Research Article: Citation date: 30/December/2021

AI Governance in Healthcare: Explainability Standards, Safety Protocols, and Human-AI Interactions Dynamics in Contemporary Medical AI Systems

Shivansh Khanna[1] , Shraddha Srivastava[2]

[1]PricewaterhouseCoopers (PwC), Chicago, IL

[2]Country Financial, Bloomington, IL

Article history:

Received: 22/July/2021
Available online: 30/ December/2021

Abstract

The fast-growing incorporation of artificial intelligence (AI) into the modern healthcare industry necessitates immediate consideration of its legal and ethical dimensions. In this research, we focused on three principal areas requiring specific, contextual direction from both governmental entities and industry participants to guide the responsible and ethical progression of AI in healthcare. First, the research discusses standards for explainability. Within healthcare, understanding AI-driven decisions is vital because of their profound implications for human health. Various participants, from patients to oversight bodies, require differing levels of transparency and explanation from AI systems. Next, we examine safety protocols. Given that employing AI in healthcare could result in decisions that carry severe ramifications, we argue for evaluating its objective criteria, search parameters, training applicability, risk for of poor data, and possible risks. Finally, the dynamics of human-AI interaction were discussed. Optimal interaction necessitates the creation of AI systems that augment human capabilities and acknowledge human cognitive processes. The involvement of AI system users in healthcare, defined through tiers of understanding, contribution, and oversight, spans from elementary to advanced engagements. Each tier relates to the depth of comprehension, the scope of data contribution, and the level of oversight exercised by the healthcare specialist regarding the AI instrument. This research emphasizes the necessity for specific guidelines for each of the three dimensions to guarantee the secure, ethical, and efficient utilization of AI in healthcare.

Keywords: *artificial intelligence, ethical considerations, explainability standards, healthcare, human-AI collaboration, legal considerations, safety considerations*

Authors:

Introduction

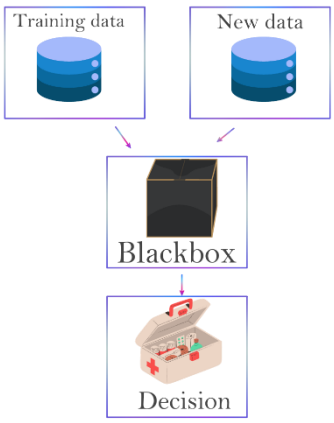
Butcher and Beridze (2019) define AI governance as "*a variety of tools, solutions, and levers that influence the development and applications of AI.*" [1]. AI governance examines the optimal ways for society to adapt to the emergence of sophisticated AI systems. This field encompasses various dimensions, such as political, military, economic, governance, and ethical concerns, all of which are relevant to the impact that advanced AI could have on society.

Perry and Uuk (2019) divided AI governance into 3 sub-components. These include technical environment, which considers how technological advancements are influenced by various

factors and how they, in turn, affect capabilities; ideal governance, which focuses on the best possible actions if full cooperation were achievable; and AI politics, which explores the influence of AI on domestic and international political systems, political economy, and international relations [2].

The integration of Artificial Intelligence (AI) into health care systems presents many ethical, legal, commercial, and social challenges. These challenges are not entirely unprecedented. The embedding of software and computing solutions into health care has been a subject of contention and examination for many years, with key stakeholders, such as developers, governments, and health care providers, constantly grappling with the evolving landscape of technological advancements and their implications. While the foundational issues with integrating technology into health care are not novel, AI introduces a set of unique ethical dilemmas that go beyond the conventional challenges previously encountered. For instance, AI systems, given their capability for self-learning and autonomy, pose questions regarding responsibility, transparency, and decision-making that are distinct from those posed by traditional software.

Figure 1. black box model



Given the profound impact AI can have on both the industrial and societal fronts, governments around the world find themselves in a unique position. Their role isn't merely to observe or regulate but to actively collaborate and guide the evolution of this technology. In many ways, governments act as a bridge, connecting the dots between AI researchers, developers, the industry at large, and the broader public. By fostering a collaborative environment, governments can not only ensure that the development of AI is in line with the broader public interest but can also gain its potential for public welfare, economic development, and societal advancement.

While governments play a central role in guiding the trajectory of AI, the foundation of the technology lies with researchers, developers, and the industry. By collaborating closely with each other, they set the technical parameters, exploring what's feasible and what's not, and laying the groundwork for the technology's practical applications. Their endeavors are crucial in ensuring that AI's evolution is rooted in scientific rigor and innovation.

Yet, while the technical development of AI is primarily in the hands of researchers and industry, its ethical, societal, and regulatory aspects are matters of broader concern. Their combined role is to ensure that as AI systems are developed and deployed, they do so within frameworks that prioritize public interest, safety, and ethics. By setting standards, regulations, and guidelines, they shape the environment in which AI operates. Their involvement ensures that AI does not just evolve as a technological force but does so in a manner that is responsible, ethical, and aligned with societal values.

