noemicajournal.org

e-ISSN: 3086-7037

ARTICLE

Responsible governance of artificial intelligence in healthcare systems: principles for equity, transparency, and institutional capacity

Governança responsável da inteligência artificial em sistemas de saúde: princípios para equidade, transparência e capacidade institucional

Gobernanza responsable de la inteligencia artificial en sistemas de salud: principios para la equidad, la transparencia y la capacidad institucional

Antonio Pires Barbosa[1]

[1] Universidade Nove de Julho (UNINOVE), São Paulo, SP, Brasil. ORCID: https://orcid.org/0000-0001-6478-6522.

SUBMITTED	ACCEPTED	PUBLISHED	DOI
12/03/2025	20/10/2025	05/12/2025	in process

ABSTRACT

The incorporation of artificial intelligence into healthcare systems raises governance challenges that go beyond the technical validation of algorithms. This article examines how normative principles can guide the responsible adoption of diagnostic, triage, and clinical management systems, with particular attention to equity in care, transparency of automated decisions, and the strengthening of institutional capacity within health organizations. The discussion draws on an integrative review of the literature on clinical governance, patient safety, distributive justice, algorithmic explainability, and organizational capabilities, engaging contributions published in journals such as Nature Medicine, The New England Journal of Medicine, Science, and BMC Medicine, as well as the World Health Organization guidance on ethics and governance of artificial intelligence for health. Building on this foundation, the article proposes a model composed of six interrelated dimensions: legitimate clinical purpose, data quality and representativeness, contextual validation, effective human oversight, contestation mechanisms, and continuous monitoring. It argues that the technical performance of an

algorithm, measured in isolation through metrics such as sensitivity or the area under the ROC curve, does not guarantee public benefit when systems are deployed across unequal care networks with disparate infrastructure, training, and resources. The article's contribution lies in integrating technical, ethical, legal, and administrative requirements into a framework applicable across the entire lifecycle of artificial intelligence in healthcare, from procurement to eventual decommissioning, offering guidance for both service managers and policy makers.

Keywords: artificial intelligence; healthcare; governance; equity; transparency; institutional capacity.

INTRODUCTION

Artificial intelligence systems are already part of the routine of many health services. They support the reading of imaging exams, help stratify cardiovascular risk, flag patients with a higher chance of clinical deterioration, organize waiting lists, and assist in planning hospital resources. Topol (2019) described this movement as part of a transition toward high-performance medicine, in which machine learning algorithms process volumes of data that exceed human capacity for manual analysis. This shift brings real gains in certain contexts, but it also introduces risks that are not limited to technical software failures.

The first risk is clinical error resulting from poorly calibrated models or models applied outside the context for which they were validated. An algorithm trained on data from a large university hospital may perform very differently when used in a basic health unit serving a different patient profile. Kelly et al. (2019), in a review published in BMC Medicine, showed that most studies on artificial intelligence in healthcare lack robust external validation, which limits the extent to which results reported in technical papers can be generalized to everyday clinical practice.

The second risk involves the opacity of the models. Deep neural networks and other machine learning methods often function as black boxes, producing recommendations without an intuitive explanation of the reasoning behind them. This characteristic makes clinical oversight and accountability difficult when something goes wrong, a problem already discussed by Verghese, Shah, and Harrington (2018), who warned that blind trust in computational recommendations may weaken rather than strengthen medical judgment.

The third risk, perhaps the most studied in recent years, is inequality. Obermeyer et al. (2019), in an article published in the journal Science, demonstrated that an algorithm widely used in the United States to identify patients with more complex care needs systematically discriminated against Black patients, because it used healthcare spending as a proxy for clinical need. Since Black patients have historically had less access to services, even with equivalent health conditions, the model underestimated the severity of their conditions. This case has become a required reference in any discussion of artificial intelligence governance in healthcare, because it shows that a system with good aggregate performance metrics can still produce serious distributive harm.

These three risks, clinical error, opacity, and inequality, are not engineering failures that can be solved merely with more sophisticated models. They stem from institutional decisions about how technology is chosen, deployed, supervised, and reviewed. This is the central argument of this article: the governance of artificial intelligence in healthcare must translate broad ethical principles into concrete responsibilities, verifiable procedures, and criteria that allow subsequent auditing. Without this translation, relevant decisions tend to migrate to technology vendors or to isolated specialists, reducing the ability of the health organization to exercise effective oversight over tools that directly affect patients' lives.

