AI-Assisted Diagnosis and Liability Attribution: Who is Responsible When Algorithms Err?

BrielleRios

Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

ARTICLE HISTORY
Received: 21 June 2026
Accepted: 28 July 2026

KEYWORDS
AI-assisted diagnosis; medical responsibility; algorithmic ethics; clinical judgment; normative ethics

ABSTRACT
The rapid development of AI-assisted diagnosis is profoundly transforming both the epistemic structure of medical diagnosis and the allocation of responsibility within clinical practice. Unlike traditional medical tools, machine-learning–based diagnostic systems engage deeply in clinical reasoning through algorithmic opacity, continuous learning, and probabilistic outputs, thereby challenging established models of medical responsibility centered on human physicians. When errors occur in AI-assisted diagnosis and result in patient harm, the question of how responsibility should be attributed has become a pressing issue in medical ethics and health law. From a normative ethical perspective, this article systematically examines the distinctive challenges of responsibility attribution in AI-assisted diagnosis. It first elucidates the technical features that position diagnostic AI systems as "cognitive collaborators," blurring the traditional boundary between instruments and decision-making agents. The article then comparatively analyzes major theoretical approaches to responsibility attribution, including physician-centered liability, manufacturer responsibility, distributed responsibility, and no-fault compensation schemes, demonstrating that no single model adequately captures the complex socio-technical realities involved. Through illustrative case analysis, the article further reveals how these approaches may lead, in practice, to unjust attribution of responsibility, dilution of accountability, or barriers to effective patient redress. Building on this analysis, the article argues for a context-sensitive, factor-oriented normative framework for allocating responsibility in AI-assisted diagnosis. Rather than presuming responsibility to rest with any single actor, responsibility attribution should take into account key variables such as the degree of algorithmic explainability, the extent of AI's practical influence within clinical workflows, the system's validation and regulatory status, the maturity of professional guidelines, the nature of the error involved, and whether patients were adequately informed of AI involvement.

I. Introduction

The integration of artificial intelligence (AI) into clinical medicine represents one of the most transformative developments in healthcare history. Over the past decade, AI systems have demonstrated remarkable capabilities in medical imaging interpretation, pathological analysis, and clinical decision support, with some algorithms achieving diagnostic accuracy comparable to or exceeding that of human experts (Esteva et al., 2017; Topol, 2019). From detecting diabetic retinopathy in fundoscopic images to identifying malignant lesions in radiology scans, AI-assisted diagnostic tools have proliferated across multiple medical specialties, fundamentally altering the landscape of clinical practice (Rajpurkar et al., 2017). The global AI healthcare market, valued at approximately $11 billion in 2021, is projected to exceed $187 billion by 2030, underscoring the rapid commercial expansion of these technologies (MarketsandMarkets, 2022).

Despite these promising advances, the incorporation of AI into diagnostic workflows has exposed critical gaps in existing legal and ethical frameworks governing medical liability. Traditional models of medical responsibility were constructed around human decision-makers operating with established tools and technologies. However, AI systems introduce unprecedented complexity: they function as active participants in clinical reasoning rather than passive instruments, they operate through opaque

algorithmic processes that even their developers may not fully comprehend, and they evolve continuously through machine learning mechanisms (Char et al., 2018; Price, 2017). When an AI-assisted diagnosis proves erroneous and results in patient harm, the question of accountability becomes profoundly complicated. Should liability rest with the treating physician who relied on algorithmic recommendations? The software developers who designed the system? The healthcare institution that implemented the technology? Or should responsibility be distributed across multiple actors in ways that current legal doctrine struggles to accommodate?

This question is not merely academic. Real-world incidents have already highlighted the urgency of establishing clear liability frameworks. The widely publicized challenges with IBM Watson for Oncology, which reportedly recommended unsafe and incorrect treatment regimens in multiple cases, illustrate how algorithmic errors can translate into tangible clinical risks (Ross & Swetlitz, 2017). Similarly, concerns about algorithmic bias in diagnostic AI systems have raised questions about who bears responsibility when machine learning models perform inadequately for underrepresented patient populations (Obermeyer et al., 2019). Without coherent liability standards, patients may face barriers to compensation when harmed by algorithmic errors, clinicians may practice defensive medicine by either over-relying on or completely avoiding AI tools, and developers may be discouraged from bringing beneficial innovations to market due to uncertain legal exposure.

The stakes of resolving these liability questions extend far beyond individual cases of medical negligence. Clear accountability mechanisms are essential for protecting patient rights and ensuring access to remedies when harm occurs. They are equally crucial for fostering responsible innovation in AI healthcare technologies by providing developers and healthcare providers with predictable legal standards. Furthermore, well-designed liability frameworks can incentivize the development of safer, more transparent AI systems while promoting their judicious integration into clinical practice (Gerke et al., 2020). Conversely, ambiguous or poorly constructed liability rules risk either stifling beneficial technological progress through excessive caution or enabling harmful applications through inadequate oversight.

This paper addresses the fundamental question: How should legal and ethical responsibility be allocated when AI-assisted diagnostic systems contribute to medical errors? To answer this question, we first analyze the distinctive technical characteristics of AI diagnostic tools that differentiate them from traditional medical devices and complicate conventional liability analysis. We then examine the theoretical frameworks for understanding medical responsibility and identify the potential parties who might bear liability in AI-mediated clinical decisions. Subsequently, we explore competing normative positions regarding liability attribution, evaluate them against real and hypothetical cases, and identify key factors that should influence responsibility assignment. Finally, we propose a comprehensive framework for allocating liability that balances patient protection, technological innovation, and practical enforceability.

This inquiry is timely and essential. As AI systems assume increasingly central roles in diagnostic medicine, the absence of clear liability standards creates uncertainty that affects all stakeholders in the healthcare ecosystem. By developing nuanced approaches to responsibility attribution that account for AI's unique characteristics while remaining grounded in fundamental principles of medical ethics and tort law, we can establish governance structures that harness AI's potential while safeguarding against its risks.

II. The Distinctive Nature Of AI-Assisted Diagnosis

To understand why AI-assisted diagnosis poses unique challenges for liability attribution, we must first appreciate how these systems differ fundamentally from conventional medical technologies. Three technical characteristics distinguish AI diagnostic tools from traditional medical devices and complicate the application of established legal and ethical frameworks: algorithmic opacity, dynamic learning capabilities, and probabilistic output generation.

The "Black Box" Problem: Algorithmic Opacity and Inexplicability

Perhaps the most widely discussed characteristic of contemporary AI systems is their opacity. Many state-of-the-art diagnostic algorithms, particularly those based on deep neural networks, function as "black boxes" whose internal decision-making processes remain inscrutable even to their developers (Castelvecchi, 2016). A radiological AI system trained on millions of chest X-rays may achieve exceptional accuracy in detecting pneumonia, yet provide no comprehensible explanation for why a particular image received a positive classification. The system identifies patterns across thousands of

variables in ways that defy human interpretation, making it impossible to trace the logical pathway from input data to diagnostic conclusion.

This opacity contrasts sharply with traditional diagnostic tools. When a physician uses a stethoscope, an X-ray machine, or even a complex CT scanner, the relationship between the device's function and its output remains transparent and comprehensible. These instruments augment human perception but do not themselves engage in reasoning or inference. The physician can explain precisely how the tool contributed to their diagnostic assessment. With AI systems, particularly those employing deep learning architectures, this explanatory capacity is often absent (London, 2019). The algorithm may detect subtle correlations in imaging data that no human observer would recognize, but it cannot articulate what features or patterns drove its conclusion.

This inexplicability has profound implications for liability. Traditional medical malpractice doctrine requires determining whether a physician's conduct conformed to professional standards of care—an inquiry that presumes the ability to examine and evaluate the reasoning behind clinical decisions. When AI systems contribute opaque recommendations based on inscrutable processes, this analytical framework becomes strained. How can we assess whether a physician exercised reasonable judgment in relying on an algorithmic recommendation if neither the physician nor anyone else can understand the basis for that recommendation? How can developers be held accountable for algorithmic flaws that cannot be identified or explained?

Continuous Learning and Dynamic Evolution

A second distinctive feature of many AI diagnostic systems is their capacity for continuous learning and autonomous evolution. Unlike static medical devices that maintain consistent functionality after manufacturing, AI systems often update their algorithms based on new data, potentially altering their diagnostic behavior over time without human intervention (Char et al., 2018). A deep learning model deployed in a hospital network may refine its pattern recognition capabilities as it processes additional images, adjusting its internal parameters to optimize performance. These updates can occur automatically, through federated learning mechanisms that allow algorithms to learn from distributed data sources while preserving patient privacy.