One of the foundational principles of medical practice is the informed consent of the patient, which is predicated on clear, transparent communication between the healthcare provider and the patient. When a medical decision is made by an AI system that is not interpretable even by professionals, it undermines the possibility of fully informed consent. Patients have the right to

understand the basis upon which medical decisions affecting their health are made. By using uninterpretable AI systems, healthcare providers risk violating this essential right, thereby compromising the ethical quality of care. Additionally, the absence of a clear rationale behind medical decisions may lead to decreased trust in healthcare providers, which could in turn affect treatment adherence and patient outcomes.

When physicians start to rely heavily on black box algorithms, there is a risk that their own clinical skills may atrophy, or that they may become less adept at making these nuanced decisions. This could lead to a form of deskilling among healthcare providers, where the human capacity to diagnose and treat is increasingly deferred to algorithms that cannot be questioned or understood.

These black box algorithms are not infallible and are susceptible to various kinds of errors and biases, often stemming from the data on which they were trained [3,4]. When an AI system makes a mistake or reflects a systemic bias, the lack of interpretability makes it exceedingly difficult to identify the root cause of the error. This is problematic both for patient care and for the broader medical community's understanding of how to improve treatment protocols.

Moreover, the unpredictability of AI is not just a factor of its design but also of the issue between machines and humans. Users can provide inputs or interact with AI in unforeseen ways, leading to unforeseen outcomes. These situations, where AI encounters unfamiliar inputs or is manipulated by users, can sometimes lead to behaviors that are not just unexpected but potentially harmful. This underscores the crucial need for governments and other governing bodies to step in and ensure that AI systems are designed, deployed, and managed with the utmost care, prioritizing the safety and well-being of their users above all else.

Explainability Standards in Healthcare

Healthcare, being a critical domain directly associated with human lives, is an area where the implications of decisions made are exceptionally consequential. The use of Artificial Intelligence (AI) in this sector has brought transformative changes, but it has also ushered in a heightened sense of responsibility, given the direct impact AI decisions can have on patient outcomes. This reality necessitates a stringent standard for explainability [5,6]. Unlike sectors where AI recommendations might dictate marketing strategies or consumer choices, in healthcare, the decisions often mean the difference between life and death, wellness and illness. While in some industries a lack of transparency might be excused for proprietary reasons, or to protect trade secrets, in healthcare, the criteria for clarity and understandability should be decidedly elevated.

The context in which explanations are sought is important. The recipient of the explanation dictates its depth and complexity. For patients, the rationale behind a specific treatment or diagnosis should be clear and understandable, devoid of any complex medical jargon. They deserve to have insights into the decisions that affect their well-being, presented in a comprehensible manner [7,8]. Healthcare providers, on the other hand, require a more detailed breakdown. For them to integrate AI assistance with their expert clinical judgment, they need to perceive the underpinnings of AI outputs. This deep understanding fosters trust and ensures alignment between human and machine decisions. On the regulatory front, auditors and regulators mandate a granular technical view. Their role is to ensure that AI solutions meet the stringent standards of the healthcare sector, making them indispensable gatekeepers of safety and quality.

Time-sensitive scenarios, like emergencies, dictate brevity and precision, ensuring that healthcare providers receive the essential information promptly to make informed decisions. However, the sacredness of medical information means that all communications must respect

privacy norms. Confidentiality isn't just a legal obligation; it's a cornerstone of the trust that patients place in the healthcare system. Additionally, in dynamic environments like operating theaters, real-time interpretations of AI decisions can be of paramount importance, aiding practitioners in making split-second decisions.

Critical decisions, such as diagnosing a life-threatening illness or suggesting an invasive surgery, undeniably call for rigorous explanation. Such decisions have profound implications, and both patients and healthcare providers should be fully informed of the AI's reasoning. In contrast, for routine decisions, like general health check-ups or dietary recommendations, explanations might be simpler, though they still need to be justifiable. Furthermore, the right to contest an AI's decision is fundamental in healthcare. Given the permanent and potentially grave consequences of certain decisions, patients and providers must have avenues to challenge and seek clarification on AI recommendations.

While advanced deep learning models, especially those used in diagnostic imaging, are complex, the burgeoning field of explainable AI (XAI) offers promising solutions to decipher these systems [9,10]. Yet, the financial implications of developing explainable models are significant. Though the stakes in healthcare make the investment in explainability indispensable, it's equally crucial to ensure that the associated costs don't deter the innovation and wider adoption of potentially life-saving AI technologies.

Given the nature of health decisions and the profound ethical considerations involved, a balanced and thoughtful approach to explainability is essential. The context, audience, nature of the decision, and feasibility all play a role in determining the right level of explanation, ensuring that AI becomes a trustworthy and valuable tool in the healthcare sector.

Table 1. Explainability Standards for AI in Healthcare: Perspectives Across Stakeholders

Criteria	Patients	Healthcare Providers	Regulators and Auditors
Recipient of Explanation	Require treatment or diagnosis in simple terms.	Seek detailed explanations to synchronize AI decision with clinical judgment.	Demand deep technical insight for compliance checks.
Timing and Mode of Delivery	Prioritize quick, concise explanations in emergencies; uphold confidentiality.	Value real-time explanations during surgeries/treatments; uphold confidentiality.	Uphold confidentiality and regulatory standards always.
Decision Significance	Expect clarity on high-importance decisions and a right to contest them.	Must align AI suggestions with clinical scenarios and have the ability to contest decisions.	Ensure AI decisions are up to healthcare standards and safety.
Feasibility of Explanation	Prefer simple terms technically; see justifiable cost financially.	Need integration with advanced models technically; prioritize safety financially.	Dive deep into model complexity technically; balance cost and compliance financially.

