation mistakes, input validation should be implemented to monitor data quality, and output constraints should be put in place to avoid nonsensical results from the AI system.
- **Mandatory Human Verification:** Create "guardrails" in the medical pathway which ensure that before any AI-based decision or action is taken, there is human sign-off.
- **Post-Market Surveillance:** AI systems will need to be monitored over time to identify "data shifts" or usage patterns that may reveal safety risks.

TRANSPARENCY AND THE "BLACK BOX" PROBLEM

A leading concern revolves around transparency in arriving at conclusions by the use of AI.

- **Interpretability:** Deep learning algorithms are "black boxes," and it is not clear to physicians how they work or what is the rationale behind a particular recommendation.
- **Communication Barriers:** A doctor who does not understand how the AI system arrived at a particular diagnosis may have trouble communicating to a patient how the treatment process works.
- **Scientific and Legal Trust:** Such unexplained decisions may impact trust, which is critical in legal and scientific spheres.

DATA PRIVACY AND SECURITY RISKS

The heavy dependence on large datasets for training AI programs has created a major weakness with regard to patient confidentiality.

- **Privacy Violations:** There is an increased risk of unauthorized data sharing during the collection of large longitudinal data sets; the AI could also potentially infer certain sensitive data about the patient that the patient did not want to share.

Breaches and Misuse: As indicated in various sources, some breaches of privacy in the past have included cases of information-sharing without explicit consent, as well as the misuse of information by third parties. Pham (n.d.)

SOLUTION

Open Reporting and Sharing: Developers must open their shareable reporting and data sets, enabling a third party to evaluate and determine the presence of bias.

Standardized Metrics: Performance metrics, such as the C-statistic or AUC, should be reported by racial or ethnic category rather than as an overall average.

Workforce Training: Training physicians on AI basics is important for enabling them to better comprehend the rationale behind AI decisions.

IMPLEMENTATION AND OPERATIONAL CHALLENGES

While a technically successful AI system may exist, it must also integrate correctly with the medical environment. Workflow Disruption: A large number of physicians dismiss AI tools as an impractical and inconvenient factor of their workflow. Alowais et al. (2023)

SOLUTION

- **Human-Centered Design:** Developers should utilize ethnographic studies to gain an understanding of the particular needs in a healthcare context or environment before creating an AI tool or system.
- **Real-World Validation:** Beyond lab validation, AI trials must be conducted in a prospective design to validate its efficacy in a real-time setting within a hospital environment.
- **Co-Production:** Involving patients in a direct capacity in the design process is essential to guarantee that the data on which the algorithm operates is informed by lived experiences, as opposed to data typically collected in a clinical capacity.

ETHICAL AND SOCIAL RISKS

- **Liability and Accountability:** Determining who is responsible when an AI makes a mistake remains an ongoing challenge. It is unclear if the liability should fall on the physician, the developer, or the healthcare institution.
- **Dehumanization of Care:** Increased reliance on machines may limit the contact and communication between healthcare workers and patients, potentially damaging the therapeutic relationship.
- **Job Obsolescence Fears:** Misconceptions that AI will replace healthcare professionals can lead to skepticism and aversion toward these technologies among staff.
- **Ambition vs. Capability:** Some programs, such as those aimed at complex cancer therapies, have been criticized for having ambitions that exceed their current capabilities, leading to integration difficulties and failed expectations.

ACCOUNTABILITY OF AI

Who is responsible when an AI system causes a failure or does damage? This question requires addressing the legal conflicts between three major entities:13

- The Physician/Clinician: Whether it is just to hold a clinician accountable for an error that was caused by a machine he or she does not necessarily comprehend.
- The Developer: Whether it should rest with those who created the program, even if they are no longer associated with a particular setting where the software failure took place.
- The Institution/Healthcare Organization: The role of the hospital or healthcare system in providing for safe implementation and oversight with respect to the human-in-the-loop guardrails.

1) Transparency and the "Right to Explanation". Egan and Ji (2025)

The "black box" problem-actually the lack of transparency regarding how deep learning algorithms reach decisions-is a central legal parameter. A legal framework must address:

- **Explainability:** The legal requirement of AI to offer a reason for its output in order for clinicians to communicate the treatment process to patients and maintain informed consent.
- **Proprietary vs. Public Interest:** This is a legal dilemma between the developer's right to protect intellectual property and the public interest in algorithms being scrutinized for structural bias or historical inequities.

2) Regulatory Standards and Compliance Frameworks

Legal liability is usually determined by one's conformity to certain set standards. A specific article should contain:

- **Sector-specific standards:** for the UK, this includes DCB0129 for suppliers and DCB0160 for the implementers.
- **The Full Product Life Cycle (TPLC):** The monitoring of AI by Regulatory Authorities right from the design stage up to post-market surveillance.
- **The "Autolearn" Challenge:** How legal frameworks handle algorithms that continually evolve, making their behavior unpredictable over time.