This dynamic nature creates temporal ambiguity regarding product identity and liability. When a diagnostic error occurs, which version of the algorithm was responsible? The initial design created by developers? The evolved system shaped by institutional data? If an algorithm performs adequately at deployment but deteriorates over time due to dataset shift or inappropriate learning from corrupted data, who bears responsibility for monitoring and preventing such degradation? Traditional product liability doctrine assumes that manufacturers can test and validate their products before release, ensuring they meet safety standards. With continuously learning systems, however, the "product" in question is never truly finalized and may behave unpredictably in ways its creators could not have anticipated (Gerke et al., 2020).

Moreover, the capacity for autonomous evolution challenges the notion of clear causation. If an algorithm's diagnostic accuracy depends partly on the data it encountered during post-deployment learning, then errors might be attributable not solely to original design flaws but also to the particular patient populations and clinical contexts in which the system was deployed. This raises questions about institutional responsibility for providing appropriate training environments and oversight for learning algorithms.

Probabilistic Output and Inherent Uncertainty

Third, AI diagnostic systems typically generate probabilistic rather than deterministic outputs. Instead of definitively declaring "pneumonia present" or "pneumonia absent," an AI system might report "87% probability of pneumonia" or rank multiple potential diagnoses by likelihood. This probabilistic framing reflects the statistical nature of machine learning, which identifies patterns and correlations rather than applying deterministic rules (Rajpurkar et al., 2017).

While probabilistic reasoning is inherent to all medical diagnosis, AI systems make this uncertainty explicit and quantifiable in ways that complicate decision-making and liability attribution. When a physician examines a patient and forms a diagnostic impression, that conclusion incorporates implicit probability assessments based on clinical experience, but is typically expressed with apparent confidence. When an AI system outputs a probability score, it explicitly acknowledges uncertainty and shifts interpretive burden to the human clinician. A physician must then decide what threshold of algorithmic confidence justifies clinical action—is 75% sufficient? 90%? 99%? Different threshold selections will optimize for different outcomes, trading sensitivity against specificity.

This probabilistic framing creates challenges for determining when algorithmic recommendations

constitute "errors" warranting liability. If an AI system assigns 65% probability to a diagnosis that ultimately proves incorrect, did the algorithm fail? Or did it accurately communicate uncertainty that the physician subsequently mismanaged? If a physician follows an AI recommendation with 85% confidence that proves wrong, while ignoring an alternative diagnosis with 15% probability that proves correct, how should we evaluate the reasonableness of that decision? Unlike traditional diagnostic tools that provide objective measurements, AI systems often provide subjective probability assessments that require interpretation—and that interpretation introduces additional sources of potential error and disagreement.

From Passive Instrument to Active Participant

These three characteristics—opacity, dynamic learning, and probabilistic output—collectively transform AI systems from passive diagnostic instruments into active participants in clinical reasoning. A thermometer or blood pressure cuff simply measures physiological parameters; the physician performs all interpretive and integrative work. An AI diagnostic system, by contrast, analyzes complex data, identifies patterns, generates differential diagnoses, and may even recommend specific therapeutic interventions. It engages in tasks that have traditionally defined the core of medical expertise (Topol, 2019).

This transformation from tool to collaborator fundamentally challenges existing liability frameworks predicated on clear separation between human decision-makers and the instruments they employ. When diagnosis emerges from human-AI collaboration rather than human judgment assisted by tools, responsibility becomes inherently shared and difficult to disentangle. Neither purely physician-centered nor purely product-focused liability models adequately capture this collaborative reality.

Clinical Applications and Real-World Complexity

These theoretical considerations manifest in concrete clinical contexts. In radiology, AI systems now routinely assist with interpreting X-rays, CT scans, and MRIs, identifying potential abnormalities and prioritizing worklists (Esteva et al., 2017). Some systems operate as "concurrent readers," providing real-time second opinions alongside radiologists' interpretations. Others function as screening tools, flagging potentially abnormal studies for human review while clearing normal studies without radiologist oversight. The appropriate liability framework may differ significantly between these use cases.

In dermatology, AI applications trained on thousands of skin lesion images can classify suspicious lesions as likely benign or malignant with accuracy rivaling board-certified dermatologists (Esteva et al., 2017). However, these systems have demonstrated concerning performance disparities across different skin tones, performing less accurately for darker-skinned patients due to underrepresentation in training datasets (Daneshjou et al., 2022). When such algorithmic bias contributes to delayed melanoma diagnosis in minority patients, liability analysis must grapple with both technical design choices and systemic inequities.

In pathology, AI systems analyze tissue specimens to detect cancerous cells, grade tumor severity, and predict treatment response. These applications increasingly employ whole-slide imaging and neural networks capable of identifying subtle morphological patterns invisible to human observers. Yet they also depend critically on specimen preparation quality, staining consistency, and imaging parameters—factors outside the algorithm developers' control but within institutional purview (Niazi et al., 2019).

These examples illustrate that AI diagnostic systems exist within complex sociotechnical ecosystems involving multiple actors: algorithm developers, healthcare institutions, regulatory agencies, and clinicians, each contributing to diagnostic outcomes in ways that defy simple causal attribution. Understanding liability in this context requires frameworks sophisticated enough to account for distributed agency while maintaining clear accountability.

III. Theoretical Frameworks for Liability Attribution

Having established the distinctive characteristics of AI-assisted diagnosis, we now examine the theoretical and legal frameworks applicable to analyzing responsibility when these systems contribute to medical errors. This analysis proceeds in three stages: first reviewing traditional medical liability doctrine, then exploring how AI challenges conventional categories of responsibility, and finally identifying the range of potential liability-bearing entities.

Traditional Medical Liability: The Foundation

Medical malpractice law in most common law jurisdictions rests on negligence principles requiring

plaintiffs to establish four elements: duty, breach, causation, and damages (Berlin, 2017). A physician owes patients a duty of care defined by professional standards—specifically, the care that a reasonably competent practitioner in the same specialty would provide under similar circumstances. Breach occurs when the physician's conduct falls below this standard. The plaintiff must then demonstrate that this breach directly caused compensable harm.

Central to this framework is the concept of the "reasonable physician" standard, typically established through expert testimony regarding customary professional practice. Courts evaluate whether the defendant physician's diagnostic reasoning, treatment decisions, and technical execution conformed to what competent peers would have done. This inquiry focuses on the physician's decision-making process: Did they gather adequate information? Consider appropriate differential diagnoses? Exercise sound clinical judgment? The analysis is inherently retrospective but applies a prospective standard—judging decisions based on what was knowable at the time, not what became clear in hindsight (Morreim, 2015).

This traditional framework assumes that physicians exercise substantial autonomy and control over diagnostic and therapeutic decisions. While physicians consult various information sources—medical literature, laboratory results, imaging studies, specialist opinions—ultimate decisional authority and responsibility reside with the treating physician. Medical devices and diagnostic tests serve as tools that inform but do not dictate clinical judgment. The physician remains the primary cognitive agent, integrating diverse information streams into coherent clinical reasoning.

Fault-based liability serves multiple functions within this traditional model. It provides compensation for injured patients while theoretically deterring substandard care by imposing financial consequences for negligence. It also reinforces professional norms and standards, as malpractice litigation helps define the boundaries of acceptable practice. However, critics have long noted limitations of this approach, including barriers to patient compensation, defensive medicine practices, and inadequate attention to systemic factors contributing to medical errors (Mello & Studdert, 2008).

How AI Challenges Traditional Categories

AI-assisted diagnosis strains traditional liability frameworks in multiple respects. First, the opacity and complexity of AI algorithms complicate the determination of whether reliance on algorithmic recommendations constitutes reasonable professional conduct. If a widely adopted AI system generates a diagnostic suggestion that proves erroneous, did physicians who followed that suggestion exercise inadequate independent judgment, or did they reasonably rely on a reputable decision support tool? Conventional doctrine offers limited guidance for evaluating the reasonableness of deferring to algorithmic authority, particularly when the algorithm's reasoning remains opaque (Price, 2017).

Second, AI systems blur the distinction between diagnostic tools and decision-makers. Traditional liability analysis treats medical devices as passive instruments whose proper or improper use affects physician liability but which do not themselves bear responsibility. However, when AI systems actively generate diagnostic hypotheses, assign probabilities to clinical conditions, and recommend interventions, they perform functions previously reserved for human clinical judgment. This makes it conceptually unclear whether AI systems should be governed by device regulation and product liability or by professional liability standards applicable to diagnostic services (Gerke et al., 2020).

Third, the continuous learning capabilities of AI systems create temporal discontinuities that complicate causation analysis. If an algorithm that performed adequately during validation subsequently produces errors due to dataset drift or inappropriate learning from biased data, establishing clear causal links between original design decisions and ultimate harm becomes extremely difficult. The chain of causation extends across time and encompasses multiple contributing factors, including institutional data governance practices and post-deployment monitoring (Char et al., 2018).