One viable strategy for enhancing accountability is the development of flagging facilities within the system. In a healthcare, AI applications like diagnostic imaging systems are used by radiologists and doctors to make informed medical decisions. It is crucial for these professionals to have the capability to flag any outputs that appear erroneous or incongruent with their clinical judgement [11,12]. This is particularly important when the AI suggests a diagnosis

that the healthcare provider considers unlikely. Such flagging mechanisms serve dual purposes. On one hand, they allow immediate re-evaluation of the diagnosis or treatment plan to safeguard patient well-being. On the other hand, they provide invaluable feedback for the continual refinement and improvement of the AI system itself, thereby aligning the technology more closely with real-world clinical needs.

If an AI system suggests a specific course of treatment or medication for a patient, it is imperative for patients and their caregivers to have the option to question or contest that decision. For example, if a machine-learning model recommends a particular medication that a patient has previously reacted negatively to, there should be an easily navigable process for expressing these concerns. This procedural avenue not only contributes to the immediate safety of the individual patient but also fosters a broader sense of trust and reliability in AI-assisted healthcare [13,14].

Healthcare is characterized by many variables, including unique patient physiology and complex disease presentations. To account for this, "red teams" of experts can simulate rare or complex clinical scenarios to test the AI system's ability to handle these edge cases effectively. By subjecting the AI system to these stress tests, developers can ascertain the model's ability to make accurate predictions and decisions even in less common or more challenging situations. This type of testing is integral for confirming that the AI system maintains a high level of performance and safety across a range of clinical scenarios.

Given the sensitive nature of medical data and the critical impact of healthcare decisions, audits of AI systems should be rigorous and comprehensive. These audits should scrutinize not just the algorithm but also the data on which it was trained, ensuring diversity and representativeness. Audits should also examine whether the system complies with relevant medical standards and regulations. Such auditing can help to identify and mitigate any unintentional biases in the AI system, ensure it meets legal and ethical guidelines, and ultimately instill confidence among its users.

Documentation of the AI systems employed in clinical settings must be thorough and transparent, covering aspects like validation, efficacy, and potential risks. Documentation is essentially the backbone of auditing and provides the necessary resources for all other accountability measures [15]. In addition, it serves as a historical record, allowing for traceability and offering insights into the evolution of the system's capabilities and limitations. Detailed, accurate documentation is essential for both regulatory compliance and for building user trust in AI applications in healthcare.

Table 2. Alternative ways if providing explanation of AI system is not possible.

Strategy	Detail	Benefit
Flagging Opportunities	AI applications in diagnostic imaging are used by healthcare providers like radiologists and doctors. These professionals should have the ability to flag any questionable or inconsistent outputs for further review.	Allows for immediate re-evaluation of patient diagnosis or treatment, and provides developers with feedback for system refinement.
Opportunities to challenges outcomes	If an AI system suggests a specific treatment or medication, patients and caregivers must have a formalized process to question or challenge these decisions. For example, a process should be in place to	Ensures immediate patient safety and fosters a broader sense of trust and reliability in AI-assisted healthcare.

	raise concerns if a recommended medication has previously caused adverse reactions.	
Adversarial Testing	Due to the complexity and variability in healthcare, "red teams" can simulate rare or complex clinical scenarios to stress-test AI systems. This examines the AI's capability to handle atypical cases effectively.	Confirms the AI system's robustness across a range of clinical scenarios, which is integral for maintaining a high level of performance and safety.
Auditing	Rigorous audits should be conducted to scrutinize the AI system, the training data, and its compliance with relevant medical standards and regulations.	Helps identify and mitigate unintentional biases, ensures that the system meets legal and ethical guidelines, and instills confidence among users.
Documentation	Thorough and transparent documentation must cover aspects like system validation, efficacy, and potential risks. This documentation should be readily available for audit.	Serves as the backbone for all other accountability measures, allows for traceability, and offers insights into the system's capabilities and limitations.

Key Considerations to Ensure Healthcare safety

The AI system may excel in detecting conditions that are already well-understood and have been extensively documented in the training data. However, this approach might be inadequate for identifying rare, uncommon, or novel conditions that deviate from known patterns [16]. Misdiagnosis or missed diagnoses can occur as a result, leading to potentially serious consequences for patient care.

Similarly, an AI tool designed to recommend treatments based on cost-effectiveness may encounter limitations if the objective function is not carefully tailored. Cost-effectiveness is undoubtedly an important factor, especially in healthcare systems where resources are limited. Yet, focusing solely on cost could neglect other important variables such as the patient's individual health condition, co-morbidities, and overall well-being. In such scenarios, the objective function, if too narrowly defined, may lead to recommendations that are economically sound but medically or ethically questionable.