A second thread of the argument concerns the unequal distribution of capabilities across health organizations. University hospitals, large private networks, and well-funded public systems have data science teams, legal departments, and bargaining power with vendors. Smaller units, especially in regions with lower health investment, generally lack this structure. If governance of artificial intelligence is treated only as a technical problem to be solved internally by each organization, the likely outcome is a concentration of benefits in institutions that are already better resourced, widening inequalities between territories and populations. Support policies, shared standards, and public evaluation infrastructure can mitigate this asymmetry, but they require deliberate political decision-making.

A third thread concerns the need for continuous monitoring. The performance of an artificial intelligence system is not static. Changes in the population served, in clinical protocols, in image-capture equipment, or even in medical record coding habits can substantially alter the quality of predictions over time, a phenomenon known in the technical literature as data drift. Sendak et al. (2020), in a study published in NPJ Digital Medicine on the implementation of a predictive sepsis model within an American

hospital system, showed how the process of validation, adjustment, and continuous monitoring was as important as the algorithm's initial development. Without periodic review, a system validated at one point in time may become inadequate without anyone noticing, until the harm has already occurred.

Finally, it is worth highlighting the dimension of accountability. The greater a system's potential to affect rights, resources, or treatment opportunities, the greater the level of documentation, scrutiny, and possibility of appeal that should be available to affected patients. This proportionality between risk and control is a recurring principle both in the literature on medical device regulation and in recent artificial intelligence guidelines, including the World Health Organization document on the ethics and governance of artificial intelligence for health (World Health Organization, 2021) and the risk-based approach adopted by the European Artificial Intelligence Regulation.

This article is organized as follows. After this introduction, the theoretical framework underpinning the proposal is presented, bringing together clinical governance, patient safety, distributive justice, and the theory of organizational capabilities. Each of the six dimensions of the proposed model is then developed, discussing clinical purpose, data, validation, transparency, human oversight, contestation, monitoring, and institutional capacity. A specific section addresses procurement, regulation, professional training, interoperability, and economic evaluation, topics that are often left out of strictly technical discussions of artificial intelligence in healthcare. The article closes with an extended discussion of the model's institutional implications, an analysis of its limitations, and final considerations regarding the research and policy agenda.

THEORETICAL FRAMEWORK

Artificial intelligence as a sociotechnical technology

A common mistake in discussions of artificial intelligence in healthcare is treating the algorithm as an isolated object whose value can be assessed solely through statistical performance metrics. The literature on sociotechnical systems, applied to health informatics since the classic work of Coiera (2007) on clinical information systems, argues the opposite: any technology operates within a network made up of people, work routines, physical infrastructure, and organizational rules. A machine learning model with excellent performance under test conditions can fail completely when

introduced into a workflow that professionals cannot follow, or when it generates alerts at a volume that exceeds the team's capacity to respond.

Cabitza, Rasoini, and Gensini (2017), in an article published in JAMA on the unintended consequences of machine learning in medicine, drew attention to the phenomenon of automation complacency, in which professionals begin to accept algorithmic recommendations without critical questioning, even when the clinical context calls for caution. The opposite also occurs: systems perceived as unreliable are simply ignored, rendering the technological investment useless. Both outcomes result from poor sociotechnical integration, not necessarily from a flaw in the model itself.

This perspective explains why the evaluation of an artificial intelligence system in healthcare cannot be limited to bench testing. It is necessary to consider how professionals interpret recommendations, how the system fits into existing workloads, and how the organization responds when the algorithm gets it wrong. This is the theoretical backdrop against which the six dimensions proposed later are built.

Clinical governance and patient safety

The tradition of clinical governance, consolidated in the United Kingdom from the 1990s onward as a response to serious care failures, offers a useful vocabulary for thinking about artificial intelligence in healthcare. Clinical governance can be understood as the set of structures through which health organizations are made accountable for continuously improving the quality of their services and safeguarding high standards of care. Applied to artificial intelligence, this tradition suggests that the adoption of an algorithm should follow the same rigor required for introducing a new drug or procedure: evidence of efficacy, safety assessment, staff training, and post-implementation monitoring.