Multiple Responsibility Domains

AI-related medical errors may trigger several distinct forms of legal responsibility, each with different standards, procedures, and consequences.

Criminal liability represents the most severe form but applies only in cases of gross negligence or reckless indifference to patient safety. Criminal prosecution of physicians for medical errors remains rare and typically requires egregious departures from standard care combined with clear causation of serious harm or death. AI introduces novel questions about criminal culpability: Could a physician be criminally liable for blindly following clearly unreasonable AI recommendations? Could developers face criminal charges for deploying AI systems known to contain serious flaws? While these scenarios seem remote under current doctrine, they illustrate how AI may eventually require

reconsidering criminal liability standards (Price, 2017).

Civil liability through medical malpractice claims constitutes the primary mechanism for addressing AI-related errors. Patients injured by negligent diagnosis or treatment may seek compensation through tort litigation, typically alleging physician negligence. However, product liability claims against AI developers represent an alternative or complementary avenue, potentially based on design defects, manufacturing defects, or failure to warn about known risks. The relationship between these liability theories—whether they operate independently or interact in complex ways—remains largely unresolved (Gerke et al., 2020).

Administrative liability encompasses regulatory sanctions, professional discipline, and licensure consequences. Physicians who inappropriately rely on or disregard AI systems might face investigation by medical boards or disciplinary action affecting their ability to practice. Healthcare institutions might incur penalties for inadequate oversight of AI implementation. Developers and manufacturers face regulatory requirements through medical device approval processes, with potential sanctions for non-compliance. These administrative mechanisms operate independently of civil liability but significantly influence behavior and establish safety standards (Berlin, 2017).

Finally, ethical responsibility extends beyond legal requirements to encompass professional obligations grounded in medical ethics and fiduciary duties to patients. Even when conduct does not violate legal standards or regulatory requirements, it may breach ethical commitments to patient welfare, informed consent, or professional competence. Ethical responsibility analysis considers whether AI use aligns with fundamental values of medical practice, regardless of legal liability—a crucial distinction when law lags behind technological change (London, 2019).

Potential Responsibility-Bearing Entities

AI-assisted diagnosis involves multiple actors whose decisions and actions collectively shape diagnostic outcomes. Allocating responsibility requires identifying which entities potentially bear liability and under what circumstances.

Treating physicians represent the most obvious candidates for responsibility. They make ultimate treatment decisions, interact directly with patients, and possess professional expertise obligating them to exercise independent clinical judgment. Traditional medical liability doctrine would presumptively assign responsibility to physicians who rely on AI recommendations, under the theory that professional autonomy entails professional accountability. However, this presumption becomes questionable when physicians lack meaningful ability to evaluate algorithmic recommendations due to opacity, or when institutional policies effectively mandate AI use (Price, 2017).

AI developers and manufacturers constitute a second potential locus of responsibility. Under product liability principles, manufacturers may be held strictly liable for defective products that cause harm, regardless of negligence. If an AI diagnostic system contains design flaws, programming errors, or inadequate validation, injured patients might seek compensation from developers. However, applying product liability doctrine to AI systems raises complex questions about what constitutes a "defect" in probabilistic algorithms, how to establish causation when multiple factors contribute to errors, and whether learned behavior after deployment constitutes a manufacturing defect (Gerke et al., 2020).

Healthcare institutions occupy an intermediate position, potentially bearing vicarious liability for employed physicians' negligence as well as direct liability for inadequate AI system selection, implementation, or oversight. Hospitals and clinics decide which AI systems to adopt, establish protocols for their use, provide training to clinicians, and monitor performance. When institutional decisions contribute to AI-related harm—such as deploying inadequately validated systems or failing to establish appropriate safeguards—direct institutional liability may be appropriate (Berlin, 2017).

Regulatory agencies, while not typically liable for specific injuries, bear responsibility for establishing safety standards, validating AI systems before market authorization, and conducting post-market surveillance. Regulatory failures—such as approving defective systems or inadequately monitoring deployed algorithms—might constitute governmental negligence in some circumstances, though sovereign immunity doctrines generally shield agencies from liability (Char et al., 2018).

Finally, a provocative question: Could AI systems themselves bear responsibility? Some scholars have proposed recognizing advanced AI systems as limited legal persons capable of bearing certain obligations and liabilities, analogous to corporate personhood. Under this theory, AI diagnostic systems might maintain insurance, face financial penalties for errors, or be "decommissioned" for persistent poor performance. While this remains largely speculative, it illustrates how thoroughly AI challenges conventional liability categories (Solaiman, 2017).

The proliferation of potential responsibility-bearing entities creates both opportunities and challenges.

It expands the range of legal theories injured patients might invoke to seek compensation, potentially improving access to remedies. However, it also creates complexity, uncertainty, and the risk that responsibility becomes so diffused across multiple parties that meaningful accountability erodes—a concern we explore in subsequent sections examining competing theoretical approaches to liability allocation.

IV. Competing Theoretical Positions on Liability Attribution

Having established the complexity of AI-assisted diagnosis and the inadequacy of traditional liability frameworks, we now examine four major theoretical positions that have emerged in scholarly discourse regarding how responsibility should be allocated when algorithmic errors contribute to patient harm. Each approach reflects different normative commitments regarding the appropriate balance between physician autonomy, technological innovation, patient protection, and practical enforceability.

The Physician-Centered Approach: Preserving Professional Accountability

The physician-centered approach maintains that treating physicians should retain primary responsibility for diagnostic decisions, even when those decisions involve substantial reliance on AI recommendations. Proponents argue that this position best preserves the fundamental structure of medical professionalism and the physician-patient relationship (Kesselheim et al., 2019).

The core argument proceeds from the principle of professional autonomy and its inseparable corollary, professional accountability. Physicians undergo extensive training that qualifies them to exercise independent clinical judgment. This expertise confers both authority to make consequential decisions and responsibility for their outcomes. When physicians choose to incorporate AI recommendations into their diagnostic reasoning, they exercise professional discretion—deciding which tools to employ, how to interpret their outputs, and how much weight to assign algorithmic suggestions. This decisional authority necessarily entails accountability for resulting patient outcomes (Kesselheim et al., 2019).

Advocates emphasize that physicians retain ultimate decision-making power. No AI system can compel a physician to accept its recommendations; the physician always possesses the authority to override algorithmic suggestions based on clinical judgment, additional information, or patient-specific factors. This preserved autonomy justifies continued responsibility. Moreover, physician-centered liability aligns with patients' reasonable expectations. When individuals seek medical care, they enter relationships with human professionals whom they trust to exercise competent judgment on their behalf. Patients expect their physicians, not distant software developers, to bear primary responsibility for their care.

From a practical standpoint, physician-centered liability leverages existing legal infrastructure. Medical malpractice doctrine provides well-established mechanisms for evaluating professional conduct, determining causation, and awarding compensation. Courts and expert witnesses possess expertise in assessing whether physicians met applicable standards of care. Extending this framework to AI-assisted diagnosis requires relatively modest adaptations—clarifying that reasonable physician conduct may include appropriate reliance on validated AI tools—rather than constructing entirely novel liability regimes.

However, critics identify substantial weaknesses in physician-centered approaches. Most fundamentally, they may unfairly burden physicians with responsibility for algorithmic errors beyond their control or comprehension. When AI systems operate as opaque black boxes, physicians cannot meaningfully evaluate the soundness of algorithmic reasoning. Holding physicians liable for accepting recommendations they cannot scrutinize effectively punishes them for limitations inherent in the technology (Price, 2017). This seems particularly unjust when institutional policies or clinical workflows effectively mandate AI use, constraining physician autonomy in practice even if preserving it formally.

Furthermore, physician-centered liability may produce perverse incentives that impede beneficial AI adoption. If physicians face legal exposure for algorithmic errors while receiving no liability protection for appropriate AI use, they may practice defensive medicine by avoiding AI tools entirely, even when those tools would improve diagnostic accuracy. Alternatively, physicians might over-rely on AI to establish that they consulted all available resources, undermining rather than supporting independent clinical judgment. Either response would frustrate AI's potential to enhance healthcare quality (Gerke et al., 2020).

The Manufacturer Responsibility Approach: AI as Defective Product

The manufacturer responsibility approach treats AI diagnostic systems as medical products subject to product liability doctrine. Under this framework, developers and manufacturers bear primary responsibility for algorithmic errors attributable to design defects, inadequate validation, or failure to warn about known limitations (Challen et al., 2019).

Product liability law recognizes three types of defects. Design defects exist when a product's intended design renders it unreasonably dangerous, even if manufactured perfectly according to specifications. An AI diagnostic algorithm might contain design defects if its architecture, training methodology, or validation approach creates systematic inaccuracies. Manufacturing defects occur when specific product units deviate from intended design due to production errors. For AI systems, manufacturing defects might include coding errors, corrupted training data, or improper implementation. Failure to warn defects arise when manufacturers do not adequately inform users about known risks or proper usage limitations. AI developers might incur liability for inadequately disclosing algorithmic limitations, performance variations across patient populations, or appropriate clinical contexts for deployment (Challen et al., 2019).