In drug discovery applications, an AI system's exploration space must be cautiously constrained to avoid suggestions that could be harmful to human health. While the system may be highly capable of exploring an extensive range of chemical combinations, it must be programmed to exclude compounds known to be toxic to humans. Such a constraint is non-negotiable even if the excluded compounds might otherwise be effective in treating a particular disease. The potential for effectiveness cannot supersede the basic ethical requirement of ensuring patient safety. Therefore, the exploration space must be sufficiently limited to omit dangerous options, a step that requires careful planning and consideration from the outset of the algorithm's development.

Similarly, surgical robots present another context where the constraints on the exploration space are crucial for ensuring safe and successful operations. A robot tasked with performing surgery should operate within well-defined parameters to mitigate risk. For example, it should not make incisions outside a predetermined region, even if its algorithms identify a potential benefit for doing so. Straying from the defined exploration space in a surgical context can introduce considerable risk of harm, including unintended damage to healthy tissues or organs. Therefore, the operational constraints on surgical robots must be rigorously established and adhered to, regardless of what the algorithmic calculations might otherwise suggest.

In both cases, whether drug discovery or surgical applications, the limitation of the exploration space is a critical component for the safety and efficacy of the AI system. These constraints are not merely technical considerations but are deeply entwined with ethical implications and patient well-being [17,18]. An unconstrained or inadequately constrained exploration space can lead to recommendations or actions that are not only ineffective but potentially harmful.

AI models that are designed to diagnose or predict diseases must be updated regularly to account for evolving strains or new manifestations of diseases. If an AI system is trained on older datasets, its efficacy and reliability may be compromised because it may not recognize newer variations of the diseases it is designed to diagnose. This limitation could result in misdiagnoses, improper treatment plans, or even failure to identify new outbreaks, all of which have severe implications for public health and individual patient care. Hence, ensuring that the training data are up-to-date is essential for maintaining the accuracy and reliability of diagnostic AI systems. Similarly, the diversity of the training data is another significant factor that impacts the model's performance. An AI system trained primarily on data from a specific demographic might perform poorly when applied to different populations. This can result in biased outcomes, exacerbating existing healthcare disparities. Such shortcomings can lead to unequal quality of care and could be considered ethically problematic.

In real-time patient monitoring applications where the AI system is continually updating its model based on incoming data, the introduction of incorrect data could be particularly detrimental. This could occur due to faulty equipment, software bugs, or even malicious actors intending to compromise the system. Given that healthcare decisions are often time-sensitive and have immediate consequences for patient well-being, erroneous data could lead to incorrect clinical decisions [19,20]. Therefore, robust safeguards and validation checks need to be implemented to ensure that the data fed into the system are accurate and reliable. Any anomalies or outliers should trigger alerts for manual review, and the system should be designed to resist incorporating such data until verified.

Similarly, the integrity of patient data is of utmost importance in any healthcare AI application. Data poisoning could lead to corrupted or altered patient records, with serious repercussions that may include misdiagnoses and inappropriate treatments. For example, if a patient's allergy information is incorrectly modified, it could lead to the administration of medications that cause severe allergic reactions. To mitigate this risk, healthcare databases must be secured with the highest levels of encryption and access control. Regular audits should also be conducted to ensure data integrity. In addition, machine learning models should be designed to identify and flag potential data irregularities that might suggest poisoning or corruption.

One area of concern is medical image manipulation, where adversarial attacks can subtly alter the image data in a way that deceives the AI system. For instance, these manipulations can make the system diagnose a condition that isn't present or overlook one that is, leading to incorrect clinical decisions. Such scenarios could have severe consequences, ranging from unnecessary treatments to missed opportunities for early intervention. Therefore, it is crucial to test healthcare AI systems against a range of adversarial inputs designed to exploit their vulnerabilities. These tests can reveal weaknesses in the algorithm's design or in its training data, allowing developers to make necessary adjustments to improve the system's resilience against such attacks.

Healthcare AI systems often manage and store a large volume of sensitive information, making them attractive targets for malicious actors. Adversarial testing can help identify vulnerabilities in how the data is stored, accessed, and transmitted, thereby providing insights into areas that need strengthening to prevent unauthorized access or data breaches. Such testing can also

examine the system's ability to maintain data integrity, ensuring that patient information remains accurate even in the face of attempted attacks.

Given the sensitive nature of healthcare data and the immediate impact of medical diagnoses and decisions, adversarial testing is not merely an optional step but a requirement for healthcare AI systems. Both for ensuring the accurate interpretation of medical images and for safeguarding the enormous quantities of patient data, adversarial testing provides valuable information to improve the AI system's reliability and security. Systematically exposing and addressing vulnerabilities can significantly enhance the overall effectiveness and trustworthiness of these critical technologies.