3) Data Governance and Privacy Rights

Accountability also involves how patient data are processed both prior to and following algorithm development. Key parameters include:

- **Consent and Misuse:** Legal repercussions for misusing personal information or sharing patient records without explicit consent, including any cases involving authorities such as the Royal Free London NHS Foundation Trust.
- **Data Integrity:** A legal requirement to ensure the representativity of training data, in order to avoid the "New Jim Code"-that is, where algorithms encode racial bias through proxies for current or historical underdiagnosis or insurance status.

4) Algorithmic Bias as a Legal Liability

Bias is not only a technical error but also a possible legal violation of equity.

- **Historical Error as Precedent:** Legal frameworks should take into consideration the fact that race correction, such as in eGFR or lung function testing, might be at the core of systemic bias leading to delayed diagnosis and withholding treatment.
- **Fairness Metrics:** the use of legal "guardrails" to help establish whether an AI model predicts parity or equality of odds across racial and ethnic groups.

5) Clinical Risk Management and Failure Modes

A legal framework must define the "standard of care" when using unstable technologies.

- **Anticipating Failure Modes:** Accountability for known risks such as hallucinations, misinterpretation of prompts, or calculation errors in drug dosing.
- **Residual Risk Evaluation:** Determining what level of "residual risk" is legally acceptable in a clinical context after technical safeguards (like input validation and output constraints) have been implemented.

GLOBAL FRAMEWORK

DETAILED REVIEW OF LEGISLATIVE TEXTS

Wide variance exists in the approach of sovereign states to AI legislation, from AI-specific comprehensive legislation to adapting existing technology-neutral frameworks. European Commission (n.d.)

1) **The European Union AI Act:** This is considered the world's first comprehensive AI law, embracing a risk-based approach where AI systems are divided into four layers of risk: prohibited risk, high risk, AI triggering transparency requirements, and general-purpose AI. For health, the middle categories are the most relevant, imposing particular obligations that ensure safe and ethical use. It has been enacted, but most requirements are set to take effect from August 1, 2026. UK Government (n.d.)
2) **US Regulatory Position:** The US has no overall federal law that applies particularly to the development of AI. Instead, there are technology-neutral laws at the federal level, such as those related to data protection and equality, in addition to guidance provided by specific agencies. In this regard, the FDA has been a pioneer in approving over 950 AI/ML-enabled medical devices and publishing discussion papers on AI in drug development and manufacturing. The following are some proposed federal legislation: SAFE Innovation AI Framework and AI Research Innovation and Accountability Act.
3) **India's Evolving Strategy:** India's approach is guided by the National Digital Health Blueprint (NDHB), which promotes ethical AI use and equitable care. A unique feature of India's strategy is its focus on scalable, cost-effective AI designed for underserved rural areas. Furthermore, India's draft Personal Data Protection Bill emphasizes data sovereignty, requiring sensitive health data to be stored domestically to protect patient privacy.
4) **China's Governance:** China has published guidelines on the registration of AI-driven medical devices and permissible use cases. These regulations strongly call for human supervision and the "assisting role" of AI in diagnosis and treatment.
5) **Adaptive Model of Japan:** Japan has employed "regulatory sandboxes" managed by the Pharmaceuticals and Medical Devices Agency for controlled testing of new technologies, balancing regulation with the advancement of technology.

International Instruments and Guidelines: International bodies work towards unified global structures for better human lives.

1) **OECD AI Principles:** The 2024 update represents the first intergovernmental standard on AI. It aims at a fine balance between innovation with democratic values and human rights.

2) **WHO Guidelines:** The World Health Organization has published best practices in support of patient privacy and the mitigation of bias. Most notably, the WHO recommendations include setting up no-fault, no-liability compensation funds where patients are compensated for harm without any need for proving fault.

COURT DECISIONS AND ENFORCEMENT TRENDS

Legal accountability through the testing of regulations and the scrutiny of courts is on the rise. Gandhi and Talwar (2023)

- **Data Privacy Violations:** Authorities in charge of privacy protection have emerged as the frontrunners in the AI enforcement race. For instance, Clearview AI has been under heavy scrutiny and enforcement from EU, UK, Canada, and Australia for alleged privacy violations.
- **Training Data Disputes:** The EU's Irish Data Protection Commission investigates whether companies like Meta, Google, and X are allowed to cite "legitimate interests" as the legal basis for training LLMs on users' personal data.
- **Consent Breach:** Historical legal concerns include the Royal Free London NHS Foundation Trust, which came under criticism for having shared patient data without explicit consent.