Wiens et al. (2019), in a landmark article published in Nature Medicine titled "Do no harm: a roadmap for responsible machine learning for health care," systematized this analogy by proposing a roadmap for the responsible development of machine learning in healthcare, covering everything from formulating the clinical problem to post-implementation maintenance. The authors argue that most of the ethical problems associated with artificial intelligence in healthcare do not stem from bad faith, but from gaps in processes, such as the absence of prospective validation, lack of continuous monitoring, and scarcity of mechanisms to report and correct failures.

Distributive justice and algorithmic bias

The discussion of equity in artificial intelligence in healthcare engages directly with the tradition of distributive justice in bioethics, which asks how the benefits and risks of an intervention should be distributed among different social groups. Machine learning models are trained on historical data, and historical data reflect the inequalities of access, record-keeping, and quality of care that already exist within the health system. When a population group appears less frequently, has lower-quality records, or shows systematically worse outcomes in the training data, the model tends to reproduce, and in some cases amplify, these differences.

The case of Obermeyer et al. (2019), already mentioned, illustrates this mechanism clearly, but it is far from an isolated instance. Panch, Mattie, and Atun (2019), in an article on artificial intelligence and algorithmic bias published in the Journal of Global Health, argue that bias can enter the development cycle of an artificial intelligence system at several stages, from the choice of the problem to be solved to the definition of outcome variables, as well as through the composition of the training dataset and the validation criteria. For this reason, the authors contend that equity auditing cannot be a single step tacked onto the end of development, but must accompany the entire process.

Rajkomar, Hardt, Howell, Corrado, and Chin (2018), in an article published in the Annals of Internal Medicine on fairness in machine learning applied to healthcare, propose a useful distinction among different technical notions of equity, such as statistical parity, equality of opportunity, and balanced calibration across subgroups, showing that these definitions can conflict with one another and that the choice of an equity criterion is ultimately a value judgment, not merely a technical matter. This finding reinforces this article's argument that the governance of artificial intelligence in healthcare requires explicit deliberation about values, not just metric optimization.

Explainability and trust

The ability to explain why a system arrived at a particular recommendation is often treated as synonymous with transparency, but the recent literature suggests greater caution. Detailed technical explanations, such as saliency maps in images or variable importance coefficients, can be incomprehensible to professionals without training in data science, and even more so to lay patients. Amann et al. (2020), in an article on explainability of artificial intelligence in healthcare published in BMC Medical Informatics and Decision Making, argue that explainability needs to be understood across multiple dimensions, including technical explainability, clinical

justification, and communication appropriate to the audience, and that these dimensions serve the different needs of different actors.

From the standpoint of institutional trust, transparency serves a function that goes beyond explaining each individual prediction. It underpins the possibility of external auditing, comparison between competing systems, and accountability when harm occurs. Price and Cohen (2019), discussing privacy and medical big data in Nature Medicine, also warn that transparency about how patient data are used is a precondition for the legitimacy of any artificial intelligence system in healthcare, especially when sensitive data circulate between public and private institutions.

Organizational capabilities and state capacity theory

Finally, the model proposed in this article draws on the literature on organizational capabilities and state capacity, which argues that the effective implementation of any policy or technology depends on qualified human resources, adequate infrastructure, coordination processes, and mechanisms for institutional learning. Sun and Medaglia (2019), in a study on the adoption of artificial intelligence in Chinese public hospitals published in Government Information Quarterly, found that organizational barriers, such as professional resistance, lack of data standardization, and the absence of committed leadership, were just as relevant as technical limitations in explaining the success or failure of artificial intelligence projects in healthcare.

This literature supports the argument that responsible governance of artificial intelligence in healthcare cannot be reduced to an ethical checklist applied once. It requires continuous capacity for evaluation, adjustment, and decision-making, unevenly distributed across organizations, which makes public policy support for institutional capacity an inseparable component of the ethical and technical agenda.