Table 3 Safety Factors for AI in Healthcare

Key Safety Factors for AI in Healthcare	Specific Considerations
Appropriateness of Objective Function	**Clinical Context**: The objective function for a diagnostic AI tool needs careful selection. For instance, if it is configured to identify anomalies based on their resemblance to known conditions, there is a risk of misdiagnosing rare or new diseases. **Treatment Recommendations**: When AI recommends treatments based on a metric like cost-effectiveness, it must also incorporate individual patient variables and overall well-being to avoid suboptimal recommendations.
Limitations of Exploration Space	**Drug Discovery**: An AI tool developed for discovering new drug combinations should exclude substances that are known to be hazardous to human health. **Surgical Robots**: The operational parameters for a surgical robot should be strictly defined to prevent actions such as making incisions outside of a specified area.
Training Data Relevance	**Evolving Diseases**: AI models must be updated regularly with current data to adapt to new strains or variations of diseases. **Diverse Populations**: The data set used for training must include a range of demographics to eliminate biases and improve performance across different patient groups.
Data Poisoning Risks	**Continuous Learning in Hospitals**: For AI systems that learn in real-time from patient monitoring, measures must be in place to protect against data corruption, whether it originates from faulty equipment or malicious activities. **Patient Data Integrity**: Safeguarding the integrity of patient data is crucial to prevent errors such as incorrect diagnoses or treatments.
Adversarial Testing	**Medical Image Manipulation**: Adversarial tests should be performed to identify vulnerabilities in interpreting medical images. **Patient Data Security**: Adversarial testing can also reveal weaknesses in the storage and transmission of patient data, thereby contributing to its security.

Human-AI Collaboration

The division of labor between humans and machines in healthcare settings is based on their respective strengths and limitations. On one hand, machines excel in data processing tasks that require speed and precision. For instance, machine learning algorithms can sift through vast amounts of medical data—ranging from patient records to laboratory tests—in an incredibly short period [21,22]. These algorithms can identify patterns or anomalies with higher accuracy, thereby aiding in early diagnosis and suggesting potential treatment options. Such capabilities are particularly beneficial in handling complex conditions, where timely and accurate data analysis can substantially influence patient outcomes.

On the other hand, healthcare professionals offer indispensable skills that machines cannot replicate. Emotional intelligence, for example, is critical when it comes to interpreting patients' symptoms and understanding their experiences, which can sometimes be subjective and laden with emotional nuances. A doctor's capacity to empathize with a patient can facilitate better

communication and consequently, more effective treatment plans. Additionally, the experience healthcare professionals accumulate over years of practice equips them with a nuanced understanding of patient care, which includes not just medical knowledge but also ethical considerations and an awareness of the socioeconomic factors that might influence health outcomes.

Table 4. Considerations for Successful Human-AI Collaboration in Healthcare

Design for the different strengths of people and machines	**Clinical Judgment vs. Data Processing**: Machines analyze data; healthcare professionals provide emotional intelligence and experience. Patient Interaction: Chatbots handle routine queries, but human professionals are crucial for emotionally-sensitive discussions.
Successful collaborations are built on communication	**Explaining Diagnostics**: AI systems must clearly explain findings to healthcare providers. Feedback Loop: Medical professionals should provide feedback for system improvement.
Flexibility in role assignment is a boon	**Augmenting, Not Replacing**: AI aids professionals but final judgment rests with humans. Skill Retention: Surgeons use AI for precision, retaining hands-on skills and intervening when necessary.
Design processes with human psychology in mind	**Alarm Fatigue**: AI alert systems must minimize false positives to prevent staff from ignoring genuine emergencies. Staff Training & Acceptance: Staff training and clear communication are vital to foster trust and confidence in AI technology.

Delivering serious diagnoses or discussing end-of-life care are instances where the limitations of machine interactions become palpable. In such situations, the nuances of human emotion, the need for empathy, and the complexities of ethical considerations come into play. These are areas where healthcare professionals have the advantage due to their training in patient-centered care and their ability to understand and manage emotional nuances. A physician can pick up on a patient's non-verbal cues, tailor the delivery of difficult news to the patient's emotional state, and offer immediate, personalized support [23,24]. This sort of nuanced interaction goes beyond the capabilities of current machine-based healthcare interfaces, which are primarily designed for information processing rather than emotional support.

Effective communication forms the foundation of successful collaborations between humans and machines, particularly in healthcare settings where the stakes are often high. For instance, when an AI tool is employed to analyze medical imaging and it detects a potential issue, it is crucial that this tool can articulate its findings in a manner that is easily comprehensible to healthcare providers. Clear and transparent reporting can include the AI providing not just the diagnostic result but also supplementary data such as a confidence score, to indicate the level of certainty of its findings. This enables the healthcare provider to weigh the machine-generated information against other factors like patient history, symptoms, and their own professional judgment. Therefore, system designers should prioritize creating user interfaces that facilitate this level of nuanced communication between machine and human operators.

Feedback mechanisms are another vital component of this collaborative relationship. Healthcare professionals should have an intuitive and straightforward way to provide feedback to the AI system. Whether the AI's diagnostic suggestion was accurate or missed the mark, that information is invaluable for the machine's learning process. By providing feedback, medical professionals can contribute to the system's ongoing training, helping to refine its algorithms and improve its future performance. This creates an iterative feedback loop that

serves to enhance the capabilities of the AI system while also familiarizing healthcare providers with the tool's strengths and limitations.