Advocates argue that product liability appropriately incentivizes safety-oriented development. Manufacturers facing potential liability for defective products will invest in rigorous validation, transparent documentation, and ongoing performance monitoring. This aligns responsibility with control—developers possess far greater ability than physicians to ensure algorithmic accuracy, identify potential failure modes, and implement safeguards. Product liability also addresses the inherent power imbalance between individual physicians and large technology companies. Physicians lack resources to thoroughly evaluate proprietary algorithms, but manufacturers can conduct extensive testing before market release (Price, 2017).

Additionally, product liability provides injured patients with viable compensation mechanisms. Manufacturers typically maintain substantial assets and liability insurance, making them capable of satisfying damage awards. Physician-centered liability may leave patients undercompensated if individual practitioners lack sufficient resources, whereas manufacturer liability ensures adequate compensation sources.

However, critics raise several objections. First, product liability may inadequately account for AI's distinctive characteristics, particularly continuous learning and context-dependent performance. If an algorithm performs adequately during pre-market validation but degrades due to dataset shift in specific clinical environments, classifying this as a product defect seems questionable. The "defect" arguably results from deployment context rather than inherent product flaws (Gerke et al., 2020).

Second, strict manufacturer liability might excessively burden innovation by imposing uninsurable risks on developers. If manufacturers face potentially unlimited liability for unpredictable algorithmic behavior in diverse clinical settings, they may decline to develop or market AI diagnostic tools, particularly for rare conditions or underserved populations where profit margins cannot justify liability exposure. This could impede technological progress and deprive patients of beneficial innovations.

Third, manufacturer liability may inadequately account for physician agency in applying algorithmic recommendations. Even if an AI system contains limitations, physicians might use it appropriately by recognizing those limitations and exercising independent judgment. Conversely, physicians might misuse well-designed systems by applying them in inappropriate contexts or blindly following recommendations without critical evaluation. Pure product liability approaches struggle to incorporate these use-dependent factors into responsibility allocation.

Distributed Responsibility: Shared Accountability Across Multiple Actors

Distributed responsibility frameworks reject the premise that liability must rest primarily with a single party, instead advocating for shared accountability across multiple actors whose decisions collectively shape diagnostic outcomes. This approach acknowledges that AI-assisted diagnosis involves complex collaborations among developers, clinicians, institutions, and regulators, with each contributing to both successful and unsuccessful outcomes (Matthias, 2004).

Proponents argue that distributed responsibility accurately reflects the sociotechnical reality of AI healthcare. Diagnostic errors rarely result from isolated failures by single actors but instead emerge from interactions among multiple contributing factors: algorithmic design choices, institutional implementation decisions, clinical application judgments, and regulatory oversight gaps. Allocating responsibility solely to physicians or manufacturers artificially simplifies this complexity and fails to incentivize safety improvements across the entire system (Matthias, 2004).

Under distributed frameworks, liability allocation depends on each party's causal contribution to harm. Developers might bear responsibility for design flaws or inadequate validation. Healthcare institutions might be liable for deploying algorithms in inappropriate contexts, providing insufficient clinician training, or failing to monitor performance. Physicians might incur liability for negligent application of AI recommendations, such as blindly accepting implausible suggestions or ignoring relevant patient-specific information. Regulatory agencies might face accountability for inadequate oversight or approval of defective systems. The relative responsibility of each party would be determined through causal analysis examining which decisions and actions contributed to the ultimate harm.

This approach offers several advantages. It creates comprehensive accountability by ensuring that all actors whose negligence contributed to harm face appropriate consequences. It generates multi-dimensional incentives for safety, encouraging developers to build robust systems, institutions to implement them carefully, physicians to use them judiciously, and regulators to maintain rigorous oversight. It also provides flexibility to accommodate diverse scenarios—the appropriate liability distribution for algorithmic design flaws differs from that for physician misuse or institutional implementation failures, and distributed frameworks can recognize these distinctions.

However, distributed responsibility faces significant practical and conceptual challenges. Most fundamentally, diffusing liability across multiple parties may dilute accountability to the point of ineffectiveness. When everyone is partially responsible, no one may feel fully accountable, potentially creating moral hazard where each party assumes others will ensure safety (Matthias, 2004). Plaintiffs seeking compensation face increased litigation complexity and costs when they must pursue claims against multiple defendants, each of whom may attempt to shift blame to others. This complexity may render the legal system inaccessible to injured patients who cannot afford protracted multi-party litigation.

Furthermore, distributed responsibility requires sophisticated causal analysis to determine each party's contribution to harm—an exceptionally difficult task given AI's opacity and the multifactorial nature of diagnostic errors. Courts may lack expertise to parse complex technical questions about algorithmic behavior, institutional policies, and clinical decision-making processes. The resulting uncertainty may produce inconsistent outcomes that fail to provide clear guidance for future conduct.

No-Fault Compensation: Risk Spreading Without Individual Liability

The most radical departure from traditional tort principles, no-fault compensation schemes propose eliminating individual liability determinations in favor of collective risk-spreading mechanisms. Under this approach, injured patients would receive compensation from dedicated funds supported by contributions from healthcare institutions, AI manufacturers, physicians, or general taxation, without requiring proof of negligence or causal attribution (Sage, 2004).

Advocates draw parallels to workers' compensation systems, which provide injury benefits without fault determinations, and to New Zealand's comprehensive accident compensation scheme, which covers all personal injuries regardless of cause. Applied to AI-related medical errors, no-fault systems would compensate patients who suffer harm during AI-assisted diagnosis without demanding that they identify responsible parties or prove negligence. Compensation would be triggered by demonstrable injury causally linked to healthcare delivery, not by fault (Sage, 2004).

This approach offers compelling advantages. It eliminates barriers to patient compensation created by the difficulty of proving negligence and establishing causation in complex AI contexts. Patients need not navigate opacity, distributed agency, or technical complexity to receive remedies—they simply demonstrate injury during AI-assisted care. No-fault systems also reduce litigation costs and delays, allowing faster compensation through administrative processes rather than adversarial proceedings.

From a policy perspective, no-fault compensation might encourage broader AI adoption by reducing physicians' and institutions' liability concerns. If healthcare providers know that injured patients will receive compensation without individual fault determinations, they may feel more comfortable using AI tools that could improve outcomes. Similarly, developers might innovate more freely without fearing catastrophic liability for unforeseeable algorithmic behavior. The system spreads costs of inevitable errors across stakeholders rather than concentrating them on particular defendants (Sage, 2004).

However, critics identify serious weaknesses. Most prominently, eliminating individual liability removes powerful incentives for safety and quality improvement. If developers and physicians face no consequences for negligent conduct, they may invest inadequately in algorithmic validation, clinical training, and careful implementation. Tort liability, despite its imperfections, theoretically motivates

actors to prevent harm by creating financial consequences for negligence. No-fault systems sacrifice this deterrent function.

Additionally, no-fault schemes raise challenging questions about funding sources and compensation levels. Who should pay into compensation funds—manufacturers, healthcare institutions, physicians, insurers, taxpayers? How should contributions be calibrated to reflect risk? What compensation levels are appropriate—full damages as in tort, or reduced benefits as in workers' compensation? These design choices profoundly affect fairness and sustainability but lack obvious answers.

Finally, no-fault compensation may be politically unfeasible in jurisdictions with strong tort traditions and public expectations of individual accountability. Many societies expect wrongdoers to face consequences for their negligence. Eliminating this accountability dimension may encounter substantial resistance, even if no-fault systems offer practical advantages.

Each of these theoretical positions illuminates important considerations while exhibiting significant limitations. The physician-centered approach preserves professional accountability but may unfairly burden clinicians for algorithmic failures. Manufacturer responsibility aligns liability with development control but may excessively constrain innovation. Distributed responsibility reflects complex causation but creates practical implementation challenges. No-fault compensation ensures patient remedies but eliminates individual deterrence. The optimal framework likely incorporates elements from multiple approaches, calibrated to specific contexts and use cases—a possibility we explore through case analysis and ultimately in our proposed framework.

V. Case Analysis: Applying Theory to Practice

To ground theoretical discussions in practical reality and evaluate competing liability frameworks against concrete scenarios, we now examine three cases representing different manifestations of AI-related diagnostic errors. These cases—two drawn from actual events and one hypothetical but realistic—illustrate the complexity of responsibility attribution and test the adequacy of various theoretical approaches.