Common Legal Gaps and Unique Regulatory Approaches: Several critical gaps where current laws struggle to keep pace with AI innovation. United Nations (2025)

- **The "Autolearn" Challenge:** Classical regulations for medical devices presume that a medical device does not change over time, when in fact AI solutions can and do change over time and can "learn" based on new information fed into them.
- **Accountability of Autonomous Systems:** In cases of autonomous systems, establishing fault can be a complex endeavor. When dealing with fully autonomous systems, negligence charges can be attributed to the party developing the AI system because users cannot control it as a human. However, strict liability for the consequences of AI systems can be implemented to safeguard patients, even if this means stifling advancements.
- **Transparency and the "Black Box" Problem:** Because of the "black box" nature of deep learning, it is hard for clinicians to defend their decisions to patients; this elusiveness is a problem because clinicians are legally and emotionally responsible for their decisions, as well as because of its implications for patient trust.

UNIQUE APPROACHES TO REGULATION

- **UK Approach:** The UK does not propose enacting AI-specific legislation but instead encourages the regulators to support the extension of existing software and medical device regulations, known as the "AI as a Medical Device" AIaMD.
- **Shared Responsibility Frameworks:** Some jurisdictions have started to shift liability for flawed algorithms onto the developer, but impose responsibility on healthcare providers in respect of how they incorporate the results from those tools into clinical judgment.

The tripartite model for allocating AI liability offers a conceptual framework for addressing the difficulty of assigning fault when an AI system causes an error in a medical setting. Given the multi-stakeholder nature of AI, accountability is typically distributed among three main entities: the developer of the AI, the healthcare provider (clinician), and the healthcare institution (hospital). Habli et al. (2020)

1) The AI Developer

The primary responsibility may lie with the developer in cases where there was a fault because of design or technical flaws of the algorithm.

- **Autonomous Decision-Making:** In cases in which AI decisions are completely autonomous and not supervised by humans, negligence would likely be ascribed to the developer as there would be no liability for the user, who could not have prevented a failure of duty in something he or she does not control.
- **Flawed Algorithm:** If the algorithm is faulty, such as being biased or having incomplete data, then the developer bears responsibility.
- **Product Liability:** Some frameworks propose that software vendors must secure insurance for product liability claims for errors embedded in software.

2) The Healthcare Provider (Clinician)

- Generally speaking, traditional medical malpractice statutes hold healthcare providers liable for decisions regarding patient care, and this is sometimes true regarding their use of AI systems.
- **Clinical Judgment:** In this case, healthcare provider judgment rather than an AI recommendation is required.
- **Hybrid Decision-Making:** In a "human-in-the-loop" situation, the practitioner may be liable if he or she relies too much on an AI recommendation that causes harm.

3) The Healthcare Institution (Hospital)

The critical layer of responsibility falls on the institution or organization deploying the AI for the environment where AI operates.

- **Implementation and Oversight:** Organizations have the responsibility to see that AI systems are extensively tested, validated, and continuously monitored with due diligence to reduce errors.
- **Systemic Safety:** The sources suggest that often the risks are not just in the AI but in how it interacts with the clinical environment and staff behavior; hence, the institution needs to manage these operational risks.
- **Accountability Standards:** Hospitals can be held liable for failure to establish adequate frameworks of responsibility or deploying technologies that are not adequately transparent.
- **Alternative:** The "No-Fault" Approach

Because identifying a single responsible party in this tripartite structure can be legally complicated and may discourage innovation, the WHO recommended an alternative: no-fault, no-liability compensation funds. This model ensures that patients get faster relief for harm without the need to prove specific fault against any of the three parties.

CONCLUSION

What is needed to meet the rapidly changing challenges posed by AI in highly critical sectors like healthcare is a multi-layered accountability framework that balances innovation with patient safety and human rights. The required framework must integrate legal responsibility, ethical oversight, and international cooperation in a way that will assure AI acts as a trusted "digital health promoter" and not a source of harm. Singer et al. (2016)

Layer 1: Legal Responsibility and Liability

First, clear legal definitions and a structured approach must be considered to deal with the complexity of the "black box" nature of AI.

Shared and Strict Liability: In hybrid human-AI systems, legal frameworks should divide responsibility between developers (for faulty algorithms) and clinicians (for how the tool was used). In fully autonomous systems, negligence should rest squarely with the developer.

Compensation and Insurance: Also, strict liability standard imposed by regulators or no-fault, no-liability compensation funds - as recommended by WHO - to facilitate quick relief for harm. Developers must be required to hold liability insurance.

- **Standardized Definitions:** Legal systems should adopt clear, standardized definitions of key terms—such as "developer," "autonomous system," and "residual risk"—to reduce ambiguity when determining liability.

Layer 2: Ethical Oversight and Fairness-Aware Design

The second layer emphasizes the mitigating actions to combat bias, and human autonomy, focusing on multidisciplinary governance.