These technologies should serve to augment human capabilities rather than replace them entirely. For example, in the field of radiology, AI can be extraordinarily useful for flagging potential issues in scans that might require closer scrutiny. However, the final judgment must rest with the human expert—the radiologist. This approach not only leverages the machine's capability for quick and accurate pattern recognition but also ensures that the healthcare professional remains engaged, vigilant, and ultimately responsible for the patient's care. Delegating tasks to machines in this manner can also reduce the cognitive load on healthcare providers, allowing them to focus on more nuanced aspects of diagnostics and patient interaction [25].

Similarly, in the surgical theater, robotic systems have been increasingly incorporated to assist with highly precise tasks. These systems can be particularly useful in minimally invasive procedures or those that require a level of accuracy difficult to achieve by human hands alone. However, it's essential that surgeons remain hands-on with critical aspects of surgical procedures to ensure skill retention. Having the surgeon actively involved also allows for immediate human intervention in case of unexpected complications, something that robotic systems are not yet capable of managing autonomously. Skill retention is essential because dependency on automated systems could lead to atrophy of critical surgical skills that are cultivated over years of rigorous training.

Efforts to investigate the viability of introducing AI robots into the domain of individual care, accompanied by emotional support from human caregivers, merit examination for their transformative potential and possible shortcomings. This exploration can serve as an initial component within the broader context of Human-AI Medical Systems. The proposition of an AI-assisted in-home robot capable of identifying medical exigencies and initiating contact with human caregivers represents a promising avenue for enhancing critical care services.

In hospitals, monitoring systems are ubiquitous and are intended to alert healthcare professionals to potential emergencies. However, if these systems are prone to setting off false alarms, there is a risk that healthcare providers may begin to ignore them, possibly overlooking genuine critical situations. Therefore, when designing AI-based alert systems, it is crucial to minimize false positives. Algorithms should be finely tuned to strike a balance between sensitivity and specificity, ensuring that alerts are both accurate and actionable. This approach takes into account the cognitive load and stress levels of healthcare professionals, thus reducing the risk of alarm fatigue.

The introduction of AI systems in healthcare settings is often met with a mix of anticipation and apprehension among staff. To facilitate a smoother transition, comprehensive training and consultation with healthcare providers are essential. Such initiatives can alleviate fears, clear any misconceptions, and foster a more nuanced understanding of how the technology is meant to assist in their daily tasks. The training process should be designed to make it clear that AI is intended to be a tool that augments human capabilities, rather than a replacement that could make human skills obsolete.

The Nature of AI System Operator's Role in Healthcare

The concept of Basic Interaction with an AI tool in healthcare is characterized by limited user awareness, input, and control. In this paradigm, the healthcare staff has a minimal understanding of the AI's inner workings. For example, a nurse involved in patient scheduling may know that an AI tool is in operation but lacks insights into the algorithms or criteria used to allocate schedules. The level of input from the nurse is confined to observing the already

generated schedule and adding patient details into predetermined fields. Consequently, the nurse operates on a "need-to-know" basis and acts within the bounds set by the AI system. The level of control is also minimal; the nurse cannot alter the AI's parameters or challenge its decisions. This scenario highlights a hierarchical relationship between the AI system and the nurse, where the latter is subservient to the algorithmic decisions made by the former [26]. Basic Interaction can be efficient for tasks that are repetitive and do not require deep clinical insights, but it leaves no room for human judgment or adaptability in complex situations.

In Guided Interaction, healthcare professionals like doctors have a more interactive relationship with the AI tool. Here, the level of awareness extends beyond mere knowledge of the AI's existence; the doctor understands that the AI tool makes use of patient data and symptom analysis to suggest possible diagnoses. When it comes to the level of input, the doctor is enabled to add specific details such as symptoms and patient history to guide the AI's analytical process. This interactive model allows for a balanced approach to decision-making. The level of control is also elevated; the doctor has the discretion to consult the AI tool when deemed necessary and can accept or reject its suggestions. While the AI tool provides a valuable second opinion or shortcuts through medical literature, the ultimate clinical decision rests with the doctor. Guided Interaction thus harmonizes algorithmic assistance and human expertise.

Advanced Interaction signifies a more symbiotic relationship between the AI tool and the healthcare professional. This is evident in scenarios involving radiologists who employ AI for image analysis. In such cases, the level of awareness is advanced, often incorporating a deep understanding of how the AI's algorithms function, including their limitations. The radiologist is trained to recognize conditions where the AI might err, enhancing the reliability of the overall diagnostic process. Regarding the level of input, the radiologist has the flexibility to input detailed image data and even manipulate algorithmic parameters to tailor the AI's analysis to individual patient needs. In terms of control, the radiologist possesses a higher degree of oversight. They can instruct the AI to reassess certain parts of the image or even decide to rely solely on their expertise. The Advanced Interaction paradigm supports a co-piloting model where both the human expert and the AI tool work in tandem for optimal outcomes [27].