Case 1: IBM Watson for Oncology—Unsafe Treatment Recommendations

In 2017, internal documents revealed that IBM's Watson for Oncology, an AI system designed to assist oncologists with treatment recommendations, had suggested "unsafe and incorrect" cancer therapies in multiple instances (Ross & Swetlitz, 2017). The system, trained partly on hypothetical cases provided by a single hospital (Memorial Sloan Kettering Cancer Center) rather than diverse real-world patient data, reportedly recommended treatments that could have exacerbated patients' conditions or caused serious harm. For example, Watson allegedly recommended prescribing a patient with severe bleeding a drug that would worsen bleeding, a suggestion any competent oncologist would recognize as dangerous.

The Watson case illustrates multiple liability complexities. First, it raises questions about training data adequacy and validation rigor. Watson's reliance on hypothetical cases from a single institution, rather than diverse real-world data, arguably constitutes a design defect—the system's architecture and training methodology rendered it systematically unreliable. Under product liability frameworks, IBM might bear responsibility for marketing a diagnostic product that lacked adequate validation for its intended use. The company's aggressive marketing claims about Watson's capabilities, coupled with insufficient disclosure of validation limitations, might support failure-to-warn claims (Ross & Swetlitz, 2017).

However, physician-centered frameworks would emphasize oncologists' professional obligations to critically evaluate treatment recommendations. The allegedly dangerous suggestions—such as prescribing bleeding-enhancing drugs to hemorrhaging patients—were sufficiently implausible that competent specialists should have recognized them as erroneous. Even if Watson bore responsibility for generating unsafe recommendations, physicians who implemented those recommendations without independent clinical judgment might share liability for failing to exercise reasonable professional care.

Institutional responsibility also deserves consideration. Hospitals and cancer centers that deployed Watson bore obligations to ensure adequate validation before implementation, provide appropriate clinician training, and monitor performance in practice. If institutions marketed Watson to patients as providing cutting-edge AI-assisted care without adequately vetting the system, they might incur liability for negligent implementation. Memorial Sloan Kettering's dual role as Watson's training data provider and commercial partner raises additional conflict-of-interest concerns that might affect

institutional accountability (Ross & Swetlitz, 2017).
Distributed responsibility frameworks appear well-suited to Watson-type scenarios. The unsafe recommendations resulted from multiple contributing factors: IBM's inadequate validation and training methodology, physicians' potential over-reliance on algorithmic authority, and institutional failures to properly vet and monitor the system. Allocating responsibility across these parties based on their respective contributions seems more accurate than assigning primary liability to any single actor. However, this approach faces practical challenges—how should a court apportion responsibility among IBM, individual oncologists, and healthcare institutions? What evidence would establish each party's causal contribution?
Interestingly, the Watson case apparently did not result in major liability claims, partly because the unsafe recommendations were reportedly caught before implementation in most cases. This highlights an important dimension: potential harm versus actual harm. Should liability attach to algorithmic errors that could have caused injury but were intercepted through safety mechanisms? This question has significant implications for incentive structures and accountability.
Case 2: Google DeepMind and the Royal Free Hospital—Data Governance and Algorithmic Bias
In 2017, the United Kingdom's Information Commissioner's Office found that the Royal Free NHS Foundation Trust had inappropriately shared patient data with Google DeepMind for developing an AI system to detect acute kidney injury (Powles & Hodson, 2017). The collaboration involved transferring identifiable information for approximately 1.6 million patients without adequate legal basis or patient notification. Subsequently, research revealed that the developed AI system exhibited performance disparities across different patient demographic groups, with reduced accuracy for certain populations underrepresented in training data.
This case raises distinct liability questions centered on data governance, informed consent, and algorithmic equity. The unauthorized data sharing violated data protection regulations and patient privacy rights. Responsibility for this violation appears to rest primarily with the Royal Free Trust, which controlled the data and bore legal obligations regarding its use. However, Google DeepMind, as the recipient organization that allegedly encouraged or facilitated the inappropriate data sharing, might bear contributory responsibility (Powles & Hodson, 2017).
The algorithmic bias dimension introduces additional complexity. If the AI system performs less accurately for minority populations due to training data deficiencies, who bears responsibility for resulting diagnostic errors? Developer-focused frameworks would emphasize Google DeepMind's obligation to ensure representative training data and validate performance across diverse patient groups. Product liability principles might classify inadequate representation as a design defect that renders the algorithm unreasonably dangerous for certain populations (Challen et al., 2019).
However, the data sharing arrangement complicates simple developer liability. Royal Free Trust provided the training data, and its patient population characteristics shaped the resulting algorithm's capabilities and limitations. If the Trust's patient population inadequately represented broader demographics, the institution arguably contributed to algorithmic bias through data provision. This suggests shared responsibility between developers (who should have recognized representation issues) and data-providing institutions (who should have ensured diverse, representative datasets).
Physician liability in bias scenarios raises particularly troubling questions. If an AI system performs poorly for minority patients due to training data limitations, should clinicians who rely on the system incur liability for resulting errors? On one hand, physicians might argue they cannot reasonably be expected to recognize subtle statistical performance variations across demographic groups, especially when these disparities remain unknown or undisclosed. On the other hand, professional obligations to provide equitable care might require skepticism about AI recommendations for patients whose characteristics differ substantially from typical training data (Obermeyer et al., 2019).
The Royal Free case also illustrates regulatory responsibility. The Information Commissioner's Office ultimately sanctioned the Trust for data protection violations, but only after improper data sharing had occurred and AI development was substantially complete. This reactive approach suggests inadequate proactive oversight. Should regulators bear some accountability for failing to prevent foreseeable problems through prospective governance frameworks?
Informed consent failures add another dimension. Patients whose data trained the algorithm received no notification and provided no consent, violating fundamental medical ethics principles. While this primarily constitutes a privacy violation rather than diagnostic error, it affects the legitimacy of any AI system developed from improperly obtained data. Should algorithmic recommendations derived from non-consensual data be considered per se unreliable or inadmissible in clinical decision-making?

If physicians unknowingly rely on such algorithms, does this affect their liability?
This case supports distributed responsibility approaches that recognize multiple parties' contributions to complex harms involving data governance failures, algorithmic bias, regulatory gaps, and clinical deployment decisions. However, it also reveals how distributed frameworks create accountability diffusion—when Royal Free Trust, Google DeepMind, regulators, and potentially individual physicians all bear partial responsibility, achieving meaningful accountability becomes exceptionally difficult.
Case 3: Hypothetical Scenario—The Dueling Recommendations Dilemma
To explore liability issues not fully captured by existing cases, consider a hypothetical but realistic scenario. Dr. Chen, a radiologist, uses an FDA-approved AI system to assist with interpreting chest CT scans. The AI algorithm flags a subtle pulmonary nodule in a 58-year-old patient's scan and assigns 72% probability of malignancy, recommending biopsy. However, based on the nodule's characteristics, location, and the patient's clinical history, Dr. Chen judges the finding likely represents benign scarring from prior infection. She classifies it as probably benign and recommends watchful waiting rather than immediate biopsy.
Six months later, the nodule has grown substantially. Biopsy confirms advanced lung cancer that has metastasized. The patient requires aggressive treatment and faces significantly worsened prognosis due to the diagnostic delay. Had biopsy occurred when the AI first flagged the nodule, earlier intervention would likely have improved outcomes substantially.
This scenario inverts the typical pattern where physicians uncritically accept erroneous AI recommendations. Here, the AI correctly identified suspicious findings that human judgment dismissed. Who bears responsibility for the resulting harm?
Physician-centered frameworks would emphasize Dr. Chen's professional autonomy and accountability. She exercised independent clinical judgment, disagreeing with the AI recommendation based on her radiological expertise and patient-specific considerations. If her judgment fell below the standard of care—if a reasonable radiologist would have pursued biopsy given the AI's recommendation and clinical context—then she bears liability for negligent diagnosis regardless of AI involvement. The AI's accurate identification of malignancy might actually strengthen the malpractice case by establishing that available diagnostic tools detected the cancer that Dr. Chen missed (Kesselheim et al., 2019).
However, this analysis becomes more complex upon closer examination. What if Dr. Chen's interpretation accorded with majority radiological opinion, such that most peers would have made the same judgment? The AI's superior accuracy in this instance would not necessarily establish that human judgment was negligent—it might simply reflect AI's statistical advantage across large populations. Should the standard of care incorporate AI capabilities, such that failure to defer to algorithmic recommendations constitutes negligence? Or should physicians retain discretion to exercise independent judgment based on expertise and patient-specific factors not captured in algorithms?
Manufacturer liability appears minimal in this scenario. The AI system functioned as designed, correctly identifying a suspicious lesion with appropriate probability assessment. No product defect is apparent. However, a subtle argument might emerge regarding the 72% probability assignment. Was this confidence level adequately calibrated? Did it accurately reflect true malignancy likelihood, or did the algorithm systematically under-estimate certainty? If the AI's probabilistic output misleadingly suggested greater uncertainty than warranted, potentially influencing Dr. Chen's decision, this might constitute a design flaw warranting developer responsibility (Challen et al., 2019).
Institutional responsibility might arise if the hospital provided inadequate guidance regarding how radiologists should respond to AI recommendations. If institutional protocols failed to specify appropriate thresholds for acting on algorithmic suggestions, or if Dr. Chen received insufficient training in integrating AI into diagnostic workflows, the hospital might share liability for creating an environment where physician judgment and AI recommendations could conflict without clear resolution mechanisms.
This case powerfully illustrates distributed responsibility's appeal. The diagnostic delay resulted from complex interactions between algorithmic probability assessment, physician judgment, institutional policies, and the inherent uncertainty of medical diagnosis. No single party acted with clear negligence, yet harm occurred. Distributed frameworks could allocate responsibility based on subtle causal contributions: Dr. Chen's judgment that deviated from AI recommendations, potential AI calibration issues, institutional guidance deficiencies, and perhaps even regulatory failures to establish

