International Journal of Healthcare Sciences ISSN 2348-5728 (Online)
Vol. 14, Issue 1, pp: (549-570), Month: April 2026 - September 2026, Available at: www.researchpublish.com

Protecting Human Judgment in AI-Enabled Healthcare: A Policy Framework for Responsible Artificial Intelligence Use

David Bull

Ed.D., PhD, DBA, MBA, MSc, BCMHC, PMP.
American InterContinental University System.

DOI: https://doi.org/10.5281/zenodo.22092819

Published Date: 25-August-2026

Abstract: **Artificial intelligence (AI) is increasingly integrated into healthcare decision-making, offering significant opportunities to improve diagnostic support, operational efficiency, risk prediction, access, and quality of care. Existing AI governance frameworks appropriately emphasize safety, effectiveness, transparency, equity, privacy, accountability, and human oversight. However, comparatively less policy attention has been directed toward whether healthcare professionals retain the cognitive, professional, and institutional capacity to exercise meaningful independent judgment as AI assumes greater influence over consequential decisions. This white paper identifies this emerging governance gap and proposes the Human Judgment Protection Standard (HJPS), a risk-based framework for protecting meaningful human judgment in AI-enabled healthcare. The HJPS comprises six interrelated domains: Human Decision Authority, Cognitive Independence, AI Competency and Critical Evaluation, Meaningful Human Oversight, Accountability and Traceability, and Continuous Human-Judgment Monitoring. A four-level risk classification model is proposed to align the intensity of human-judgment safeguards with the potential consequences and degree of AI influence associated with specific healthcare use cases. The framework further introduces a Human-Judgment Impact Assessment (HJIA) and a continuous organizational implementation cycle incorporating AI inventory, risk classification, workflow design, competency development, controlled deployment, dual-domain monitoring, audit, corrective action, and reassessment. The paper argues that maintaining a human formally "in the loop" is insufficient when automation bias, cognitive offloading, algorithmic authority, professional deskilling, workload, or organizational conditions undermine the individual's practical capacity to evaluate or challenge AI-generated recommendations. The HJPS therefore advances a transition from human-in-the-loop to human-capacity-in-the-loop governance. Policy recommendations emphasize risk-proportionate oversight, professional AI competency, meaningful override authority, shared accountability, organizational monitoring, accreditation integration, and continued empirical evaluation. The framework offers a human-centered approach for ensuring that advances in healthcare AI strengthen rather than inadvertently diminish the human capabilities necessary for safe, ethical, accountable, and resilient healthcare.**

Keywords: **artificial intelligence, healthcare AI, human judgment, human-centered AI, Human Judgment Protection Standard, meaningful human oversight, cognitive offloading, automation bias, AI governance, healthcare policy.**

I. INTRODUCTION

Artificial intelligence (AI) is rapidly becoming embedded across the healthcare continuum, influencing clinical decision support, diagnostic interpretation, risk prediction, documentation, patient monitoring, resource allocation, administrative operations, and population health management. These applications offer substantial opportunities to improve efficiency, expand access to expertise, reduce administrative burden, and strengthen the quality and timeliness of healthcare decisions. At the same time, the increasing sophistication and autonomy of AI systems are changing the relationship between technology and the professionals expected to interpret and act upon their outputs. Consequently, responsible AI governance must address not only the safety and performance of AI technologies but also the conditions under which humans interact with them. The World Health Organization (WHO, 2021) has emphasized that AI for health should preserve human autonomy, promote well-being and safety, ensure transparency, foster responsibility, and advance equity. Similarly, the

National Institute of Standards and Technology (NIST, 2023) frames trustworthy AI as requiring continuous governance and risk management throughout the technology lifecycle.

U.S. healthcare AI governance is also evolving through regulatory, professional, and organizational mechanisms. The U.S. Food and Drug Administration (FDA) has developed guidance addressing AI-enabled medical devices and clinical decision-support software, while federal health-information policy increasingly emphasizes transparency surrounding predictive algorithms incorporated into certified health information technology. Professional organizations have likewise emphasized that AI should augment rather than replace professional judgment. Collectively, these developments represent important progress toward safer and more accountable healthcare AI. Nevertheless, much of the existing governance discussion concentrates primarily on the technology, its accuracy, reliability, transparency, bias, privacy, security, and validation. Comparatively less attention has been directed toward a parallel question: What happens to human decision-making capacity as professionals increasingly rely upon AI to perform cognitive functions that previously required independent professional reasoning?

This question is particularly important because keeping a human formally involved in an AI-supported process does not necessarily guarantee meaningful human oversight. Automation bias may predispose professionals to accept algorithmic recommendations even when contradictory information is available, while cognitive offloading can transfer memory, interpretation, and reasoning activities from the individual to technological systems. Over time, repeated substitution of AI for professional cognitive activity may also create concerns about dependency, diminished vigilance, or deterioration of selected skills. These effects should not be interpreted as arguments against AI. Cognitive offloading and automation can be beneficial when they reduce unnecessary cognitive burden and allow professionals to concentrate on higher-value activities. The governance challenge is instead to distinguish beneficial cognitive augmentation from forms of technological dependence that weaken capabilities still necessary for safe and accountable practice.

This distinction is consistent with a human-centered approach to artificial intelligence. Human-centered AI seeks to develop and deploy technology in ways that strengthen human agency, capability, responsibility, and well-being rather than treating technological performance as an end in itself (Shneiderman, 2022). Bull's (2026a, 2026b) work on human-centered AI and human presence similarly emphasizes intentional human agency within AI-mediated environments, while the Conscious Intelligence Integration Framework extends this concern to reflective awareness, discernment, metacognitive regulation, ethical reasoning, and epistemic responsibility (Bull, 2026c). Although these concepts have broader applications, they are particularly consequential in healthcare, where professionals may remain ethically, legally, and organizationally accountable for decisions increasingly shaped by intelligent technologies.

The resulting policy gap concerns the difference between human presence and human capacity. A healthcare professional may remain technically "in the loop" while lacking sufficient information to evaluate an AI recommendation, adequate time to consider alternatives, competency to recognize system limitations, authority to override the technology, or sufficient independent practice to maintain the skills necessary for intervention. Recent scholarship has accordingly questioned whether conventional human-in-the-loop arrangements provide meaningful oversight when humans lack the epistemic and practical capacity to intervene effectively (van de Sande et al., 2026). This concern becomes more consequential as AI systems become increasingly accurate and embedded in routine workflows. The greater the perceived reliability of a system, the easier it may become for verification to evolve into habitual acceptance rather than substantive professional review.

The central policy question is no longer simply whether healthcare AI is accurate, but whether humans remain cognitively, professionally, and institutionally capable of exercising meaningful judgment when AI participates in consequential decisions. This white paper addresses that question by proposing the Human Judgment Protection Standard (HJPS), a risk-based framework for preserving meaningful human judgment within AI-enabled healthcare. The HJPS does not seek to maximize human intervention or constrain beneficial automation. Rather, it establishes a framework for determining when human decisional capabilities require protection, what those capabilities entail, and how healthcare organizations can monitor them alongside technological performance. The framework incorporates six domains of human-protected judgment, a four-level risk classification model, a Human-Judgment Impact Assessment (HJIA), and an organizational implementation cycle designed to integrate human-capability considerations into existing AI governance.

The paper proceeds from the premise that the appropriate unit of analysis for responsible healthcare AI is increasingly the human–AI system, rather than either the algorithm or professional considered independently. It first examines the emerging healthcare AI governance landscape and identifies limitations of conventional human-in-the-loop approaches. It then considers automation bias, cognitive offloading, algorithmic authority, dependency, and potential professional deskilling as

interconnected threats to meaningful judgment. Building upon human-centered AI and Conscious Intelligence principles, the paper introduces the HJPS and its risk-proportionate governance structure. Subsequent sections address organizational implementation, governance and accountability, ethical and legal considerations, policy recommendations, and implementation trade-offs. The ultimate objective is not to protect humans from artificial intelligence, but to ensure that increasingly capable AI strengthens healthcare without inadvertently weakening the human judgment upon which safe, ethical, accountable, and resilient care continues to depend.

Healthcare AI Governance and the Unresolved Policy Gap

The governance of artificial intelligence in healthcare is developing rapidly as AI moves from relatively discrete analytical applications toward technologies embedded directly within clinical and organizational workflows. Current governance approaches appropriately emphasize safety, effectiveness, transparency, fairness, privacy, cybersecurity, validation, and accountability (National Institute of Standards and Technology [NIST], 2023; World Health Organization [WHO], 2021). These safeguards are essential because AI-supported decisions may affect diagnosis, treatment, prognosis, triage, resource allocation, and access to healthcare. Yet the expanding role of AI creates a second governance challenge that is less directly addressed: how healthcare organizations should protect the human capacity for meaningful judgment as increasingly capable technologies assume portions of professional cognitive work. The distinction is consequential because safe technology does not automatically produce a safe human–AI decision system; meaningful oversight also depends on the human decision-maker's epistemic capacity, cognitive space, decisional authority, and practical ability to intervene (van de Sande et al., 2026).

The Emerging Governance Environment

International and U.S. policy frameworks provide an important foundation for responsible healthcare AI. The World Health Organization (WHO, 2021) identifies protection of human autonomy, promotion of human well-being and safety, transparency, responsibility, inclusiveness, and sustainability as central principles for AI in health. Similarly, the NIST Artificial Intelligence Risk Management Framework emphasizes governance across the AI lifecycle and organizes risk management around the functions of Govern, Map, Measure, and Manage (NIST, 2023). Together, these approaches establish that responsible AI requires more than technical accuracy; it requires ongoing institutional governance.