CLINICAL PURPOSE AND PROPORTIONALITY

The first element of the proposed model is the requirement of a legitimate clinical purpose. Before discussing model architecture or data quality, an organization needs to answer a simple question: what clinical or organizational problem the system is intended to solve, for which population, and with what expected benefit. Systems adopted without this clear definition tend to be evaluated only for their technical sophistication, which does not necessarily correspond to clinical usefulness.

Proportionality follows directly from this requirement. A triage system that prioritizes patients in an emergency room queue has far greater potential to cause harm than a tool that organizes administrative scheduling. For this reason, the level of evidence required, the intensity of human oversight, and the rigor of documentation should be proportional to the potential risk of the application, rather than uniform for every tool labeled as artificial intelligence. This logic of risk-based grading is present both in the regulation of software-based medical devices, such as the Food and Drug Administration guidelines for software as a medical device, and in the approach of the European Artificial Intelligence Regulation, which classifies systems used in healthcare predominantly as high-risk, subject to heightened requirements for documentation, risk management, and human oversight.

Char, Shah, and Magnus (2018), in an influential article published in the New England Journal of Medicine on implementing machine learning in healthcare, warned of the risk of adopting technology out of competitive pressure or fashion, without the organization having clarity about the problem to be solved. The authors argue that decisions about adopting artificial intelligence in healthcare should follow a process similar to that used to evaluate new therapies, with stages of evidence-gathering, controlled trials where possible, and review by independent committees before large-scale adoption.

In institutional practice, this requirement translates into documents that record purpose, target population, alternatives considered, expected benefit, and success criteria, approved before the system is acquired. This record serves a dual function: it guides the choice among competing vendors and creates a baseline for later evaluation, making it possible to compare what was promised with what actually occurred after implementation.

DATA, REPRESENTATIVENESS, AND EQUITY

Machine learning models learn patterns from historical data. When that data reflects inequalities of access, record-keeping, or quality of care, the model tends to reproduce those inequalities in its recommendations. This problem, already discussed in the theoretical framework in connection with the case reported by Obermeyer et al. (2019), requires institutional governance to treat the composition of the dataset as a central issue, not as a technical detail to be delegated entirely to the development team.

A recommended practice, present in several recent guidelines, is to evaluate performance by population subgroup, such as race, gender, age, geographic region, and socioeconomic status, in addition to aggregate performance. A system may show excellent overall accuracy and still perform significantly worse for a given subgroup, which would go unnoticed in reports that present only averages. Vayena, Blasimme, and Cohen (2018), in an article on the ethical challenges of machine learning in medicine published in PLOS Medicine, argue that transparency about the composition of the training dataset, including its representativeness limitations, should be a minimum requirement for any system submitted for regulatory approval or institutional acquisition.

Governance must also consider the feedback loop between a system's use and the production of new data. If a triage algorithm directs fewer resources to a particular group, that group's outcomes tend to worsen, generating data that, if used to retrain the model in the future, reinforces the initial pattern of discrimination. This feedback loop is particularly dangerous because it can occur silently, without managers noticing the source of the problem. Documenting the origin of the data, periodically reviewing its composition, and maintaining equity auditing processes are concrete measures to reduce this risk.

CONTEXTUAL CLINICAL VALIDATION

Performance metrics obtained in a development environment, such as sensitivity, specificity, and the area under the ROC curve calculated on a retrospective test set, do not substitute for validation in the real-world context of use. The difference between performance under controlled conditions and performance in everyday clinical practice is a recurring theme in the literature, often referred to as the gap between technical validation and clinical utility.

Kelly et al. (2019) reviewed studies on artificial intelligence in healthcare and found that most of them do not carry out external validation, that is, testing the model on a population or institution different from the one used in development. Without this step, it is not possible to know whether the reported performance holds up when the system is transferred to a different care setting, with a different patient profile, different data-capture equipment, and different work routines.

Contextual validation must also assess the system's impact on actual clinical decisions and patient outcomes, not just the isolated accuracy of the

prediction. A model may correctly predict the risk of a given complication and still produce no benefit if the recommendation does not change clinical practice, or if professionals do not trust the tool enough to act on it. Prospective studies comparing outcomes before and after implementation, and gradual rollout strategies, starting with pilot units before expanding across the whole network, are ways to reduce the risk of prematurely generalizing results obtained under artificial conditions.