clear standards for AI-assisted diagnosis.
Conversely, this case reveals distributed responsibility's weaknesses. If liability becomes too attenuated across multiple parties, it may inadequately compensate the patient or provide insufficient deterrence. The patient suffered real harm and deserves compensation, but from whom? Dividing responsibility among Dr. Chen, the AI manufacturer, and the hospital might leave the patient with fragmented, inadequate remedies while imposing burdensome litigation requirements.
No-fault compensation frameworks would bypass these complexities entirely. The patient suffered injury during AI-assisted healthcare delivery; compensation would be provided without determining whether Dr. Chen, the AI manufacturer, the hospital, or some combination acted negligently. This ensures patient remedy but eliminates any accountability-based incentive for preventing similar errors in future cases (Sage, 2004).
These three cases demonstrate that no single liability framework adequately addresses all AI-related diagnostic error scenarios. Watson illustrates clear design and validation deficiencies suggesting manufacturer responsibility, but also physician and institutional failures. Royal Free highlights data governance and algorithmic bias issues implicating developers, institutions, and regulators. The hypothetical dueling recommendations scenario shows how physician judgment and AI capabilities can conflict in ways that challenge any simple liability assignment. The diversity of these scenarios suggests that flexible, context-sensitive frameworks may be necessary—an insight that informs our subsequent regulatory recommendations.

VI. Key Factors Influencing Liability Attribution

The preceding case analyses reveal that appropriate liability allocation for AI-related diagnostic errors cannot follow a one-size-fits-all approach. Instead, responsibility should be determined contextually based on several critical factors that influence both the causal dynamics of errors and the normative appropriateness of assigning liability to different parties. We identify six key variables that should inform liability attribution in specific cases.
AI Transparency and Explainability
The degree to which an AI system's reasoning process can be understood and evaluated fundamentally affects liability assessment. Highly transparent systems that provide comprehensible explanations for their recommendations enable physicians to exercise meaningful independent judgment. When a diagnostic AI articulates that it identified a pulmonary nodule based on specific morphological features, size characteristics, and statistical correlations with malignancy, clinicians can critically evaluate this reasoning against their own observations and clinical knowledge (London, 2019). In such cases, physician-centered liability becomes more justifiable—the physician possessed adequate information to make an informed decision about whether to accept or reject the algorithmic recommendation.
Conversely, opaque "black box" systems that provide no explanation beyond probability scores severely constrain physicians' ability to exercise independent judgment. When an AI simply declares "78% probability of pneumonia" without indicating what features drove this assessment, physicians cannot meaningfully evaluate the recommendation's validity. Holding physicians fully liable for accepting or rejecting such opaque recommendations seems unfair, as they lack the information necessary for informed decision-making. In these scenarios, manufacturer responsibility becomes more appropriate, as developers created systems that preclude meaningful human oversight (Amann et al., 2020).
Regulatory frameworks increasingly recognize this distinction. The European Union's AI Act proposes heightened requirements for "high-risk" AI systems, including transparency and explainability mandates for healthcare applications. These regulatory developments suggest that explainability should affect not only liability allocation but also market authorization—systems below minimum transparency thresholds might be deemed per se inappropriate for clinical deployment.
Physician Dependence and AI's Role in Clinical Workflow
The manner in which AI systems are integrated into clinical workflows significantly affects appropriate liability distribution. AI applications exist on a spectrum from purely consultative tools that provide optional second opinions to deeply embedded systems that effectively mandate algorithmic recommendations (Finlayson et al., 2021).
At the consultative end, physicians voluntarily choose to seek AI input as one of multiple information sources informing their independent judgment. When AI functions this way—similar to requesting a

specialist consultation—physician-centered liability seems appropriate. The clinician retains clear decisional autonomy and can weigh algorithmic recommendations against other considerations. Failure to critically evaluate AI input under these circumstances reflects inadequate professional diligence.

However, many institutional implementations create substantial pressure or requirements for physicians to follow AI recommendations. Workflow designs may route AI-flagged cases differently, create documentation burdens for deviating from algorithmic suggestions, or subject physicians who frequently override AI to administrative scrutiny. In such environments, physicians' nominal autonomy becomes practically constrained. Liability frameworks must account for these institutional dynamics—assigning full responsibility to physicians who face significant institutional pressure to defer to AI seems unjust and ignores hospitals' role in shaping clinical decision-making (Price, 2017).

The most extreme scenario involves fully automated diagnostic systems where AI recommendations bypass human review entirely for certain cases. Some radiology AI systems automatically clear normal chest X-rays without radiologist interpretation, routing only abnormal studies for human review. In such workflows, physicians exercise no judgment over AI-cleared cases. Manufacturer and institutional liability become predominant, as physicians have no opportunity to prevent errors in automated pathways.

Validation Status and Regulatory Approval

An AI system's validation status and regulatory clearance significantly affect liability analysis. FDA-approved medical devices undergo rigorous premarket review demonstrating safety and efficacy. Physicians who rely on properly approved AI systems can reasonably expect them to meet established performance standards. If an FDA-cleared algorithm produces erroneous recommendations despite proper use, manufacturer liability becomes more compelling—the regulatory approval process should have identified significant defects (Benjamens et al., 2020).

Conversely, physicians who employ non-validated, experimental, or "off-label" AI systems assume greater responsibility for evaluating their reliability. Using an AI algorithm that lacks regulatory clearance or applying an approved system outside its validated scope constitutes a higher-risk practice requiring enhanced professional diligence. Errors occurring in such contexts may primarily reflect physician negligence in using unproven tools rather than manufacturer defects.

However, regulatory status alone cannot determine liability. Even FDA-cleared devices may contain defects that emerge post-market or perform inadequately in specific populations despite aggregate validation data. The approval process provides important information about AI reliability but does not guarantee perfection or eliminate all uncertainty about appropriate use.

Clinical Practice Guidelines and Professional Standards

The existence and content of professional guidelines regarding AI use substantially influence the standard-of-care analysis central to physician liability. When specialty societies issue clear recommendations about appropriate AI applications—specifying validated use cases, necessary safeguards, and circumstances requiring human override—these guidelines establish benchmarks for evaluating physician conduct (Geis et al., 2019).

Unfortunately, professional guidance for AI-assisted diagnosis remains underdeveloped in most specialties. This creates uncertainty about what constitutes reasonable physician behavior when using AI tools. Should radiologists routinely defer to AI recommendations that exceed certain confidence thresholds? Must they independently verify all algorithmic findings? Are they obligated to understand AI systems' technical details, or can they reasonably rely on vendor representations and regulatory approvals? Absent clear professional consensus, courts struggle to define the applicable standard of care.

As AI becomes more prevalent, professional societies must develop specific guidance addressing these questions. Well-crafted guidelines would clarify physicians' responsibilities when using AI, including: required competencies and training, appropriate reliance thresholds, circumstances necessitating independent verification, and documentation standards. Such guidance would provide physicians with clearer expectations while giving courts more definitive standards for evaluating professional conduct.

Informed Consent and Patient Knowledge

The quality of informed consent regarding AI involvement affects ethical if not strictly legal dimensions of liability. Medical ethics requires that patients receive adequate information about their care to make informed decisions. When AI systems substantially influence diagnosis or treatment, do patients have a right to know about algorithmic participation? Should they be informed about AI's

capabilities, limitations, and potential biases? (Gerke et al., 2020)

Current practice varies widely. Some institutions prominently disclose AI use and even market it as a quality enhancement. Others employ AI without patient notification, treating it as an internal clinical tool comparable to laboratory equipment. This inconsistency raises questions about informational duties and their relationship to liability.