In the United States, healthcare AI oversight is distributed across regulatory and professional mechanisms rather than governed by a single comprehensive healthcare AI statute. The FDA regulates AI-enabled technologies that meet applicable medical-device requirements while recognizing that certain clinical decision-support (CDS) software functions may fall outside the statutory definition of a device. Its January 2026 final guidance clarifies the criteria used to distinguish certain non-device CDS functions from software functions subject to device regulation. FDA's broader digital-health portfolio also includes final guidance addressing predetermined change-control plans for AI-enabled device software and draft lifecycle-management guidance for AI-enabled device software functions, reflecting increasing attention to AI across the product lifecycle.

Federal health-information policy provides another layer of governance. The Office of the National Coordinator for Health Information Technology's HTI-1 final rule introduced algorithm-transparency requirements for AI and other predictive algorithms incorporated into certified health IT. The requirements are intended to provide clinical users with baseline information that can help them evaluate predictive interventions for fairness, appropriateness, validity, effectiveness, and safety. Importantly, HTI-1 addresses characteristics such as intended use, intended users and populations, known limitations, and the decision-making role for which a predictive intervention was designed. These developments represent significant movement toward transparency and responsible deployment.

From Technology Governance to Human–AI Governance

Although healthcare AI governance has advanced substantially in addressing technological safety, validation, transparency, fairness, privacy, security, and accountability, an important gap remains. Most existing governance mechanisms understandably begin with the AI system itself: Is it accurate? Is it appropriately validated? Is it biased? Is it secure? Are its intended uses and limitations transparent? Is its performance monitored over time? These questions are essential, but they do not fully establish whether the human–AI decision system as a whole is safe and capable of supporting responsible decisions.

As AI assumes a greater role in professional reasoning and decision-making, governance must therefore extend beyond the performance of the technology to consider the capabilities and conditions of the humans expected to oversee it. The relevant

questions become not only whether the AI performs safely, but whether healthcare professionals retain the knowledge, cognitive independence, professional competence, authority, and practical ability to recognize errors, question recommendations, and intervene when necessary. This shift—from governing the technology alone to governing the interaction between human and artificial intelligence—provides the foundation for addressing the human-capability gap in healthcare AI governance.

Consider a highly accurate clinical decision-support system operating as designed. The technology may satisfy technical and regulatory expectations while clinicians gradually become less likely or less able to independently evaluate its recommendations. Similarly, professionals may formally possess override authority but face insufficient information, time pressure, workflow constraints, or organizational expectations that make disagreement difficult in practice. Under such conditions, the AI may remain technically reliable while the quality and effectiveness of human oversight deteriorate.

This possibility exposes the limitations of treating human-in-the-loop as a sufficient governance endpoint. Human involvement can range from substantive independent evaluation to little more than procedural confirmation. A clinician who reviews an AI recommendation against relevant clinical evidence, understands the system's limitations, considers reasonable alternatives, and can freely question or override the recommendation exercises fundamentally different oversight from one who routinely approves AI recommendations because the system is presumed correct. Both arrangements technically place a human in the loop, but only the former clearly demonstrates meaningful human oversight.

The policy concern is therefore not simply whether a human remains involved, but whether that individual retains the capacity and conditions necessary to exercise meaningful judgment. These include adequate knowledge and professional competency, cognitive independence, sufficient information and time, appropriate decision authority, practical ability to intervene, and organizational support for questioning AI-generated recommendations. When these conditions are absent, human oversight may remain formally present while becoming functionally weak or merely symbolic.

The Human-Capability Gap

The human-capability gap in healthcare AI governance is the difference between formally requiring healthcare professionals to oversee AI-supported decisions and ensuring that they retain the cognitive capacity, professional competence, knowledge, information, authority, and practical ability necessary to exercise that oversight meaningfully.

The gap may emerge gradually as AI assumes a greater share of cognitive work. AI can reduce cognitive workload by performing calculations, synthesizing information, identifying abnormalities, generating documentation, prioritizing cases, and recommending actions. Such cognitive offloading can be beneficial and should not, by itself, be regarded as a threat to professional competence. However, when AI repeatedly substitutes for reasoning capabilities that professionals must still retain to recognize errors, manage exceptional circumstances, respond to system failures, or assume responsibility for consequential decisions, efficiency gains may be accompanied by diminished human capability and reduced professional resilience.

The human-capability gap becomes clearer when conventional AI governance questions are considered alongside questions about the continuing capacity of the professionals expected to provide oversight:

Table 1

Conventional AI governance question	Human-capability question
1. Is the AI accurate?	Can the professional recognize when the AI may be inaccurate or inappropriate?
2. Can the AI recommendation be overridden?	Does the professional have the competence, authority, and practical ability to override or escalate it when necessary?
3. Is a human in the loop?	Does the human retain the capacity to exercise meaningful independent judgment?
4. Is AI performance monitored?	Are the effects of AI use on human judgment, competency, and performance also monitored?
5. Is the algorithm sufficiently transparent?	Does the professional have sufficient understanding to interpret and use that transparency meaningfully?
6. Who approved the final decision?	Who possessed meaningful knowledge, authority, and control over the conditions that produced the decision?

These distinctions illustrate why formal human involvement cannot, by itself, establish meaningful human oversight. An AI system may be accurate, transparent, monitored, and technically subject to human override while the professional responsible for oversight has insufficient capability, information, time, independence, or authority to recognize when intervention is necessary. The policy concern is therefore not simply whether safeguards exist, but whether the humans upon whom those safeguards depend remain capable of using them effectively.

The final distinction has particular implications for accountability. A professional may provide final authorization while the healthcare organization determines which technology is purchased, how it is configured, how much time is available for review, what training is provided, how staffing is structured, and whether challenging the AI is practically supported. Developers, meanwhile, control important aspects of system design, performance information, updates, and technical limitations. Assigning responsibility solely to the person providing final approval may therefore create an accountability–control mismatch, in which formal responsibility exceeds the individual's actual knowledge, authority, or control over the conditions that produced the decision.

***Why Transparency and Explainability Are Not Enough*.** Increasing transparency is an important response to AI risk, but transparency alone cannot preserve human judgment. HTI-1's requirements illustrate the value of providing users with information about intended use, populations, limitations, and predictive decision-support functions. Yet information becomes protective only when professionals have sufficient competency, time, cognitive independence, and authority to act upon it.

Similarly, explainability does not necessarily guarantee meaningful oversight. An understandable AI recommendation may still be accepted uncritically. Conversely, some highly complex models may provide clinically useful outputs even when their internal computations cannot be completely interpreted by individual users. The relevant governance objective is therefore not absolute technological explainability but actionable understanding: professionals should have sufficient information about an AI system's intended role, limitations, uncertainty, and relevant performance characteristics to determine when reliance is reasonable and when additional evaluation is warranted.

From Human-in-the-Loop to Human-Capacity-in-the-Loop. The unresolved policy gap suggests the need for a conceptual shift. Human-in-the-loop governance asks whether a person remains involved in a technologically supported decision. Human-capacity-in-the-loop governance asks whether that person remains capable of performing the function that justifies keeping a human there.

This distinction does not imply that humans must independently reproduce every AI-supported analysis. Such duplication would undermine many of the benefits of automation. Nor does it assume that human judgment is always superior. AI may outperform professionals on specific tasks, and appropriate reliance on validated technology can improve healthcare. The objective is instead to identify the human capabilities that remain necessary when AI encounters uncertainty, unusual circumstances, ethical complexity, conflicting evidence, system failure, or conditions beyond its validated use.

The resulting policy proposition is straightforward: the intensity of human-judgment protection should increase as the consequences of a decision, the influence of AI over that decision, and the potential consequences of diminished human capability increase. Existing technology-centered safeguards should therefore be complemented by governance mechanisms that assess professional capacity, decision authority, cognitive independence, competency, and changes in human performance over time.

This is the specific gap addressed by the Human Judgment Protection Standard (HJPS). Rather than replacing FDA regulation, federal health-information requirements, NIST risk management, professional standards, or organizational quality systems, the HJPS is intended to complement them by addressing the human side of AI-enabled decision-making. Its central premise is that responsible healthcare AI governance must protect not only the reliability of intelligent technologies, but also the human capabilities necessary to govern their use responsibly.

Human-Centered AI and the Transition to Human-Protected Judgment

The risks associated with automation bias, cognitive offloading, algorithmic authority, and professional deskilling point toward a broader challenge in healthcare AI governance: technological performance cannot be separated from the human system in which the technology operates. Artificial intelligence may be highly accurate and technically reliable while still being implemented in ways that diminish professional agency, critical evaluation, or meaningful oversight. Addressing this challenge requires moving beyond technology-centered governance toward a human-centered model in which AI performance and human capability are considered jointly. Human-centered artificial intelligence provides an important conceptual foundation for this transition because it positions technology as a means of augmenting human capability while preserving human agency, responsibility, and control (Shneiderman, 2022).

Human-Centered AI as a Governance Foundation

Human-centered AI challenges the assumption that technological advancement should be measured primarily by increasing automation. Shneiderman (2022) argues that high levels of automation and high levels of human control need not be mutually exclusive. Properly designed systems can combine powerful computational capabilities with meaningful human agency. This perspective is especially relevant to healthcare, where AI may identify patterns, generate predictions, synthesize evidence, and support decisions while professionals contribute contextual understanding, ethical reasoning, patient knowledge, communication, and responsibility for care.