TRANSPARENCY AND EXPLAINABILITY

Health professionals and patients need to understand, to some degree, the role that an artificial intelligence system plays in a clinical decision. This does not mean that every user needs to understand the model's mathematical architecture, but that the organization must ensure explanations appropriate to the context and the audience, avoiding both excessive technical jargon and oversimplifications that induce either excessive trust or unfounded rejection.

Amann et al. (2020) propose distinguishing between different layers of explainability: the technical explanation aimed at developers and auditors, the clinical justification aimed at health professionals, who need to understand why a given recommendation makes sense given the patient's clinical picture, and the communication aimed at the patient, which requires accessible language and attention to differences in health literacy. The absence of any one of these layers undermines the system's governance, even if the others are well developed.

Institutional documentation plays a central role in this dimension. Technical fact sheets describing the system's purpose, the dataset used, known limitations, the population subgroups for which performance was assessed, and the update history should be available for consultation by professionals and, where relevant, by patients and regulatory bodies. This type of documentation, analogous to a drug package insert, turns transparency from an abstract principle into a verifiable institutional practice.

HUMAN OVERSIGHT

The presence of a professional between the system and the final decision is often presented as sufficient guarantee of safety, but the literature shows that this assumption is fragile. Effective human oversight requires that the professional have the time, autonomy, and competence to critically review the recommendation, rather than merely rubber-stamping it. When workload is excessive, when the organization treats disagreement with the system as a

deviation to be penalized, or when the algorithmic output is presented with an authority that discourages questioning, oversight becomes symbolic.

Cabitza et al. (2017) describe this phenomenon as a form of automation complacency, in which the human presence in the process stops functioning as a safety layer and instead legitimizes decisions that, in practice, were never truly reviewed. Institutional protocols need to explicitly define in which situations an algorithmic recommendation may be accepted directly, in which it requires mandatory review, and in which it should be disregarded in the face of contradictory clinical signs. These protocols should also require that the professional's decision be recorded, making it possible later to assess whether human oversight is actually influencing outcomes or functioning merely as a formality.

Ongoing training is a necessary condition for human oversight to be substantive. Professionals need to understand, in general terms, how the system works, what its known limitations are, and in which situations its performance tends to be less reliable. Without this minimum level of knowledge, the formal autonomy to disagree with an algorithmic recommendation does not translate into a real capacity to exercise it.

CONTESTATION AND PATIENT PARTICIPATION

Patients affected by decisions supported by automated systems need access to information about this involvement and to effective channels for raising questions. This right to contest serves a dual function: it protects patients' individual interests and functions as a correction mechanism for the organization itself, since complaints and questions often reveal flaws that do not appear in aggregate performance metrics.

Communication about the use of artificial intelligence needs to take into account differences in health literacy and familiarity with technology among different groups of patients. A standardized explanation, written in technical language, may meet the formal requirements of transparency without actually informing the patient about what is at stake. Social participation initiatives, such as user councils and public consultations on new systems before their large-scale adoption, have been recommended by international guidelines as a way of incorporating this perspective in a structured, rather than merely reactive, manner.

The World Health Organization (2021), in its document on the ethics and governance of artificial intelligence for health, highlights public participation as one of the central principles for the responsible adoption of

these technologies, arguing that decisions about systems that directly affect healthcare should not be made exclusively by technology developers and managers, without involving the communities and patients affected.

MONITORING AND INCIDENT MANAGEMENT

The performance of an artificial intelligence system can deteriorate after implementation, even without any change to the model's code. Changes in the population served, in clinical protocols, in the equipment used to capture exams, or in medical record documentation practices can alter the distribution of input data, a phenomenon that the technical literature calls data drift. Without continuous monitoring, this kind of deterioration can go unnoticed for long periods.

Sendak et al. (2020) describe, drawing on the experience of implementing a sepsis prediction model, how the post-implementation monitoring stage demanded as much institutional attention as the development of the algorithm itself, including adjustments to alert thresholds, periodic performance review, and channels for professionals to report cases in which the recommendation seemed inappropriate. This kind of monitoring structure should be a mandatory part of any artificial intelligence project in healthcare, not an optional add-on.