Robust informed consent might affect liability in several ways. First, it could establish shared decision-making that distributes responsibility between physicians and patients. If patients knowingly choose AI-assisted diagnosis understanding its benefits and risks, they arguably assume some risk of algorithmic error—though this hardly eliminates physician or manufacturer liability for negligence. Second, inadequate disclosure might constitute independent ethical violation or battery, creating liability regardless of diagnostic outcomes. Third, patient knowledge affects causation analysis—informed patients might have declined AI-assisted diagnosis or sought second opinions, potentially avoiding harm.

As AI becomes routine in clinical practice, establishing clear informed consent standards becomes essential. Patients deserve transparency about how AI affects their care, and this transparency should inform liability frameworks that balance patient autonomy, professional responsibility, and practical feasibility.

Error Typology: Design Defects, Data Bias, or User Error

Finally, liability attribution should reflect error etiology—the root cause of algorithmic failures. Three primary error categories warrant distinct treatment.

Design defects arise from flawed algorithmic architecture, inadequate validation methodology, or poor software engineering. An AI system trained on insufficient data, using inappropriate machine learning techniques, or lacking basic error-checking mechanisms contains design defects attributable primarily to developers. Product liability frameworks appropriately address such cases (Challen et al., 2019).

Data bias reflects systematic inaccuracies resulting from non-representative training data or discriminatory patterns in historical datasets. When AI performs poorly for minority populations because training data underrepresented those groups, responsibility may be shared between developers (who should ensure representative data) and data-providing institutions (who may have supplied biased datasets). Physicians using biased algorithms face complex questions about their obligations to recognize and compensate for systematic performance variations (Obermeyer et al., 2019).

User error involves inappropriate AI application by clinicians or institutions—using algorithms outside validated scope, ignoring clear contraindications, or failing to exercise reasonable judgment when accepting recommendations. Such errors primarily reflect physician or institutional negligence rather than product defects, supporting physician-centered or institutional liability.

Distinguishing these error types in practice can be challenging, particularly given AI's opacity. However, this taxonomy provides a useful framework for causal analysis that should inform responsibility allocation. Clear design defects warrant manufacturer liability; obvious user errors support physician/institutional liability; and data bias scenarios may require distributed responsibility recognizing multiple parties' contributions to systematic inaccuracies.

VII. Normative Recommendations: Constructing a Comprehensive Liability Framework

Having analyzed competing theoretical positions, examined illustrative cases, and identified key variables affecting liability attribution, we now propose a comprehensive framework for allocating responsibility when AI-assisted diagnostic systems contribute to patient harm. Our recommendations synthesize insights from preceding sections while recognizing that effective governance requires balancing multiple objectives: protecting patient welfare, incentivizing safety and innovation, preserving physician professionalism, and ensuring practical enforceability.

Foundational Principles

Any viable liability framework for AI-assisted diagnosis must rest on four foundational principles that guide specific policy choices.

First, patient protection must remain paramount. Liability rules serve multiple functions—deterring negligence, compensating victims, establishing professional standards—but their primary justification lies in safeguarding patient welfare. When design choices create tensions between patient protection and other objectives like innovation or professional autonomy, patient interests should presumptively prevail. This means maintaining accessible compensation mechanisms, avoiding liability rules that enable harmful AI deployment, and ensuring that responsibility gaps do not leave injured patients

without remedies (Gerke et al., 2020).

Second, liability frameworks should balance innovation incentives with safety imperatives. Excessive manufacturer liability might deter beneficial AI development, particularly for applications serving small patient populations or addressing rare conditions where profit margins cannot justify major liability exposure. Conversely, inadequate liability allows harmful products to reach markets and patients. The optimal framework encourages responsible innovation through predictable liability standards that reward safety investments while ensuring accountability for negligent or reckless conduct (Price, 2017).

Third, responsibility should align with control and capability. Parties possessing greater ability to prevent harm should bear commensurate responsibility. Developers who design algorithms can implement validation protocols, ensure training data quality, and provide adequate user guidance—they should be accountable for failures in these domains. Physicians who make clinical decisions should answer for professional judgment errors. Institutions that deploy AI systems should be responsible for implementation quality and oversight. This principle of matching responsibility to control promotes both fairness and efficacy (Amann et al., 2020).

Fourth, liability frameworks must be sufficiently clear and operational to guide behavior and enable consistent adjudication. Excessively vague or complex liability rules fail to provide the predictability necessary for prospective compliance and create unacceptable arbitrariness in outcomes. While some contextual flexibility is essential given AI's diversity, core liability principles should be comprehensible to developers, healthcare providers, and patients, and administrable by courts without requiring exceptional technical expertise in every case.

Stakeholder-Specific Responsibilities

Building on these principles, we propose differentiated responsibilities for each major stakeholder group.

Physician Responsibilities: Treating physicians should retain primary responsibility for clinical decisions involving AI recommendations, but this responsibility should be calibrated to their realistic capabilities. Physicians must exercise competent professional judgment in deciding whether to accept, modify, or reject algorithmic recommendations based on patient-specific factors, clinical context, and their expertise. However, physicians should not be held liable for failing to identify algorithmic flaws that require technical expertise beyond clinical training or for following recommendations from properly validated, FDA-approved systems used within their authorized scope (Kesselheim et al., 2019).

This framework requires establishing an updated "reasonable physician" standard that incorporates AI literacy while remaining realistic about clinicians' technical limitations. Physicians should be expected to: (1) understand AI systems' general capabilities and limitations, (2) critically evaluate whether specific patients fall within an algorithm's validated scope, (3) exercise independent judgment rather than reflexively accepting all AI recommendations, and (4) recognize obviously implausible or contradictory algorithmic outputs. However, physicians need not understand proprietary algorithmic details, identify subtle statistical biases, or second-guess properly validated systems absent specific reasons for concern.

Professional societies should develop specialty-specific guidance clarifying these expectations, and licensing boards should ensure that continuing medical education includes AI competency components. Malpractice doctrine should then evaluate physician conduct against these articulated standards.

Manufacturer Responsibilities: AI developers and manufacturers should bear strict liability for design defects, inadequate validation, and failure to warn about known limitations. This means injured patients can recover damages by demonstrating that an AI system contained defects that caused harm, without proving developer negligence. This strict liability standard appropriately reflects developers' superior ability to ensure product safety and their profit from AI commercialization (Challen et al., 2019).

However, several important qualifications apply. Manufacturers should be liable only for uses within an AI system's validated and authorized scope. If physicians deploy algorithms in contexts explicitly identified as inappropriate in labeling or training materials, manufacturer liability should be reduced or eliminated. Additionally, manufacturers should receive safe harbor protection for disclosed limitations—if an AI system's documentation clearly warns about specific failure modes or performance limitations, and harm results from such disclosed risks, manufacturer liability should be reduced absent evidence that the disclosure itself was inadequate.

To operationalize this framework, regulatory standards must clearly define minimum validation requirements, transparency expectations, and disclosure obligations. The FDA and comparable international regulators should establish tiered requirements based on AI risk levels, specifying validation study designs, performance benchmarks, and post-market surveillance obligations. Compliance with these regulatory standards should create a rebuttable presumption of adequate care, while non-compliance should support liability findings.

Institutional Responsibilities: Healthcare institutions should bear direct liability for negligent AI system selection, implementation, and oversight. This includes: (1) conducting independent validation of AI systems before clinical deployment, (2) ensuring adequate clinician training in AI use, (3) establishing appropriate clinical workflows that preserve human judgment while leveraging AI capabilities, (4) monitoring AI performance in institutional contexts, and (5) implementing error-reporting and quality-improvement mechanisms (Finlayson et al., 2021).

Institutional liability should be particularly robust when hospitals or health systems create structural pressures that effectively mandate physician deference to AI recommendations. If institutional policies penalize physicians for overriding AI or create workflow designs that make independent judgment practically impossible, institutions should share liability for resulting errors even if physicians nominally retained decisional authority.

Institutions should also bear responsibility for algorithmic bias that emerges from or is exacerbated by institutional data practices. If a hospital's patient population or data collection practices create training datasets that perpetuate discriminatory patterns, the institution shares responsibility for resulting performance disparities across demographic groups.

Regulatory Responsibilities: While regulatory agencies typically enjoy sovereign immunity from liability, they bear crucial responsibility for establishing and enforcing AI safety standards. Regulators should: (1) develop clear, risk-based classification schemes for AI medical devices, (2) establish stringent premarket validation requirements for high-risk applications, (3) mandate ongoing post-market surveillance and performance monitoring, (4) require transparency about algorithmic limitations and validation study details, and (5) enforce compliance through meaningful penalties for violations (Benjamens et al., 2020).

Regulatory standards should be dynamic, evolving as AI technology advances and evidence about real-world performance accumulates. This requires agencies to invest in technical expertise, maintain active post-market surveillance programs, and update requirements as needed rather than treating initial approvals as permanent authorizations.