The World Health Organization (2021) similarly places human autonomy at the center of AI governance for health, emphasizing that humans should remain in control of healthcare systems and medical decisions. However, preserving formal decision authority does not necessarily ensure that professionals remain capable of exercising it. As AI becomes more deeply embedded in workflow, a human-centered approach must therefore address not only who formally makes the decision, but whether the individual continues to possess the cognitive and professional resources necessary to make that decision meaningfully.

This distinction extends human-centered AI from a design philosophy toward a human-capability governance principle. An AI system should not be considered human-centered merely because it is usable, interpretable, or subject to human approval. Human-centered governance must also consider how repeated interaction with the technology affects professional reasoning, autonomy, competence, and accountability over time.

Human Agency in AI-Mediated Environments

This concern is consistent with Bull's (2026a, 2026b) work on human-centered AI and teaching presence, which conceptualizes AI as an amplifier of human capability rather than a substitute for intentional human agency. Although developed within educational contexts, the underlying human-centered proposition has broader relevance: AI should strengthen rather than displace the uniquely human functions upon which responsible professional practice depends.

The transfer to healthcare should be made cautiously because teaching and healthcare decision-making involve different professional, ethical, and regulatory environments. The relevant conceptual connection is therefore not the direct application of teaching-presence theory to clinical care. Rather, it is the broader principle that AI-mediated systems remain dependent upon purposeful human presence, interpretation, contextualization, and responsibility. In healthcare, those functions become particularly consequential when decisions involve uncertainty, competing evidence, patient preferences, ethical considerations, or potentially irreversible consequences.

Human agency consequently involves more than retaining permission to override AI. It includes the capacity to recognize *when* override or further investigation is warranted. A clinician cannot meaningfully challenge a recommendation if the cognitive capabilities necessary to recognize the problem have substantially deteriorated. Likewise, institutional permission to disagree with AI provides little protection when organizational workflows, time pressures, or cultural expectations make disagreement practically difficult. Meaningful agency therefore requires both individual capability and enabling organizational conditions.

Conscious Intelligence and Critical AI Engagement

The Conscious Intelligence Integration Framework (CIIF) provides an additional conceptual foundation for understanding the human capacities required within AI-mediated environments. Bull (2026c) conceptualizes Conscious Intelligence through five interrelated dimensions: Reflective Awareness, Discernment, Metacognitive Regulation, Ethical Reasoning, and Epistemic Responsibility. Together, these dimensions describe capabilities through which individuals can consciously evaluate their engagement with AI rather than passively accepting technologically generated information.

These dimensions have direct conceptual relevance to the problem of human judgment in healthcare. Reflective Awareness enables professionals to recognize how AI may be influencing their own thinking. Discernment supports critical evaluation of AI-generated information against evidence, context, and alternative explanations. Metacognitive Regulation involves monitoring and adjusting one's reasoning when interacting with technological assistance. Ethical Reasoning becomes necessary when technically optimized recommendations intersect with patient values, fairness, autonomy, or competing obligations. Epistemic Responsibility requires professionals to consider the quality and limitations of information before accepting or acting upon it.

These capabilities are particularly important because AI can alter not only the information available to a professional but also the process through which conclusions are formed. A system that consistently recommends a diagnosis, prioritizes a patient, or proposes a treatment may gradually become an influential participant in the professional's reasoning architecture. Conscious engagement therefore provides a conceptual counterweight to passive reliance.

From Human-Centered AI to Human-Protected Judgment

Human-centered AI and Conscious Intelligence provide important foundations for responsible engagement with artificial intelligence, but healthcare governance requires a further step: translating these principles into mechanisms that actively preserve the human capacity for meaningful judgment as AI influence increases. Human-centered AI establishes the normative principle that technology should augment human capabilities while preserving human agency and well-being. Conscious Intelligence extends this foundation by identifying the reflective, evaluative, ethical, and epistemic capacities needed for deliberate and responsible engagement with AI. Human-Protected Judgment moves from these individual and normative foundations to a governance obligation: healthcare systems should create and maintain the conditions necessary for professionals to exercise meaningful judgment when AI influences consequential decisions. The Human Judgment Protection Standard (HJPS) operationalizes that obligation through specific governance domains, risk classification, assessment, monitoring, and accountability mechanisms.

Together, these concepts represent a progression from human-centered design, to conscious human engagement, to protection of human judgment, and ultimately to organizational implementation. Table 2 summarizes this progression and clarifies the distinct contribution of each concept.

Table 2. Conceptual Progression From Human-Centered AI to the Human Judgment Protection Standard

Concept	Primary question	Contribution to healthcare AI
Human-Centered AI	How should AI augment rather than unnecessarily displace humans?	Establishes human agency, well-being, and control as design priorities
Conscious Intelligence	How should humans consciously and critically engage with AI?	Identifies reflective, evaluative, ethical, and epistemic capacities for responsible AI engagement
Human-Protected Judgment	How should healthcare systems preserve the human capacity for meaningful judgment as AI influence increases?	Establishes preservation of necessary human capability as a governance responsibility
Human Judgment Protection Standard (HJPS)	How can human-protected judgment be operationalized?	Provides governance domains, risk levels, assessment, monitoring, and accountability mechanisms

This progression is important because individual vigilance alone cannot address a systems-level problem. Healthcare professionals may be encouraged to think critically and independently, but their ability to do so is shaped by training, workload, workflow design, access to information, decision authority, opportunities for independent practice, organizational culture, and the technological environment. Human judgment should therefore be understood not solely as an individual professional attribute, but as a system-dependent capability that healthcare organizations have a responsibility to support and protect. This systems-level responsibility provides the conceptual basis for the HJPS.

Evidence increasingly suggests that these concerns extend beyond theoretical possibilities. Automation bias has long been documented in clinical decision-support environments, where erroneous technological advice can influence professional decisions despite contradictory information (Goddard et al., 2012). More recent reviews of AI in healthcare have identified concerns involving automation bias, overreliance, changes in diagnostic reasoning, and potential professional deskilling as AI assumes greater cognitive responsibility (Al-Anezi, 2026; Heudel et al., 2026). These findings do not establish that AI use inevitably diminishes professional capability; rather, they demonstrate that the effects of sustained AI reliance on human reasoning and expertise constitute a legitimate patient-safety and governance concern requiring systematic attention.

The Human-Capacity-in-the-Loop Principle

The preceding distinction between formal human involvement and meaningful human capability gives rise to a central principle of the proposed framework: *A human should be considered meaningfully "in the loop" only when that person retains sufficient cognitive capacity, professional competence, relevant information, decisional authority, and practical ability to evaluate and intervene in the AI-supported process.*

The Human-Capacity-in-the-Loop Principle shifts the focus of governance from simply ensuring human presence to ensuring that the human remains capable of performing the oversight function assigned to them. The principle does not require healthcare professionals to reproduce every computational task performed by AI, nor does it assume that humans should retain final authority in every AI-supported application. Instead, healthcare organizations should identify which human capabilities remain necessary for the safe and responsible use of a particular AI application and protect those capabilities whenever human oversight is relied upon as a safeguard.

The conceptual development of the proposed framework can therefore be represented as: *Human-Centered AI → Conscious and Critical AI Engagement → Human-Protected Judgment → Human Judgment Protection Standard*

Each stage serves a distinct function. Human-Centered AI establishes the value orientation by emphasizing human agency, well-being, and appropriate augmentation. Conscious and Critical AI Engagement identifies the reflective, evaluative, ethical, and epistemic capacities needed for responsible interaction with AI. Human-Protected Judgment establishes preservation of necessary human decision-making capability as a governance objective. The Human Judgment Protection Standard (HJPS) translates that objective into an operational policy framework.

This conceptual progression also clarifies the distinctive contribution of the HJPS. Existing AI governance frameworks appropriately emphasize the safety, reliability, transparency, fairness, and accountability of AI technologies. The HJPS complements these objectives by addressing an additional question: Do the healthcare professionals expected to interpret, oversee, question, and intervene in AI-supported decisions retain the capabilities and organizational conditions necessary to fulfill those responsibilities meaningfully? Human-Protected Judgment therefore serves as the conceptual bridge between responsible AI governance and meaningful human oversight.

The Human Judgment Protection Standard (HJPS)

The preceding analysis establishes a central limitation in contemporary healthcare AI governance: requiring human involvement does not necessarily ensure that meaningful human judgment remains available. As artificial intelligence assumes increasingly sophisticated cognitive functions, healthcare organizations need a mechanism for determining whether professionals retain the capabilities, authority, and organizational conditions necessary to evaluate and intervene in AI-supported decisions. To address this gap, this white paper proposes the Human Judgment Protection Standard (HJPS).

The Human Judgment Protection Standard (HJPS) is defined in this white paper as a risk-based governance framework designed to preserve the cognitive capacity, professional competence, institutional authority, and accountability of human decision-makers when artificial intelligence materially influences consequential healthcare decisions. The framework builds upon human-centered AI, meaningful human oversight, risk-based AI governance, and conscious and critical engagement with AI (Bull, 2026c; National Institute of Standards and Technology [NIST], 2023; Shneiderman, 2022; van de Sande et al., 2026; World Health Organization [WHO], 2021).

The HJPS is not intended to replace existing AI safety, regulatory, accreditation, privacy, cybersecurity, or quality-improvement requirements. Rather, it adds a complementary dimension to healthcare AI governance: systematic protection of the human capacity upon which meaningful oversight depends. Its central premise is that organizations cannot rely upon human judgment as a safeguard while simultaneously allowing the capabilities and conditions required for that judgment to deteriorate.