Incident management requires the systematic recording of adverse events related to the system's use, investigation of causes, and transparent communication about corrective measures. When an identified risk is serious and cannot be quickly mitigated, temporarily suspending use of the system should be an institutional option that is available and actually exercised, not merely a theoretical possibility written into a contract.

INSTITUTIONAL CAPACITY

Responsible governance of artificial intelligence in healthcare depends on concrete institutional capacity: multidisciplinary teams capable of evaluating technical proposals, adequate data infrastructure, well-drafted contracts, and leadership committed to the ongoing oversight of the systems adopted. Organizations need to develop the competence to critically evaluate vendors, which includes the ability to request and interpret validation evidence, negotiate contractual clauses on auditing and performance, and train professionals for the proper use of the technology.

This capacity is not distributed equally among health organizations. Resource-poor networks, especially in regions with lower economic

development, depend almost entirely on external vendors to evaluate and adjust artificial intelligence systems, which reduces their decision-making autonomy and their bargaining power. Sun and Medaglia (2019), in their study on the adoption of artificial intelligence in Chinese hospitals, found that organizational barriers, not just technical limitations, were decisive in explaining the success or failure of different initiatives, reinforcing the idea that investment in institutional capacity is as important as investment in the technology itself.

Public support policies, including shared technical standards, public infrastructure for independent evaluation, and training programs, can reduce this asymmetry and prevent artificial intelligence in healthcare from becoming yet another factor that concentrates advantages among institutions that are already well resourced.

PROPOSED FRAMEWORK

The framework developed in this article brings together the six dimensions discussed: legitimate clinical purpose, data quality and representativeness, contextual validation, transparency and explainability, effective human oversight and contestation mechanisms, and continuous monitoring, which runs through all the others. These dimensions should not be understood as sequential, isolated steps, but as aspects that feed back into one another throughout the system's entire lifecycle, from acquisition to eventual decommissioning.

The decision to adopt an artificial intelligence system in healthcare should be treated as reviewable from the outset, not as a definitive commitment made at the moment of acquisition. This means keeping records that make it possible to compare predicted performance with observed performance, periodically reviewing the technology's continued relevance in light of available alternatives, and, when necessary, discontinuing its use. This orientation toward public value, rather than toward technical efficiency in isolation, is the thread that ties the model's six dimensions together.

PROCUREMENT, REGULATION, AND ACCOUNTABILITY

Procurement of artificial intelligence systems in healthcare requires criteria that combine clinical, technical, and legal evaluation. Tenders and contracts must specify minimum performance requirements, required technical documentation, information security standards, interoperability requirements with existing systems, and conditions for ongoing technical support. Without these elements, the purchasing organization is left at an

informational disadvantage relative to the vendor, unable to adequately assess whether the delivered product matches what was promised.

The allocation of responsibilities among technology developers, health institutions, and the professionals who use the system needs to be explicit. The multiplication of actors involved in developing and using an artificial intelligence system must not result in accountability gaps, situations in which, when harm occurs, each party claims that the fault lies with another. Incident-investigation protocols, defined in advance rather than improvised after a problem occurs, help clarify causes and assign responsibility fairly.

External regulation and internal institutional governance are complementary, not substitutes for one another. Formal compliance with regulatory requirements, such as registration with a health authority or quality certification, does not eliminate the need for local assessment of the system's suitability to the specific context of use. Organizations need to keep track of regulatory updates and of changes in how the system is actually used in practice, which may drift away from its originally intended and approved use.

Contracts should also provide for independent auditing mechanisms, data portability in the event of a vendor change, and clear conditions for terminating the contractual relationship. Excessive dependence on a single vendor, without clauses guaranteeing continuity and transparency, undermines the institution's ability to exercise effective control over the technology it has adopted.

PROFESSIONAL TRAINING AND SAFETY CULTURE

Health professionals need to understand, even at a non-technical level, the capabilities and limits of the artificial intelligence systems they work with. Training programs should include basic notions of how to interpret algorithmic results, what types of bias may affect recommendations, principles of clinical data privacy, and strategies for communicating with patients about the role of technology in the care they receive. The goal is not to turn every professional into a data science specialist, but to develop sufficient critical judgment to use the tool safely.