Expert Interaction represents the top of AI-human collaboration in healthcare. Specialists in this category have a profound understanding of the AI's decision-making algorithms. Their level of awareness extends to knowing not just the 'what' but also the 'how' of the AI's conclusions. When it comes to input, these experts can adjust parameters, even prioritizing certain genetic markers over others based on the specific case in hand. Control in this interaction level is highly flexible. Operational rules for the AI can be set, modified, or overridden by the specialist. For instance, they can instruct the AI to always flag certain genetic markers irrespective of the overall risk profile or selectively ignore markers based on a patient's unique medical history [28]. The Expert Interaction paradigm is conducive for cutting-edge research and complex medical procedures, where both AI and human expertise are needed for nuanced and context-sensitive decisions.

Table 5. Levels and their corresponding attributes and scenarios in healthcare ai applications

Interaction Level	Level of Awareness	Level of Input	Level of Control
0. Basic Interaction	A nurse knows there's an AI tool assisting with patient scheduling but doesn't know the intricacies of how it determines the schedule.	The nurse can see the schedule and knows where to input patient details, but can't alter the AI's scheduling parameters.	No direct control over the AI's decisions; the nurse operates within the AI's provided schedule.

1. Guided Interaction	A doctor has an AI tool that suggests potential diagnoses. The doctor understands the tool's basics, like that it uses patient data and symptom analysis.	The doctor can input symptoms, patient history, and other relevant information to refine the AI's analysis.	The doctor decides when to consult the AI tool for its opinion and can choose to follow or ignore the AI's suggestions.
2. Advanced Interaction	A radiologist using an AI for image analysis understands in detail how the AI identifies potential issues and has been trained to detect when the AI might make errors.	The radiologist can input detailed image data, adjust parameters based on the patient's specifics, and interpret factors contributing to the AI's conclusions.	The radiologist can override AI-detected issues or ask the AI to reconsider certain areas of the image, and can choose to rely solely on their own judgment if necessary.
3. Expert Interaction	A specialist using AI for genetic analysis can deep-dive into the AI's decision-making process, understanding not just what the AI concludes but how.	The specialist can adjust the AI's parameters, prioritize certain genetic markers over others, and provide context that might influence the AI's genetic risk analysis.	The specialist can set specific operational rules for the AI, like "always flag certain genetic markers irrespective of overall risk profile" or "ignore certain markers for patients with specific medical history".

Conclusion

Currently, AI governance in the healthcare sector represents an unstructured and fragmented domain, lacking clear guidelines or standardized protocols. The diversity of stakeholders, ranging from private corporations and research institutions to regulatory bodies and advocacy groups, contributes to the complexity. Each stakeholder group is driven by its own set of interests and objectives, which often clash with the goals of other entities. For example, companies invested in the development of AI applications for healthcare may push for less restrictive regulations to expedite product development and market entry. Reduced regulatory burdens can accelerate innovation cycles, potentially leading to more efficient and personalized healthcare solutions. However, this kind of unfettered development may also expose the system to risks related to data privacy, ethical considerations, and even patient safety.

The public sector, represented by government agencies and regulatory bodies, is primarily focused on ensuring that the deployment of AI technologies does not compromise patient safety and ethical standards. They advocate for stringent regulations that require rigorous testing and validation of AI algorithms before they are integrated into the healthcare system. Public health organizations may also push for standardization in AI applications to ensure that technology is equally accessible and beneficial across different demographics. The inclusion of ethical considerations, such as fairness and transparency, in the governance model is often more emphasized in the public sector's approach. Thus, while the private sector's objectives are more aligned with rapid innovation and profitability, the public sector's priorities lie in risk mitigation and ensuring equitable distribution of benefits.

The divergence in priorities among different stakeholders creates a challenging environment for establishing a cohesive governance structure for AI in healthcare. One potential solution to bridge this gap is through multi-stakeholder collaborations that aim to reconcile the differing objectives through dialogue and compromise. Such partnerships could facilitate the creation of governance frameworks that balance the need for innovation with the imperative for safety and ethical considerations. This approach could involve the co-creation of guidelines, ethical codes,

and standardized testing procedures for AI healthcare applications. However, achieving this harmonized governance model is not an easy and will require sustained efforts and open dialogue among all involved parties.

Traditional governance mechanisms, such as regulations, taxes, and subsidies, have been the cornerstones of technology policy for years. However, these approaches may not be as effective in addressing the unique challenges posed by AI. For one, the rapid pace at which AI technologies evolve and get adopted poses a substantial challenge for regulatory frameworks that are inherently slow to adapt. Given this fast evolution, information regarding the full scope of risks may not always be available or may become quickly outdated, rendering traditional risk-assessment models less effective. Governments globally are at a crossroads in figuring out how to govern this transformative technology effectively. There is a need for a more detailed understanding of the risks, both predictable and unforeseen, associated with the integration of AI into healthcare. This is important not just for safeguarding patient safety, data privacy, and ethical norms, but also for realizing AI's potential benefits fully.