Purpose and Scope of the HJPS

The HJPS has four primary purposes. First, it establishes meaningful human judgment as an explicit governance objective rather than assuming that human involvement automatically provides adequate oversight. Second, it identifies the human capabilities and organizational conditions necessary for effective oversight of consequential AI-supported decisions. Third, it provides a risk-proportionate structure through which organizations can determine when stronger protections are warranted. Fourth, it establishes a basis for monitoring whether human judgment remains effective after AI deployment.

The framework applies broadly to AI-enabled healthcare activities in which technological outputs may materially influence consequential decisions. These may include clinical decision support, diagnostic interpretation, medication management, patient deterioration prediction, triage, behavioral health, utilization management, resource allocation, population health, and other applications affecting health outcomes or access to care. The relevant unit of governance is the AI use case rather than the technology alone. The same AI system may pose different human-judgment risks depending on the decision it supports, the population affected, the consequences of error, the degree of automation, and the availability of independent professional review.

The HJPS therefore does not presume that every healthcare AI application requires extensive human intervention. Low-risk administrative applications may require little beyond ordinary organizational governance and basic AI competency. Stronger protections become appropriate as AI exerts greater influence over decisions involving substantial potential harm, limited reversibility, vulnerable populations, or functions for which deterioration of human capability could compromise safety.

Foundational Principles

Five principles guide the HJPS.

Proportionality. Human-judgment safeguards should correspond to the consequences of the decision, degree of AI influence, and importance of retaining independent human capability. The objective is not maximum oversight but appropriate oversight.

Capability before presence. A human should not be considered an effective safeguard merely because the individual appears at some point in the decision process. Meaningful oversight requires sufficient competence, information, cognitive independence, authority, and practical capacity to intervene.

Augmentation rather than unnecessary duplication. HJPS does not require professionals to reproduce every task performed by AI. Beneficial automation should be encouraged when it improves care. Human capabilities should be protected when they remain necessary for safety, accountability, contextual interpretation, ethical judgment, or system resilience.

Accountability aligned with control. Responsibility for AI-supported decisions should reflect the authority, information, control, and reasonable capacity of clinicians, healthcare organizations, developers, and other actors to prevent or mitigate harm. The final human approver should not automatically become the sole repository of accountability.

Continuous governance. Human-judgment protection cannot end at deployment. Just as AI performance may change through model drift, updates, changing populations, or workflow modification, human behavior and competency may also change through prolonged AI exposure. Both require continuing evaluation.

These principles align with broader human-centered approaches emphasizing human autonomy, agency, safety, responsibility, and appropriate human control (Shneiderman, 2022; WHO, 2021), while extending governance toward preservation of the capabilities necessary to exercise that control.

The Six Domains of Human-Protected Judgment

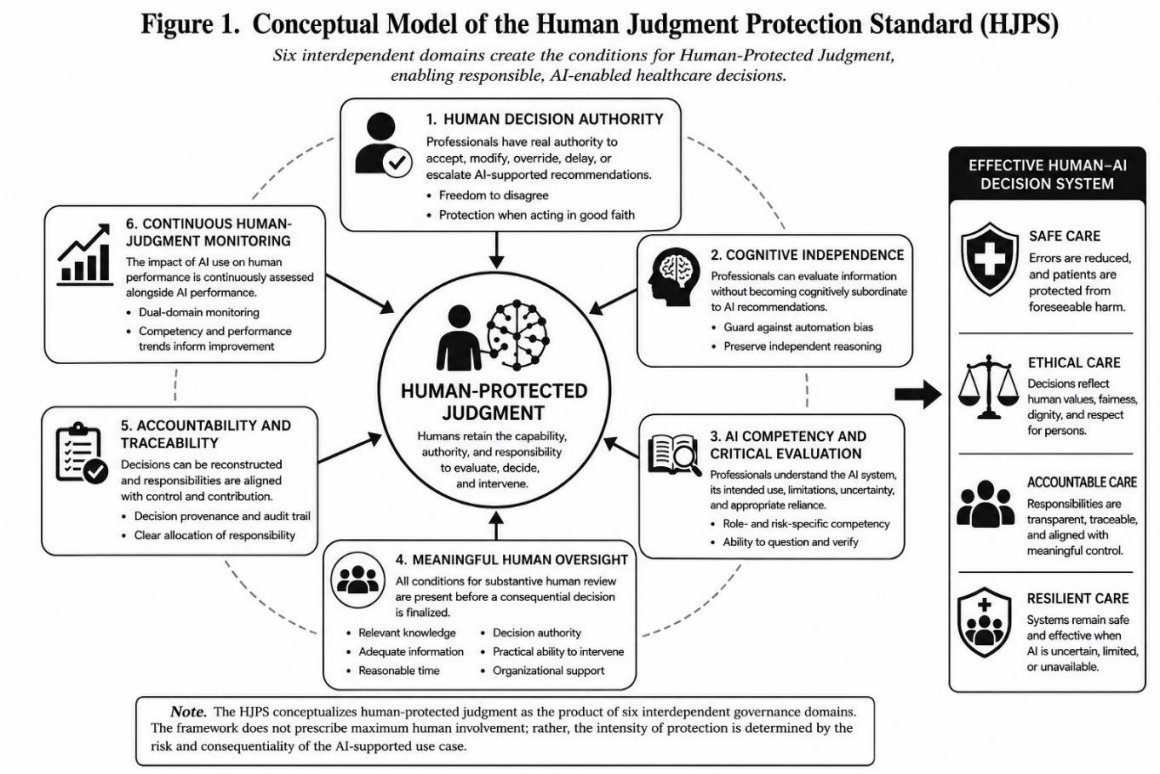
The HJPS operationalizes human-protected judgment through six interrelated domains. The domains should not be interpreted as isolated requirements; collectively, they describe the conditions under which human oversight can remain meaningful. See figure 1.

Domain 1: Human Decision Authority: Human Decision Authority concerns whether qualified professionals possess genuine authority to act when AI influences consequential decisions. Depending upon the application and professional scope, this may include the ability to accept, question, modify, override, delay, or escalate an AI-generated recommendation.

Formal authority alone is insufficient. An override mechanism that exists technically but is discouraged organizationally, difficult to access, or associated with punitive consequences does not constitute meaningful authority. Healthcare organizations should therefore ensure that professionals can exercise justified disagreement with AI without inappropriate barriers. Human authority should nevertheless remain proportional to the use case. The HJPS does not presume that every automated function requires manual authorization. Instead, explicit human authority becomes increasingly important as decisions become more consequential and difficult to reverse.

Domain 2: Cognitive Independence: Cognitive Independence refers to the professional's capacity to evaluate relevant information without becoming cognitively subordinate to the AI recommendation. It does not require ignoring AI or reaching a conclusion before viewing its output in every case. Rather, it requires preserving sufficient independent reasoning capability to recognize when an AI recommendation conflicts with clinical evidence, patient circumstances, ethical considerations, or other relevant information. This domain directly addresses automation bias, algorithmic authority, and excessive cognitive dependence. Organizations should consider whether workflow design inadvertently encourages automatic acceptance, for example, by presenting AI recommendations as default decisions, minimizing uncertainty, or requiring additional effort to select alternatives. Cognitive independence therefore represents both an individual capability and a design property of the human–AI system., See Figure 1.

Figure 1. Conceptual Model of the Human Judgment Protection Standard



Source. *Human Judgment Protection Standard (HJPS), developed by Bull (2026)*

The HJPS conceptualizes human-protected judgment as the product of six interdependent governance domains: Human Decision Authority, Cognitive Independence, AI Competency and Critical Evaluation, Meaningful Human Oversight, Accountability and Traceability, and Continuous Human-Judgment Monitoring. Collectively, these domains are intended to preserve the cognitive capacity, professional competence, institutional authority, and practical ability necessary for healthcare professionals to exercise meaningful judgment when AI influences consequential decisions. The framework is risk proportionate; the intensity of human-judgment protections should increase with the degree of AI influence, potential consequences of error, and importance of retaining independent human capability. The intended outcomes are safe, ethical, accountable, and resilient care

Domain 3: AI Competency and Critical Evaluation: Meaningful judgment requires more than technical proficiency. AI Competency and Critical Evaluation refers to the ability to use AI appropriately while understanding relevant limitations, uncertainty, potential biases, intended use, and circumstances requiring additional verification. Competency should be role- and risk-specific. A clinician using a high-risk diagnostic system requires different knowledge from an administrator using generative AI to summarize routine documents. For consequential applications, professionals should understand enough about the system to determine when reliance is appropriate and when independent evaluation or escalation is necessary.

This domain also incorporates reflective and metacognitive capacities associated with Conscious Intelligence, particularly awareness of how AI may influence one's reasoning, discernment regarding the quality of AI-generated information, and epistemic responsibility for information accepted as a basis for action (Bull, 2026).

Domain 4: Meaningful Human Oversight: Meaningful Human Oversight integrates the conditions necessary for human involvement to function as a substantive safeguard. It requires more than placing a professional between AI output and final action. Meaningful oversight exists when the responsible professional has:

1. relevant knowledge and competency;
2. access to sufficient information;
3. reasonable opportunity and cognitive space for evaluation;
4. appropriate decisional authority;
5. practical capacity to intervene; and
6. organizational support for questioning or escalating AI recommendations.

These conditions are interdependent. Competence without authority does not produce meaningful oversight; authority without information is similarly inadequate. Nor can a professional meaningfully review a recommendation when workload or workflow design reduces review to rapid procedural confirmation. The distinction can therefore be expressed simply:

Human presence + authority without capacity ≠ meaningful oversight.