An organizational culture that allows errors and disagreements to be reported without fear of punishment is a necessary condition for continuous monitoring to work in practice. Punitive environments tend to reduce the spontaneous reporting of problems, concealing failures until they become serious. Confidential reporting channels and systemic analysis of reported

incidents, rather than immediate individual blame, strengthen the organization's capacity for learning.

Multidisciplinary teams, bringing together clinical, technical, and ethical knowledge, help identify problems that would go unnoticed in purely technical analyses. Periodic performance-review meetings, with the participation of different professional categories, make it possible to track both quantitative indicators and qualitative reports about the experience of using the system.

Patients and communities also need accessible information about how artificial intelligence systems are used in their care. Communication materials written in plain language and tested with real users before release help build understanding of uses and rights, contributing to institutional trust more consistently than generic statements about technological innovation.

INTEROPERABILITY AND CONTINUITY OF CARE

Artificial intelligence systems depend on data that frequently flows between different services and levels of care. Technical interoperability, which ensures that different systems can exchange data, and semantic interoperability, which ensures that this data retains the same clinical meaning when transferred between systems, are essential conditions for avoiding fragmentation of information and, with it, a loss of quality in the recommendations generated.

Continuity of care can be undermined when recommendations generated by a system at one service do not follow the patient along their care trajectory. A risk alert generated during a hospital stay, for example, loses its usefulness if it is not communicated to the primary care team responsible for follow-up after discharge. Integration between artificial intelligence systems and electronic health records needs to be planned from the outset of the project, not treated as an afterthought.

Changes of vendor or system version require data migration processes that preserve relevant clinical history. Procurement contracts should explicitly provide for conditions of portability and record preservation, preventing a technology change from resulting in the loss of important care-related information. Well-structured interoperability also facilitates processes of external auditing and scientific research into the actual performance of systems in use, provided that requirements of privacy and legitimate purpose in the secondary use of data are respected.

ECONOMIC EVALUATION AND PUBLIC VALUE

The decision to deploy an artificial intelligence system in healthcare must consider its total cost, which includes not only the license or initial acquisition, but also data infrastructure, professional training, ongoing technical maintenance, and performance monitoring over time. Cost-savings projections presented by vendors often fail to materialize when parallel manual processes remain in place as a precaution, or when implementation requires complementary investments that were not initially foreseen.

Economic evaluation of artificial intelligence systems in healthcare needs to incorporate dimensions of safety, equity, and outcome quality, not just operational efficiency indicators. A reduction in service time, for example, does not represent a real benefit when it is achieved at the cost of a higher diagnostic error rate or additional barriers to access for certain groups of patients.

Pilot projects, conducted on a reduced scale before expansion across the entire care network, help produce more reliable estimates of real costs and of clinical and organizational effects than projections made exclusively on the basis of data supplied by the technology developer. Objective criteria for deciding on scale-up, defined before the pilot begins, prevent expansion decisions from being driven by commercial pressure or by technological enthusiasm not grounded in local evidence.

Transparency about the costs and results of artificial intelligence projects in healthcare enables social oversight and appropriate prioritization of public resources. Since resources allocated to health involve choices with a high opportunity cost, investments in technology need to be publicly justified with the same rigor applied to other decisions about allocating care resources.

RESEARCH AND POLICY AGENDA

Future research on the governance of artificial intelligence in healthcare should prioritize studies conducted in real clinical practice settings, with diverse populations, rather than relying exclusively on retrospective evaluations using controlled databases. Prospective studies, which track clinical outcomes before and after a system's implementation, offer more robust evidence of real benefit than purely technical comparisons between models.

It is necessary to deepen our understanding of how algorithmic recommendations interact with professional clinical judgment under real

working conditions, including organizational factors such as workload, institutional culture, and the relationship of trust between the team and the technology. These factors often explain variations in performance and adherence that cannot be predicted solely from the technical characteristics of the model.

Public policy has an important role to play in supporting common technical standards, shared infrastructure for independent evaluation, and certification mechanisms that reduce the information asymmetry between vendors and health organizations, especially smaller ones. Smaller networks, with limited resources, need structured support to take part in the responsible adoption of these technologies without depending entirely on the goodwill of commercial vendors.