References

1. Butcher J, Beridze I. What is the State of Artificial Intelligence Governance Globally? The RUSI Journal. 2019;164: 88–96. doi:10.1080/03071847.2019.1694260

2. Perry B, Uuk R. AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk. Big Data and Cognitive Computing. 2019;3: 26. doi:10.3390/bdcc3020026

3. Rai A. Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science. 2020. Available: https://link.springer.com/article/10.1007/s11747-019-00710-5

4. Pedreschi D, Giannotti F, Guidotti R, Monreale A, Ruggieri S, Turini F. Meaningful Explanations of Black Box AI Decision Systems. AAAI. 2019;33: 9780–9784. doi:10.1609/aaai.v33i01.33019780

5. von Eschenbach WJ. Transparency and the Black Box Problem: Why We Do Not Trust AI. Philos Technol. 2021;34: 1607–1622. doi:10.1007/s13347-021-00477-0

6. Yu R, Alì GS. What's Inside the Black Box? AI Challenges for Lawyers and Researchers. Legal Information Management. 2019;19: 2–13. doi:10.1017/S1472669619000021

7. Poon AIF, Sung JJY. Opening the black box of AI-Medicine. J Gastroenterol Hepatol. 2021;36: 581–584. doi:10.1111/jgh.15384

8. Castelvecchi D. Can we open the black box of AI? Nature News. 2016. Available: https://www.nature.com/articles/538020a

9. Holzinger A. From machine learning to explainable AI. 2018 world symposium on digital intelligence for systems and machines (DISA). IEEE; 2018. pp. 55–66. Available: https://ieeexplore.ieee.org/abstract/document/8490530/

10. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R. Explainable AI: Interpreting, explaining and visualizing deep learning. Cham, Switzerland: Springer Nature; 2019. Available: https://books.google.at/books?id=j5yuDwAAQBAJ

11. Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for Explainable AI: Challenges and Prospects. arXiv [cs.AI]. 2018. Available: http://arxiv.org/abs/1812.04608

12. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020;2: 56–67. doi:10.1038/s42256-019-0138-9

13. Gade K, Geyik SC, Kenthapadi K, Mithal V, Taly A. Explainable AI in Industry. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery; 2019. pp. 3203–3204. doi:10.1145/3292500.3332281

14. Doran D, Schulz S, Besold TR. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. arXiv [cs.AI]. 2017. Available: http://arxiv.org/abs/1710.00794

15. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI—Explainable artificial intelligence. Science Robotics. 2019;4: eaay7120. doi:10.1126/scirobotics.aay7120

16. Davahli MR, Karwowski W, Fiok K, Wan T, Parsaei HR. Controlling Safety of Artificial Intelligence-Based Systems in Healthcare. Symmetry . 2021;13: 102. doi:10.3390/sym13010102

17. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ. 2020;98: 251–256. doi:10.2471/BLT.19.237487

18. Falco G, Shneiderman B, Badger J, Carrier R, Dahbura A, Danks D, et al. Governing AI safety through independent audits. Nature Machine Intelligence. 2021;3: 566–571. doi:10.1038/s42256-021-00370-7

19. Borycki EM, Kushniruk AW. The Safety of AI in Healthcare: Emerging Issues and Considerations for Healthcare. In: Househ M, Borycki E, Kushniruk A, editors. Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges. Cham: Springer International Publishing; 2021. pp. 13–22. doi:10.1007/978-3-030-67303-1_2

20. Benrimoh D, Israel S, Fratila R, Armstrong C, Perlman K, Rosenfeld A, et al. Editorial: ML and AI safety, effectiveness and explainability in healthcare. Front Big Data. 2021;4: 727856. doi:10.3389/fdata.2021.727856

21. Grüning M, Trenz M. Me, you and AI – managing human AI collaboration in computer aided intelligent diagnosis. 2021. Available: https://aisel.aisnet.org/sighci2021/12/

22. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nat Med. 2020;26: 1229–1234. doi:10.1038/s41591-020-0942-0

23. Wang D, Churchill E, Maes P, Fan X, Shneiderman B, Shi Y, et al. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–6. doi:10.1145/3334480.3381069

24. Dong Z, Zhang H, Chen Y, Li F. Interpretable Drug Synergy Prediction with Graph Neural Networks for Human-AI Collaboration in Healthcare. arXiv preprint arXiv:210507082. 2021. Available: http://arxiv.org/abs/2105.07082

25. Park SY, Kuo P-Y, Barbarin A, Kaziunas E, Chow A, Singh K, et al. Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing. New York, NY, USA: Association for Computing Machinery; 2019. pp. 506–510. doi:10.1145/3311957.3359433

26. Zhang R, McNeese NJ, Freeman G, Musick G. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. Proc ACM Hum-Comput Interact. 2021;4: 1–25. doi:10.1145/3432945

27. Nikkhah S, Miller AD. AI in the family: Care collaboration in pediatrics as a testbed for challenges facing AI in healthcare. francisconunes.me; 2021. Available: http://francisconunes.me/RealizingAIinHealthcareWS/papers/Nikkhah2021.pdf

28. Xu L, Sanders L, Li K, Chow JCL. Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. JMIR Cancer. 2021;7: e27850. doi:10.2196/27850