Domain 5: Accountability and Traceability: Accountability and Traceability concern the ability to determine how consequential AI-supported decisions were produced and where relevant responsibilities reside. As decision-making becomes distributed across clinicians, healthcare organizations, algorithms, vendors, data systems, and workflows, conventional models of individual responsibility may become insufficient. For higher-risk applications, organizations should preserve enough information to reconstruct significant decision pathways when necessary. This may include the relevant AI recommendation, professional response, meaningful override or modification, escalation, material system information, and final action. Documentation should remain proportionate to risk so that traceability does not become excessive administrative burden.

Accountability should follow meaningful control. Healthcare professionals remain responsible for reasonable professional conduct within their authority, while organizations remain accountable for procurement, workflow design, training, staffing, monitoring, and implementation conditions. Developers and vendors retain responsibilities related to technological performance, relevant disclosures, and material system changes.

Domain 6: Continuous Human-Judgment Monitoring: The final domain, Continuous Human-Judgment Monitoring, addresses a major asymmetry in current AI governance. Organizations increasingly recognize the need to monitor AI accuracy, drift, bias, reliability, and safety after deployment. Comparable attention is rarely given to whether sustained AI use changes human performance.

The HJPS therefore proposes dual-domain monitoring: *AI Performance Monitoring + Human-Judgment Monitoring*

Human-focused indicators may include competency assessments, patterns of inappropriate reliance, error detection, meaningful overrides, escalation behavior, AI-related near misses, user-reported difficulty questioning AI, and selected performance during AI downtime. These measures should primarily support quality improvement and system learning rather than punitive surveillance. High agreement with AI should not automatically be interpreted as overreliance, nor should frequent override be considered evidence of effective judgment. Monitoring must consider context, system accuracy, case complexity, and outcomes. The relevant question is whether professionals demonstrate the ability to rely appropriately and disagree appropriately.

Table 3. The Six Domains of the Human Judgment Protection Standard

HJPS domain	Central governance question	Primary safeguard
1. Human Decision Authority	Can the professional meaningfully intervene?	Override, modification, delay, and escalation authority
2. Cognitive Independence	Can the professional evaluate the decision rather than merely accept AI output?	Independent reasoning and protection against automatic reliance

3. AI Competency and Critical Evaluation	Does the professional understand enough to use and question the AI appropriately?	Risk- and role-specific competency
4. Meaningful Human Oversight	Are the conditions for substantive human review actually present?	Knowledge, information, time, authority, and intervention capacity
5. Accountability and Traceability	Can the decision and distribution of responsibility be reconstructed?	Risk-proportionate documentation and decision provenance
6. Continuous Human-Judgment Monitoring	Does meaningful human capability remain intact over time?	Dual-domain monitoring and competency reassessment

The strength of the HJPS lies in the interaction among its domains. No single safeguard is sufficient. A clinician may have excellent AI competency but lack authority to challenge the system. Another may possess formal override authority but have become so dependent on AI that errors are rarely recognized. A third may identify an erroneous recommendation but work within a system where escalation is cumbersome or discouraged. In each circumstance, the existence of one safeguard is undermined by weakness elsewhere in the human–AI system.

The relationship can be represented conceptually as: *Human Decision Authority + Cognitive Independence + AI Competency and Critical Evaluation + Meaningful Human Oversight + Accountability and Traceability + Continuous Human-Judgment Monitoring → Human-Protected Judgment*

The six domains should not be interpreted as sequential stages or independent components. Rather, they function as an interdependent governance system. Weakness in one domain may compromise the effectiveness of the others. For example, a professional may possess AI competency and cognitive independence but still be unable to exercise meaningful judgment if organizational policies do not provide sufficient decision authority or practical ability to override an AI-supported recommendation. Accordingly, the HJPS evaluates human-protected judgment as a system-level capability rather than solely an individual professional attribute.

HJPS as a Complementary Standard

The HJPS is best understood as a complementary governance standard, not a replacement for existing frameworks. NIST's AI Risk Management Framework provides an overarching structure for managing AI risk (NIST, 2023); WHO establishes ethical principles for AI in health (WHO, 2021); FDA requirements and guidance address qualifying AI-enabled medical technologies; and federal health-information policy addresses areas such as algorithm transparency. The HJPS addresses a narrower question that cuts across these mechanisms:

If human judgment is being relied upon as part of the safety and accountability architecture, what must healthcare systems do to ensure that judgment remains meaningful?

This question becomes increasingly important as AI improves. Poorly performing AI naturally invites skepticism and verification. Highly reliable AI may create the opposite challenge: its success can make independent human evaluation appear progressively less necessary. The HJPS therefore treats preservation of human judgment not as opposition to technological advancement but as part of the infrastructure required to use increasingly capable AI responsibly. The framework also establishes the basis for the next stage of governance. Because the need for human protection varies substantially across applications, the six domains cannot be imposed with equal intensity in every setting. The HJPS therefore requires a risk-proportionate classification model that determines when basic safeguards are sufficient and when stronger requirements for independent review, competency, traceability, authorization, and monitoring become necessary.

Risk-Proportionate Human Judgment Protection

The HJPS recognizes that healthcare AI applications differ substantially in their potential consequences and degree of influence over human decision-making. Applying the same level of oversight to every AI use would create unnecessary burden, while insufficient oversight of consequential applications could compromise patient safety and meaningful professional judgment. The framework therefore uses a risk-proportionate approach in which human-judgment protections increase with the potential for harm, degree of AI influence, and importance of retaining independent human capability (NIST, 2023).

HJPS risk classification applies to the AI-supported use case rather than the technology alone. The same technology may present minimal risk when used for routine administrative assistance but substantially greater risk when used to influence diagnosis, treatment, triage, or access to care. Classification should therefore consider the severity and reversibility of potential harm, degree of AI influence, immediacy of the decision, patient vulnerability, availability of independent verification, and potential for cognitive substitution or professional deskilling.

Table 4. HJPS Risk Classification and Human-Judgment Protection Requirements

HJPS level	AI role	Human-judgment protection
Level 1: Minimal Risk	Low-consequence administrative assistance	Basic AI literacy, privacy/security safeguards, routine review.
Level 2: Moderate Risk	AI informs professional work or decision support	Role-specific competency, human verification, limitation awareness, scope monitoring.
Level 3: High Risk	AI materially influences consequential decisions	Qualified independent review, explicit override/escalation authority, competency assessment, traceability, human-performance monitoring.
Level 4: Critical Risk	AI substantially influences high-consequence or difficult-to-reverse decisions	Qualified human authorization where appropriate, enhanced monitoring and audit, strong traceability, contingency procedures, secondary review when warranted.

The classification is intended to be dynamic rather than permanent. Material changes in AI functionality, workflow, patient population, degree of automation, or evidence of increasing human reliance should trigger reassessment. Level 3 and Level 4 applications should receive particular attention through the proposed Human-Judgment Impact Assessment (HJIA) because these applications create the greatest potential mismatch between formal human responsibility and actual human capacity to intervene.

The purpose of risk classification is therefore not to maximize human involvement but to determine how much human-judgment protection is warranted for a particular AI-supported healthcare use case. This approach allows beneficial automation to proceed while concentrating stronger safeguards where diminished human capability could create significant safety, ethical, or accountability consequences.

Organizational Implementation of the HJPS and Human-Judgment Impact Assessment

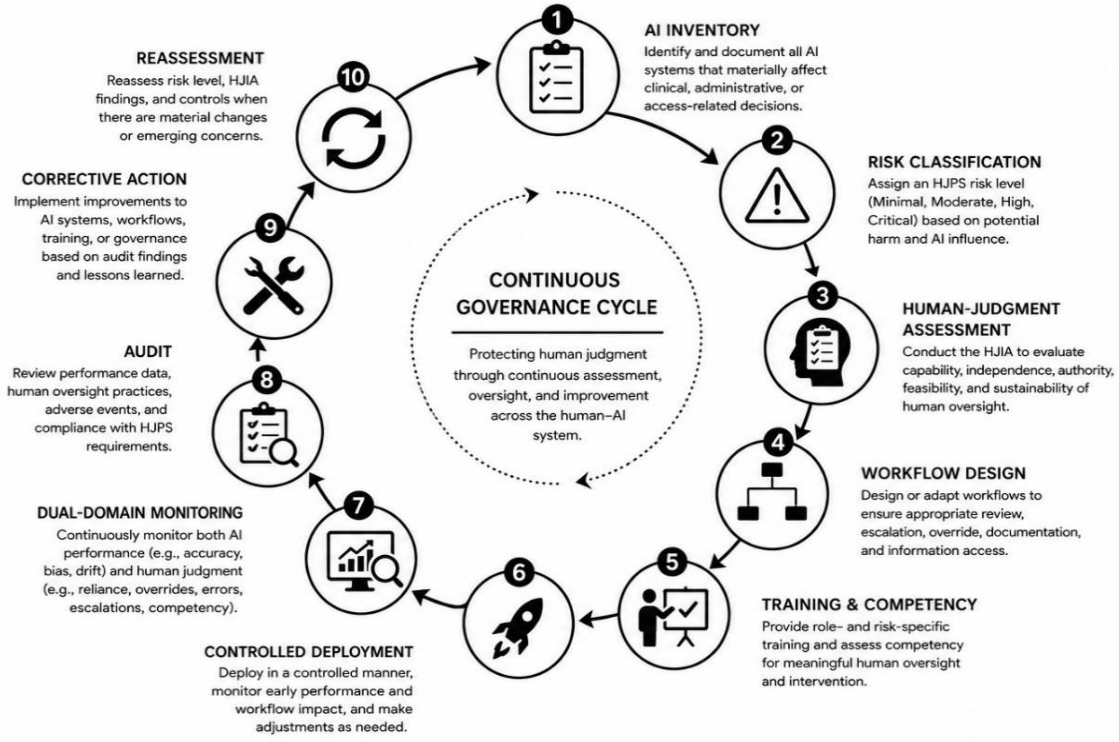
The HJPS is intended to function within existing healthcare governance, quality, patient-safety, compliance, and risk-management structures rather than as a separate administrative system. Implementation should be proportionate to the risk of the AI-supported use case and should address both technological performance and the human capacity required for meaningful oversight. For high- and critical-risk applications, organizations should determine before deployment whether professionals have the competence, information, authority, time, and practical ability necessary to evaluate and intervene in AI-supported decisions. A central implementation mechanism is the proposed Human-Judgment Impact Assessment (HJIA). The HJIA complements conventional assessments of algorithmic accuracy, bias, privacy, cybersecurity, and safety by examining how an AI application may affect the professionals expected to use or oversee it. For Level 3 and Level 4 applications, the assessment should address five core questions:

1. Capability: Do users possess the professional and AI-specific competencies necessary to critically evaluate the system's outputs?
2. Independence: Does the workflow preserve sufficient independent reasoning to recognize potentially inappropriate recommendations?
3. Authority: Can qualified professionals question, modify, override, delay, or escalate AI-supported decisions when warranted?
4. Feasibility: Do staffing, workload, information access, and workflow provide realistic conditions for meaningful oversight?
5. Sustainability: Could prolonged reliance on the AI reduce capabilities that professionals must retain for error detection, exceptional circumstances, or system failure?

The HJIA is not intended to establish that humans must independently reproduce the work performed by AI. Its purpose is to identify situations in which an organization claims human oversight as a safeguard while the conditions necessary for that oversight may be inadequate. Implementation should follow a continuous governance cycle: HJPS Organizational Implementation Cycle**.** See Figure 2.

Figure 2. HJPS Organizational Implementation Cycle

A continuous governance cycle for human-judgment–protected AI use in healthcare



Note. The HJPS Organizational Implementation Cycle represents a continuous process rather than a one-time sequence. Findings from monitoring, audit, and corrective action feed into reassessment and may result in reclassification, workflow modification, competency development, or changes in AI deployment. Source: Developed by the author.

The cycle begins with an AI inventory identifying systems that materially affect clinical, administrative, or access-related decisions. Each use case is then assigned an HJPS risk level. High- and critical-risk applications undergo an HJIA, after which workflows should be designed to provide appropriate review, escalation, override, and documentation mechanisms. Professionals should receive role- and risk-specific training before deployment, with competency assessed where the consequences of inappropriate reliance are substantial.

AI Inventory → Risk Classification → Human-Judgment Assessment → Workflow Design → Training & Competency → Controlled Deployment → Dual-Domain Monitoring → Audit → Corrective Action → Reassessment

Controlled deployment allows organizations to identify unintended workflow effects before broader implementation. Once deployed, AI-supported processes should undergo dual-domain monitoring, which evaluates both the technology and the humans interacting with it. AI monitoring may include accuracy, reliability, bias, drift, and safety events, while human-judgment monitoring may consider competency, inappropriate reliance, error detection, meaningful overrides, escalation patterns, and AI-related near misses. The distinction is important:

Healthcare organizations should ask not only, "Is the AI continuing to perform safely?" but also, "Are the humans relying on it continuing to exercise the judgment necessary for safe oversight?"

Audit findings should inform corrective actions such as workflow redesign, additional training, adjustment of decision authority, modification of AI use, or strengthened monitoring. Significant changes in the technology, workflow, population served, or observed patterns of reliance should trigger reassessment. In this way, human-judgment protection becomes a continuous quality-improvement function rather than a one-time implementation requirement.

Implementation Within Existing Governance

The HJPS should be integrated wherever possible into existing AI governance committees, quality-improvement programs, patient-safety systems, competency processes, procurement reviews, and enterprise risk-management structures. This approach minimizes unnecessary administrative burden while making human capability an explicit component of AI governance.

Implementation should also remain sensitive to organizational resources. Smaller, rural, safety-net, behavioral health, correctional, long-term-care, and other resource-constrained healthcare settings may lack dedicated AI governance infrastructure. A proportionate HJPS approach allows such organizations to concentrate resources on the AI applications presenting the greatest potential consequences rather than imposing identical requirements across all technologies.

Ultimately, implementation of the HJPS requires a shift from viewing AI governance primarily as technology surveillance toward governing the performance of the human–AI system. The purpose is not to impede automation but to ensure that healthcare organizations do not rely on human judgment as a safety mechanism without also protecting the conditions that make that judgment meaningful.

Governance, Ethics, and Accountability

The integration of AI into healthcare redistributes decision-making across professionals, healthcare organizations, technology developers, data systems, and automated processes. As this distribution becomes more complex, governance must ensure that responsibility does not become fragmented or obscured. The HJPS therefore adopts the principle that accountability should be distributed but not diffused: multiple actors may share responsibility for an AI-supported decision, but responsibility should remain identifiable and proportionate to each actor's authority, knowledge, control, and capacity to prevent harm. This approach is consistent with broader responsible-AI principles emphasizing human autonomy, accountability, transparency, safety, and fairness (NIST, 2023; WHO, 2021).

Accountability Should Follow Meaningful Control

Traditional models of professional accountability may become inadequate when clinicians retain formal responsibility for decisions while healthcare organizations and technology developers control important conditions under which those decisions occur. An organization may select the AI system, determine its configuration, establish workflows, allocate staffing, provide training, and define escalation procedures. Developers influence system design, performance, limitations, updates, and the information available to users. Healthcare professionals, meanwhile, interpret outputs and apply them to individual patients within the constraints of the environment established by these other actors.

The HJPS therefore proposes an accountability–control principle: *Responsibility for AI-supported healthcare decisions should correspond reasonably to each actor's authority, knowledge, control, foreseeable risk, and practical capacity to prevent or mitigate harm.*

Clinicians and other healthcare professionals remain accountable for appropriate professional conduct within their authority. However, the presence of a human final approver should not automatically transfer responsibility for algorithmic defects, inadequate implementation, insufficient training, unsafe staffing conditions, or system limitations that the professional could not reasonably identify or control. Similarly, healthcare organizations should not rely on a human-in-the-loop requirement as evidence of adequate governance when organizational conditions make meaningful intervention unrealistic.

Organizational and Shared Governance

Healthcare organizations bear substantial responsibility for creating the conditions under which meaningful human oversight occurs. Governance responsibilities include appropriate AI procurement, risk classification, workflow design, competency development, monitoring, escalation pathways, documentation, and periodic reassessment. High-risk AI should therefore be governed as an organizational patient-safety and quality issue, rather than treated solely as an information-technology function.

Developers and vendors have complementary responsibilities for accurately representing system capabilities, communicating known limitations, supporting appropriate monitoring, and disclosing material changes that could affect performance or use. Regulators, accreditors, and professional organizations contribute through external standards, competency expectations, and oversight mechanisms. The resulting governance model is one of shared but traceable responsibility.

Ethical Protection of Human Judgment

The HJPS also rests on an ethical proposition: preserving meaningful human judgment can be necessary for protecting patient autonomy, dignity, fairness, and individualized care. AI systems derive recommendations from data and computational processes, but healthcare decisions frequently involve considerations that cannot be reduced to prediction alone. Patient preferences, cultural and social circumstances, competing values, uncertainty, quality-of-life considerations, and ethical obligations may require contextual interpretation.

Meaningful human judgment is particularly important when AI-supported decisions affect vulnerable populations or access to care. Algorithmic systems may reproduce or amplify inequities present in historical data, making human capacity to recognize inappropriate or inequitable recommendations an important safeguard (WHO, 2021). However, requiring a human review provides limited protection if the reviewer lacks sufficient information, competency, independence, or authority to challenge the recommendation. Equity therefore depends not only on reducing algorithmic bias but also on ensuring that human oversight is substantive rather than procedural.

At the same time, human judgment should not be idealized. Human decision-makers are themselves susceptible to bias, inconsistency, fatigue, incomplete information, and error. The ethical objective of the HJPS is therefore not to privilege human judgment over AI by default. Rather, it is to combine human and artificial capabilities in ways that reduce the limitations of each. Where validated AI improves consistency or reduces human bias, its use may advance ethical healthcare. Where contextual, relational, or moral considerations exceed the appropriate scope of automation, meaningful human intervention becomes more important.

Transparency, Trust, and Professional Agency

Responsible human-AI decision-making also requires calibrated trust. Healthcare professionals should neither reject AI because it is artificial nor accept its recommendations because they appear computationally authoritative. Trust should correspond to evidence regarding system performance, intended use, limitations, uncertainty, and the circumstances of the individual case.

Transparency supports calibrated trust, but transparency is meaningful only when users can act upon the information provided. Consequently, HJPS links transparency to competency and authority: professionals must understand enough about an AI application's relevant characteristics to determine when additional verification is needed and must be institutionally empowered to act on that judgment.

Professional agency is therefore not preserved simply by allowing a clinician to click an override button. It requires an environment in which justified disagreement is possible, escalation is accessible, and professionals are not discouraged from challenging technology when evidence or patient circumstances warrant doing so.

Governance as a Human–AI System Responsibility

The broader implication is that AI governance should move beyond assigning responsibility after an adverse event toward designing conditions that make responsible decisions more likely before harm occurs. The HJPS accomplishes this by connecting human capability, organizational governance, ethical responsibility, and technological performance within the same framework.