Explainable AI (XAI) & Transparency: In order to build trust around AI systems, it is essential for them to shift away from the notion of "black-box" models and move closer to explainable AI techniques that are easily understandable to humans. Shortliffe and Sepúlveda (2018)

- **Bias Mitigation and Algorithmic Audits:** Programmers need to utilize diverse and representative data sets to avoid "The New Jim Code," which leads to the continuation of racial or ethnic disparities. Algorithmic audits or the ethical impact of the program need to become mandatory for new or existing programs.
- **Human-in-the-Loop:** A key principle needs to be designed to help healthcare professionals understand whether the tool is to replace, augment, or assist them. For critical functions like medication dosing or even surgical procedures, human verification is made mandatory.

Layer 3: International cooperation and adaptive regulation

Lastly, the layer tackles the borderless attribute of AI via global norms and access.

- Global Regulatory Harmonization: Entities like WHO, OECD, and UNESCO should take the responsibility to standardize the regulatory policies and procedures. This might include the mutual recognition of such policies to facilitate multinational entities with differing regulations in regions like the EU, US, and Asia.
- Cross Border Data Governance: Treaties can facilitate cross-border data security, and patient information can remain safeguarded through regulations like GDPR, and at the same time, AI can continue its training on diversified datasets from around the globe.
- Equitable Access: Global cooperation in ensuring that advanced countries partner with poor and middle-income countries in technology transfers will help avoid a widening gap in global health.

Thus, with adaptive regulation that follows an iterative policy based on the evolution of AI technologies, it can be ensured that such powerful technologies are used in a responsible and ethical manner. In fact, a holistic approach can make AI an enabler in democratizing healthcare knowledge and improving patient health outcomes across the world while maintaining the highest standards of accountability.

ACKNOWLEDGMENTS

None.

REFERENCES

Alhejaily, A.-M. G. (2024). Artificial Intelligence in Healthcare. [Review article]. PubMed Central. PMCID: PMC11582508; PMID: 39583770.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., and Al Yami, M. S. (2023). Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice. BMC Medical Education, 23, 689. https://doi.org/10.1186/s12909-023-04698-z

Bajwa, J., Munir, U., Nori, A., and Williams, B. (2021). Artificial Intelligence in Healthcare: Transforming the Practice of Medicine. Future Healthcare Journal, 8(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

Egan, D., and Ji, J. (2025, February 11). AI in Healthcare: Legal and Ethical Considerations in This New Frontier. International Bar Association.

European Commission. (n.d.). Artificial Intelligence in Healthcare: Transforming the Future of Medicine. European Commission.

Gandhi, P., and Talwar, V. (2023). Artificial Intelligence and ChatGPT in the Legal Context. Indian Journal of Medical Sciences, 75, 1–2. https://doi.org/10.25259/IJMS_34_2023

Habli, I., Lawton, T., and Porter, Z. (2020). Artificial Intelligence in Health Care: Accountability and Safety. Bulletin of the World Health Organization, 98(4), 251–256. https://doi.org/10.2471/BLT.19.237487

Hoffman, K. M., Trawalter, S., Axt, J. R., and Oliver, M. N. (2016). Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs About Biological Differences Between Blacks and Whites. Proceedings of the National Academy of Sciences, 113(16), 4296–4301. https://doi.org/10.1073/pnas.1516047113

Hulbert, M. (2022). Making AI Work for People of Color: Diagnosing Melanoma and Other Skin Cancers. Melanoma Research Alliance.

IBM Data and AI Team. (n.d.). Artificial Intelligence Types. IBM. https://www.ibm.com/think/topics/artificial-intelligence-types

Murdoch, T. B., and Detsky, A. S. (2013). The Inevitable Application of Big Data to Health Care. JAMA, 309(13), 1351–1352. https://doi.org/10.1001/jama.2013.393

Nagda, P. (2025). Legal Liability and Accountability in AI Decision Making: Challenges and Solutions. International Journal of Innovative Research in Technology (IJIRT), 11(11).

Pham, T. (n.d.). [Article information unavailable]. PubMed Central. PMCID: PMC12076083; PMID: 40370601.

Savchuk, K. (2024, May 15). AI Will Be as Common in Healthcare as the Stethoscope. Stanford Graduate School of Business.

Shaham, T., Schwettmann, S., Wang, F., et al. (2024). A Multimodal Automated Interpretability Agent. In Proceedings of the Forty-First International Conference on Machine Learning.

Shortliffe, E. H., and Sepúlveda, M. J. (2018). Clinical Decision Support in the Era of Artificial Intelligence. JAMA, 320(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 801–810. https://doi.org/10.1001/jama.2016.0287

UK Government. (n.d.). Introduction to the National Commission Into the Regulation of AI in Healthcare. GOV.UK.

UNESCO. (n.d.). Ethics of Artificial Intelligence. UNESCO.

United Nations. (2025, November 19). UN Calls for Legal Safeguards for AI in Healthcare. UN News.