Tiered Liability Model Based on AI Autonomy

We propose a tiered liability framework that adjusts responsibility allocation based on AI's role in clinical decision-making, recognizing three distinct categories:

Tier 1 - Consultative AI: Systems that provide optional second opinions or supplementary information without replacing human judgment. Examples include AI that identifies potentially suspicious findings for radiologist review or suggests additional differential diagnoses for clinician consideration. For Tier 1 systems, physician-centered liability predominates. Clinicians retain clear decisional autonomy and should be primarily responsible for diagnostic outcomes, though manufacturers remain liable for product defects.

Tier 2 - Collaborative AI: Systems that function as active partners in diagnosis, with physicians and algorithms jointly contributing to clinical decisions. Most contemporary AI diagnostic applications fall into this category. For Tier 2 systems, distributed liability applies, with responsibility allocated based on contextual factors discussed in Section VI. Physicians bear responsibility for reasonable clinical judgment, manufacturers for adequate design and validation, and institutions for proper implementation and oversight. The specific distribution in individual cases depends on each party's causal contribution to harm.

Tier 3 - Autonomous AI: Systems that make diagnostic or triage decisions without routine human review, such as AI that automatically clears normal imaging studies or provides diagnoses in settings lacking specialist availability. For Tier 3 systems, manufacturer and institutional liability predominates. Developers bear primary responsibility for ensuring algorithms can safely function autonomously, while institutions that deploy such systems must rigorously validate their appropriateness and monitor performance. Physician liability becomes minimal or absent for decisions made entirely by autonomous systems, though physicians might retain responsibility for appropriate system selection and scope definition.

This tiered approach provides needed flexibility while maintaining clear responsibility structures. It

acknowledges that different AI applications raise distinct liability considerations while avoiding the unworkable complexity of fully individualized liability determination for every AI system.

Complementary Mechanisms: Insurance and Compensation Funds

Beyond tort liability, we recommend two complementary mechanisms to ensure patient compensation and distribute risk appropriately.

First, mandatory liability insurance requirements for both AI manufacturers and healthcare institutions deploying AI systems would ensure that responsible parties maintain financial resources to compensate injured patients. Insurance requirements should be risk-calibrated, with premiums reflecting AI systems' risk profiles based on validation quality, deployment scope, and historical performance. This creates market-based incentives for safety while ensuring remedy availability (Sage, 2004).

Second, establishing a supplementary no-fault compensation fund for AI-related injuries would provide additional patient protection for cases where liability is unclear or responsible parties lack adequate resources. This fund, financed through contributions from AI manufacturers, healthcare institutions, and potentially general healthcare budgets, would compensate patients who suffer significant harm during AI-assisted care regardless of fault attribution. The fund would operate alongside rather than replace tort liability, offering an administrative pathway for cases where litigation proves impractical (Sage, 2004).

These mechanisms balance the compensation function of liability (ensuring patient remedies) with its deterrence function (incentivizing safety), while acknowledging that neither function is perfectly served by tort law alone.

Implementation Pathway and Adaptive Governance

Implementing this framework requires coordinated action across multiple domains. Legislatures should enact statutes clarifying AI-specific liability standards, moving beyond common law evolution that proceeds too slowly for rapidly advancing technology. Regulatory agencies must develop detailed technical requirements for AI validation, transparency, and monitoring. Professional societies should establish practice guidelines specifying appropriate AI use. Healthcare institutions need to invest in governance infrastructure for AI oversight. And courts must develop expertise in adjudicating AI-related disputes, potentially through specialized dockets or expert technical advisors.

Critically, this framework must remain adaptive. As AI technology evolves—with systems becoming more autonomous, more explainable, or more deeply integrated into clinical workflows—liability rules should evolve correspondingly. We recommend mandatory periodic review of AI liability standards, with sunset provisions requiring affirmative reauthorization rather than indefinite continuation of initial rules. This adaptive governance approach acknowledges the fundamental uncertainty surrounding AI's trajectory while establishing mechanisms for ongoing adjustment (Geis et al., 2019).

The framework we propose is admittedly complex, but this complexity reflects the genuine challenges of allocating responsibility for AI-assisted diagnosis. Simplistic approaches that assign liability exclusively to physicians or manufacturers fail to account for AI's distinctive characteristics and the distributed nature of modern healthcare delivery. By combining clear foundational principles with stakeholder-specific responsibilities, tiered liability based on AI autonomy, and complementary compensation mechanisms, this framework offers a path toward accountability that protects patients, incentivizes responsible innovation, and remains practically administrable.

VIII. Conclusion

The widespread adoption of AI-assisted diagnosis is fundamentally reshaping both the epistemic structure of medical decision-making and the allocation of responsibility within clinical practice. By examining key technical characteristics—algorithmic opacity, continuous learning, and probabilistic outputs—this article argues that diagnostic AI systems are no longer merely passive instruments in the hands of clinicians. Rather, they function as “cognitive collaborators” that actively participate in clinical reasoning. It is precisely this transformation in role that places systemic strain on existing models of medical responsibility, which were constructed around human physicians as the sole epistemic and moral agents.

Through a comparative analysis of physician-centered liability, manufacturer responsibility, distributed responsibility, and no-fault compensation models, this article demonstrates that no single liability paradigm is sufficient to capture the complex socio-technical realities of AI-assisted

diagnosis. Assigning responsibility exclusively to clinicians overlooks the epistemic opacity of algorithms and the institutional pressures that shape their use. Conversely, treating AI systems as mere products and shifting responsibility entirely to developers underestimates the irreplaceable role of clinical judgment in diagnostic practice. While distributed responsibility more closely reflects actual causal structures, it risks diluting accountability and impeding effective patient redress in practice. No-fault compensation schemes, although strengthening patient protection, weaken the normative function of responsibility attribution as an expression of moral accountability.

The central conclusion of this article is that responsibility in AI-assisted diagnosis should be governed by a context-sensitive, factor-oriented normative framework, rather than by presumptive attribution to any single actor. Responsibility allocation ought to take into account a range of critical variables, including the degree of algorithmic explainability, the extent of AI's practical influence within clinical workflows, the system's validation and regulatory status, the maturity of professional guidelines, the nature of the error involved (design defects, data bias, or misuse), and whether patients were adequately informed of AI involvement. Where algorithms are highly opaque and clinicians lack meaningful capacity to evaluate their outputs, greater normative responsibility should rest with developers and deploying institutions. By contrast, where AI functions as an advisory tool and physicians retain substantial discretionary authority, traditional principles of professional responsibility should continue to play a dominant role.

Disclosure Statement

No potential conflict of interest was reported by the author.

Data Availability Statement

Data available on request from the author.

References

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 1-9.

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. NPJ Digital Medicine, 3(1), 1-8.

Berlin, L. (2017). Medical errors, malpractice, and defensive medicine: An ill-fated triad. Diagnosis, 4(3), 133-139.

Castelvecchi, D. (2016). Can we open the black box of AI? Nature, 538(7623), 20-23.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3), 231-237.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983.

Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., ... & Zou, J. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances, 8(32), eabq6147.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., ... & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. New England Journal of Medicine, 385(3), 283-286.

Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., ... & Kohli, M. (2019). Ethics of artificial intelligence in radiology: Summary of the joint European and North American multisociety statement. Radiology, 293(2), 436-440.

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial Intelligence in Healthcare, 295-336.

Kesselheim, A. S., Woloshin, S., Lu, Z., Tessema, F. A., Ross, K. M., & Schwartz, L. M. (2019). Physicians' knowledge about FDA approval standards and perceptions of the "breakthrough therapy" designation. JAMA, 315(14), 1516-1518.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Report, 49(1), 15-21.

MarketsandMarkets. (2022). Artificial Intelligence in Healthcare Market - Global Forecast to 2030. MarketsandMarkets Research Private Ltd.

Mello, M. M., & Studdert, D. M. (2008). Deconstructing negligence: The role of individual and system factors in causing medical injuries. Georgetown Law Journal, 96, 599-623.

Morreim, E. H. (2015). The clinical standard of care: Let's toss the recipe book. Journal of Health Care Law &

Policy, 18, 95-134.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175-183.
Niazi, M. K. K., Parwani, A. V., & Gurcan, M. N. (2019). Digital pathology and artificial intelligence. The Lancet Oncology, 20(5), e253-e261.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
Price, W. N. (2017). Regulating black-box medicine. Michigan Law Review, 116(3), 421-474.
Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. Health and Technology, 7(4), 351-367.
Ross, C., & Swetlitz, I. (2017). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. STAT News, July 25, 2017.
Solaiman, S. M. (2017). Legal personality of robots, corporations, idols and chimpanzees: A quest for legitimacy. Artificial Intelligence and Law, 25(2), 155-179.
Sage, W. M. (2004). The forgoten third: Medical malpractice insurance and its role in the tort system. Health Affairs, 23(4), 10-21.
Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.