Accordingly, the governance question should not be limited to *"Who made the final decision?"* It should also ask:

Who designed the decision environment? Who controlled the technology and workflow? Who possessed the information necessary to recognize risk? Who had the authority to intervene? And were the conditions for meaningful human judgment actually present?

These questions provide a more appropriate foundation for accountability in increasingly AI-enabled healthcare and establish the basis for considering the legal and regulatory implications of human-protected judgment.

Legal and Regulatory Considerations

The HJPS is proposed within a healthcare regulatory environment that already addresses important dimensions of AI safety, transparency, and accountability but does not yet provide a unified standard specifically focused on preservation of human decision-making capacity. In the United States, oversight is distributed across medical-device regulation, health-information

technology requirements, professional standards, state law, accreditation, organizational governance, and other regulatory mechanisms. The HJPS is therefore intended to complement rather than replace existing legal and regulatory requirements.

The U.S. Food and Drug Administration (FDA) regulates AI applications that qualify as medical devices under applicable federal law and has continued to refine its approach to AI-enabled device software. FDA guidance addresses, among other areas, clinical decision-support software and predetermined change-control plans for AI-enabled device software functions, while lifecycle guidance has considered how AI-enabled devices should be managed as they evolve over time. These approaches appropriately emphasize technological safety and lifecycle management. The HJPS adds a complementary question: when safety depends partly upon human review, what evidence demonstrates that the human reviewer remains capable of performing that function?

Federal health-information policy provides another relevant foundation. The HTI-1 final rule established algorithm-transparency requirements for predictive decision-support interventions incorporated into certified health information technology. These provisions make information available concerning factors such as intended use, intended users and populations, known limitations, and the decision-making role of predictive interventions, supporting assessment of fairness, appropriateness, validity, effectiveness, and safety. Such transparency is important to meaningful oversight, but the HJPS extends the principle by proposing that organizations also evaluate whether professionals possess the competency, cognitive independence, time, and authority necessary to use that information effectively.

The emerging accreditation and healthcare-governance environment further supports this direction. Joint Commission and the Coalition for Health AI released responsible-use guidance in 2025 emphasizing organizational AI governance, risk and bias assessment, education and training, safety-event reporting, privacy, security, and ongoing quality monitoring. The HJPS could complement such approaches by making preservation and monitoring of meaningful human judgment an explicit component of organizational AI governance.

Implications for Liability and Professional Responsibility

AI also complicates conventional approaches to professional liability. A clinician may formally authorize a decision while relying upon technology selected, configured, and implemented by the healthcare organization. The HJPS does not attempt to establish new legal liability rules. Instead, its accountability–control principle provides a policy basis for examining whether responsibility is reasonably aligned with meaningful authority and control.

This distinction may become increasingly important as AI assumes greater influence over clinical decisions. If healthcare organizations rely on human oversight as a safety mechanism, they should reasonably ensure that professionals receive appropriate training, relevant information, sufficient opportunity for review, and genuine authority to intervene. Conversely, professionals remain responsible for using AI within applicable standards of practice and exercising reasonable judgment within the authority and information available to them.

Toward Human-Capability Requirements

The longer-term regulatory implication of the HJPS is therefore not necessarily the creation of an entirely new regulatory regime. Human-judgment protections could instead be incorporated into existing mechanisms such as organizational AI governance, professional competency standards, accreditation expectations, procurement requirements, quality and patient-safety programs, and post-deployment monitoring.

For high- and critical-risk AI, policymakers and healthcare organizations should increasingly ask whether human oversight is merely required or demonstrably meaningful. Where human judgment is relied upon as part of the safety architecture, preservation of the capacity to exercise that judgment should itself become a measurable governance expectation. This would move healthcare AI policy from simply requiring a human in the loop toward ensuring human capacity in the loop.

Policy Recommendations and Action Agenda

The HJPS translates the preceding analysis into a focused policy agenda for healthcare organizations, regulators, accreditors, professional bodies, and technology developers. Its purpose is not to impose additional regulatory requirements on every AI application, but to ensure that when human judgment is relied upon as a safeguard, the capacity to exercise that judgment is deliberately protected. This approach is consistent with risk-based governance principles that emphasize differentiated human–AI roles, appropriate competency, oversight, and safeguards proportionate to the level of risk (NIST, 2023; WHO, 2021).

Building on these principles, the HJPS identifies 10 priority policy actions designed to translate human-judgment protection from a conceptual governance objective into practical organizational and policy requirements. Collectively, these recommendations address the conditions necessary to preserve meaningful human oversight across the AI lifecycle, from risk classification and pre-deployment assessment to competency development, monitoring, accountability, procurement, and continuing evaluation. The following recommendations provide an actionable pathway for integrating human-judgment protection into existing healthcare AI governance structures.

1. Establish Human-Judgment Protection as a Governance Principle. Healthcare AI policies should explicitly recognize preservation of meaningful human judgment as a component of responsible AI governance. Technical accuracy, transparency, fairness, and security remain essential, but organizations should also evaluate whether professionals retain the capacity necessary to interpret, challenge, and intervene in consequential AI-supported decisions. WHO's recent policy work similarly emphasizes that AI should augment rather than replace human judgment.

2. Adopt Risk-Proportionate Human-Judgment Requirements. Healthcare organizations should classify AI-supported use cases according to potential harm, degree of AI influence, reversibility, patient vulnerability, independent verifiability, and risk of cognitive substitution. Stronger safeguards should apply as risk increases. This approach avoids imposing high-intensity oversight on low-risk administrative AI while concentrating governance resources on consequential applications.

3. Require Human-Judgment Impact Assessments for High-Risk AI. Level 3 and Level 4 applications should undergo an HJIA before full deployment. The assessment should determine whether users possess sufficient competency, cognitive independence, information, authority, and practical opportunity to exercise meaningful oversight. It should also examine whether prolonged AI reliance could weaken capabilities still required for error detection, exceptional circumstances, or system failure.

4. Treat AI Competency as a Patient-Safety Requirement. Healthcare professionals using consequential AI should receive role- and risk-specific education addressing intended use, limitations, uncertainty, bias, appropriate reliance, escalation, and override procedures. Competency should extend beyond knowing how to operate AI to knowing when its output warrants questioning. NIST similarly identifies proficiency standards and risk-management training for actors responsible for AI operation and oversight as important governance practices.

5. Protect Meaningful Override and Escalation Authority. Organizations should ensure that qualified professionals can question, modify, override, delay, or escalate consequential AI-supported decisions when warranted. These mechanisms must be practically accessible and supported by organizational culture and workflow. Formal authority that cannot realistically be exercised does not constitute meaningful oversight.

6. Monitor Human Performance Alongside AI Performance. Post deployment monitoring should assess the human–AI system, not the algorithm alone. Organizations should monitor appropriate indicators of AI accuracy, reliability, bias, drift, and safety alongside human competency, error detection, reliance, escalation behavior, and AI-related near misses. This recommendation is especially timely because NIST has identified human–AI feedback loops and methods for measuring beneficial impacts on humans as important gaps in Post deployment AI monitoring.

7. Align Accountability with Authority and Control. Responsibility should correspond reasonably to each actor's knowledge, authority, control, foreseeable risk, and capacity to prevent harm. Clinicians should remain accountable for appropriate professional practice, but healthcare organizations and technology developers should retain responsibility for implementation conditions and technological factors within their control. Recent FDA advisory discussions have likewise emphasized collaborative responsibility among regulators, manufacturers, healthcare systems, and clinicians.

8. Integrate HJPS Principles into Existing Governance Structures. Human-judgment protections should be incorporated into existing AI governance, quality improvement, patient safety, enterprise risk management, professional competency, procurement, and accreditation processes rather than creating an entirely separate administrative structure. This approach is particularly important for organizations with limited governance resources.

9. Incorporate Human-Judgment Requirements into AI Procurement. Procurement and vendor contracts for consequential AI should address intended use, limitations, performance information, material system changes, monitoring responsibilities, escalation mechanisms, and information necessary for meaningful human oversight. Organizations should avoid acquiring systems whose design or contractual arrangements prevent them from obtaining information necessary to govern high-risk use responsibly.

10. Build the Evidence Base for Human-Protected Judgment. Federal agencies, healthcare systems, academic institutions, and funders should support longitudinal research examining automation bias, calibrated reliance, cognitive offloading, deskilling, competency retention, human–AI team performance, and methods for measuring meaningful oversight. The objective should be to establish evidence-based thresholds for determining which human capabilities must be retained, under what circumstances, and how those capabilities can be measured without unnecessarily restricting beneficial automation.

These 10 recommendations provide an integrated policy agenda for protecting human judgment across the lifecycle of healthcare AI. Although each recommendation addresses a distinct governance priority, their effectiveness depends on coordinated implementation across healthcare organizations, regulatory and accreditation bodies, professional organizations, technology developers, and research institutions.

Table 5 consolidates the recommendations into an actionable framework by identifying each policy priority, its corresponding action, and the principal stakeholders responsible for advancing implementation.

Table 5. HJPS Policy Recommendations and Action Agenda

Policy priority	Primary action	Key actors
1. Human-judgment protection	Recognize meaningful human capacity as an AI governance objective	Policymakers, regulators, healthcare organizations
2. Risk classification	Apply proportionate safeguards to AI use cases	Healthcare organizations, regulators
3. HJIA	Assess human capability before high-risk deployment	Healthcare organizations
4. AI competency	Establish role- and risk-specific competency requirements	Professional bodies, educators, employers
5. Override and escalation	Protect practical authority to challenge AI	Healthcare organizations
6. Dual-domain monitoring	Monitor human and AI performance	Quality, safety, AI governance teams
7. Accountability	Align responsibility with meaningful control	Regulators, organizations, developers
8. Governance integration	Embed HJPS within existing quality and risk structures	Healthcare organizations, accreditors
9. Procurement	Include human-oversight requirements in contracts	Healthcare systems, vendors
10. Research	Develop evidence and measures of human-capability effects	Government, academia, healthcare systems

Implementation Priority

Implementation should be phased and risk proportionate. Initial efforts should focus on AI applications that materially influence high-consequence clinical or access-related decisions rather than attempting comprehensive implementation across every automated function. Particular attention is warranted for smaller, rural, safety-net, behavioral health, correctional, long-term-care, and other resource-constrained settings so that stronger AI governance does not become feasible only for large healthcare systems.

The policy objective is therefore not to slow responsible AI adoption. It is to ensure that technological advancement does not proceed faster than healthcare's capacity to govern the changing relationship between artificial and human intelligence. As WHO emphasized in its 2026 discussion of AI in health policy, AI can extend analytical capability, but judgment, contextual interpretation, and ethical deliberation remain essential human functions.

Implementation Challenges and Policy Trade-Offs

Implementation of the HJPS will require balancing the protection of meaningful human judgment against the benefits of automation, organizational efficiency, workforce capacity, and technological innovation. The objective should not be to preserve human involvement where it adds little value, but to protect those human capabilities that remain necessary for safe, ethical, and accountable healthcare. This distinction is especially important because AI may simultaneously reduce cognitive burden, improve performance, and create new forms of dependence. Accordingly, implementation should be risk proportionate, evidence based, and adaptable as AI capabilities evolve.

Avoiding Excessive Human Oversight

A principal trade-off involves determining how much human involvement is appropriate. Requiring independent review of every AI-generated output could increase workload, create alert fatigue, reduce efficiency, and undermine the benefits of automation. Conversely, insufficient review of consequential applications may allow automation bias or declining independent capability to remain undetected. The HJPS therefore does not equate responsible governance with maximum human control. Its purpose is to preserve meaningful rather than merely frequent human involvement, particularly where errors could cause substantial or difficult-to-reverse harm.

Measuring Human Judgment

A second challenge is measurement. AI performance can often be evaluated using quantifiable indicators such as accuracy, sensitivity, reliability, or error rates. Human judgment is more complex. Override frequency, for example, is not inherently a measure of good judgment: low override rates may reflect appropriate agreement with a highly accurate system or excessive reliance, while high rates may indicate either effective vigilance or inappropriate distrust. Human-judgment monitoring must therefore combine measures such as competency, error recognition, appropriate reliance, escalation behavior, near misses, and performance under selected independent or downtime conditions. These measures will require empirical refinement before standardized thresholds can be established.

Organizational Capacity and Equity

Implementation capacity will also vary substantially across healthcare organizations. Recent WHO assessments identify workforce preparedness, governance, legal uncertainty, affordability, data quality, infrastructure, training, time, and workload among continuing challenges affecting AI adoption in health systems. Large systems may have dedicated AI governance, informatics, legal, quality, and data-science resources, whereas smaller or resource-constrained organizations may not. AHRQ similarly emphasizes that hospitals differ considerably in AI readiness and resources and that safe implementation requires workforce preparation, ongoing oversight, and organizational governance.

The HJPS should therefore avoid creating a governance model that unintentionally widens disparities between well-resourced and under-resourced healthcare systems. Simplified assessment tools, shared guidance, technical assistance, and integration with existing quality and safety processes may reduce implementation burden. Equity is also relevant to the technologies themselves: WHO has warned that digital-health benefits may be distributed unevenly when structural barriers are not addressed throughout regulation, implementation, and evaluation.

Innovation, Evidence, and Adaptability

Finally, human-judgment protection must evolve with evidence. Some capabilities currently regarded as essential may become less necessary as AI demonstrates sustained superiority and reliability in narrowly defined functions, while new forms of AI may create human competencies not yet anticipated. WHO's recent policy work similarly emphasizes risk-based regulation, continuing human oversight, and the use of AI to augment rather than simply replace human judgment.

The HJPS should consequently be treated as a proposed and evolving policy framework, not a fixed prescription. Pilot implementation and longitudinal research are needed to validate its domains, determine appropriate HJIA criteria, develop measures of meaningful human judgment, and establish when safeguards can appropriately be strengthened or reduced. This adaptive approach recognizes the central policy trade-off: healthcare must protect human judgment without protecting unnecessary human tasks, and embrace automation without relinquishing capabilities that remain essential to patient safety, accountability, resilience, and ethical care.

2. CONCLUSION

Artificial intelligence is reshaping healthcare by expanding the capacity to analyze information, predict outcomes, automate processes, and support clinical and organizational decisions. While these capabilities offer substantial opportunities to improve quality, efficiency, access, and safety, responsible AI governance must address more than technological accuracy, security, transparency, and equity. As AI assumes greater cognitive responsibility, healthcare systems must also ensure that professionals retain the capacity to exercise meaningful judgment when overseeing consequential AI-supported decisions.

This white paper identifies the human-capability gap as the difference between formally requiring human oversight and ensuring that professionals retain the cognitive capacity, competence, information, authority, and practical ability necessary to provide that oversight meaningfully. The proposed Human Judgment Protection Standard (HJPS) addresses this gap by

shifting governance from merely keeping humans *in the loop* toward maintaining human capacity in the loop. Its six interdependent domains, risk-proportionate classification, Human-Judgment Impact Assessment, and continuous implementation cycle provide a structure for incorporating human capability into existing AI governance.

The HJPS does not assume that human judgment is always superior to AI or that every traditionally human function should be preserved. Rather, human capabilities warrant protection when healthcare systems continue to depend on them for recognizing errors, interpreting context, managing exceptions, incorporating patient values, exercising ethical judgment, responding to system failure, or assuming responsibility for consequential decisions. Accordingly, the HJPS should be regarded as a proposed policy framework requiring empirical evaluation, including validation of its domains and HJIA, development of measures of human capability, and pilot implementation across diverse healthcare settings.

Ultimately, responsible healthcare AI will depend not only on increasingly capable technologies but also on how deliberately healthcare systems govern the relationship between artificial and human intelligence. Protecting meaningful human judgment does not require resisting technological progress; it requires ensuring that progress strengthens rather than diminishes the human capabilities upon which safe, ethical, accountable, and resilient healthcare continues to depend. As healthcare becomes more artificially intelligent, its governance must become equally intentional about protecting the human capacity to judge when that intelligence should be trusted, questioned, or overridden.

REFERENCES

[1] Agency for Healthcare Research and Quality. (2025). *Riding the AI wave: Moving forward*. U.S. Department of Health and Human Services.

[2] Assistant Secretary for Technology Policy/Office of the National Coordinator for Health Information Technology. (2024). *Health data, technology, and interoperability: Certification program updates, algorithm transparency, and information sharing (HTI-1) final rule*. U.S. Department of Health and Human Services.

[3] Bull, D. A. (2026a). Human-centered artificial intelligence and teaching presence theory (HC-AI-TP): A theory of instructional amplification in AI-enabled education. *International Journal of Computer Science and Information Technology Research, 14*(1), 19–44. https://doi.org/10.5281/zenodo.18359914

[4] Bull, D. A. (2026b). Artificial intelligence as a "mixed blessing" in online higher education: A human-centered AI–teaching presence (HC-AI-TP) perspective. *International Journal of Computer Science and Information Technology Research, 14*(1), 88–103. https://doi.org/10.5281/zenodo.18458137

[5] Bull, D. A. (2026c). Conscious Intelligence Integration Framework: A framework for reflective, ethical, and epistemically responsible engagement with artificial intelligence. Zenodo. https://doi.org/10.5281/zenodo.20459469

[6] Joint Commission, & Coalition for Health AI. (2025). *Guidance on the responsible use of AI in healthcare*.

[7] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

[8] Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

[9] U.S. Food and Drug Administration. (2025). *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and Food and Drug Administration staff*. U.S. Department of Health and Human Services.

[10] U.S. Food and Drug Administration. (2026). *Clinical decision support software: Guidance for industry and Food and Drug Administration staff*. U.S. Department of Health and Human Services.

[11] van de Sande, D., Economou-Zavlanos, N., & van Genderen, M. E. (2026). Meaningful oversight of medical AI beyond human in the loop. *npj Digital Medicine, 9*, Article 569. https://doi.org/10.1038/s41746-026-02971-1

[12] World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization.

[13] World Health Organization. (2026). *Artificial intelligence and evidence-informed policy: Emerging challenges and opportunities: Discussion paper*. World Health Organization.

[14] World Health Organization Regional Office for Europe. (2025). *Accelerating the uptake of digital solutions by the health and care workforce in the WHO European Region*. World Health Organization.

[15] World Health Organization Regional Office for Europe. (2026). *Equity across the regulation, implementation and evaluation of digital health: Scoping review*. World Health Organization.

[16] Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

[17] Heudel, P. E., Crochet, H., Filori, Q., Bachelot, T., & Blay, J. Y. (2026). Artificial intelligence in medicine: A scoping review of the risk of deskilling and loss of expertise among physicians. *ESMO Real World Data and Digital Oncology, 12*, 100693. https://doi.org/10.1016/j.esmorw.2026.100693

[18] Al-Anezi, F. M. (2026). Generative artificial intelligence in healthcare: Automation bias, deskilling, and cognitive implications—A systematic review. *Journal of Healthcare Leadership, 18*, 590498. https://doi.org/10.2147/JHL.S590498

[19] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1