Finally, the governance of artificial intelligence in healthcare needs to be understood as a constantly evolving process, one that keeps pace with both technological advancement and the accumulation of evidence about its real effects on clinical practice. Periodic reviews of institutional and regulatory guidelines are necessary to keep the governance framework up to date in the face of changes that occur at a faster pace than the usual cycles of regulatory review.

EXTENDED DISCUSSION AND INSTITUTIONAL IMPLICATIONS

The analysis developed throughout this article indicates that technological changes produce consistent results only when accompanied by adequate institutional design. The availability of tools, data, and individual skills does not substitute for coordination rules, the definition of decision-making authority, and accountability mechanisms. Health organizations need to make their purposes explicit, assign responsibilities, and define verifiable evaluation criteria. This structure does not eliminate uncertainty, but it creates the conditions for decisions to be justified, reviewed, and improved over time, rather than remaining implicit in the informal choices of managers or vendors.

A second relevant aspect concerns the relationship between technical capability and institutional legitimacy. Artificial intelligence projects in healthcare may show good operational performance, as measured by technical metrics, and still face resistance from professionals and distrust from patients when those affected do not understand the objectives or lack real channels for participating in decisions about the technology's adoption. Legitimacy is built through the combination of technical evidence,

procedural transparency, and fair treatment of the different groups affected. Deliberative processes do not eliminate conflicts of interest or of values, but they help bring into view priorities and consequences that purely technical analyses tend to overlook.

The unequal distribution of resources among health organizations directly shapes the governance of these technologies. Larger institutions have specialized teams, robust infrastructure, and greater bargaining power with vendors, while smaller organizations depend almost entirely on external support. Policies of cooperation between institutions and of sharing evaluation capabilities can reduce this gap. Any impact assessment of an artificial intelligence system in healthcare should look not only at aggregate results, but also at whether the innovation widens or narrows asymmetries between territories, population groups, and different units within the same care network.

The proportionality between risk and control, discussed in the section on clinical purpose, has important practical implications for the design of institutional processes. Low-impact applications, such as administrative scheduling tools, can be managed with relatively simple procedures. Decisions that directly affect diagnosis, treatment, or the prioritization of access to scarce resources require detailed documentation, review by independent bodies, and clear appeal mechanisms for patients who feel wronged. A governance system that treats every application with the same degree of rigor tends to produce two opposite problems: excessive bureaucracy for low-risk cases and insufficient control for high-impact cases.

Institutional learning capacity is another decisive component for the sustainability of the proposed governance. Artificial intelligence projects in healthcare should be treated as adaptive processes, formulated with explicit hypotheses, monitoring indicators, and formal moments of review, rather than as definitive implementations evaluated only at launch. Unexpected results, whether positive or negative, need to generate concrete adjustments to models, contracts, and work routines. Organizations that hide failures, out of fear of being held accountable or of reputational damage, lose valuable opportunities for correction and tend to repeat the same problems in future projects. Environments that combine accountability with legitimate room for course correction favor this kind of learning.

From a managerial standpoint, the governance of artificial intelligence in healthcare requires genuinely multidisciplinary teams, in which technical knowledge engages with operational clinical experience, legal training,

ethical reflection, and the ability to communicate with patients and communities. This diversity of expertise reduces the blind spots that tend to appear when decisions are concentrated within a single professional group, whether technical, clinical, or administrative. At the same time, multidisciplinary teams need a common language and well-defined decision-making processes, or they risk fragmentation and contradictory decisions across different areas of the same organization.

The financial and organizational sustainability of artificial intelligence projects in healthcare depends on recurring resources, not just initial implementation investment. Initiatives funded only for the launch phase tend to deteriorate once the project period or external funding ends. Maintenance costs, model updates, ongoing professional training, and performance monitoring need to be included from the initial budget planning stage. The decision to continue, adjust, or discontinue an artificial intelligence initiative in healthcare should explicitly take into account the value actually produced, the risks identified during use, and the available alternatives, including the possibility of returning to non-automated processes when these prove better suited to the context.

Finally, the development of artificial intelligence systems in healthcare must remain oriented toward public and organizational purpose, rather than toward technological sophistication for its own sake. The value of any tool derives from its ability to solve concrete problems, strengthen the capacity of health institutions, and produce benefits distributed fairly across the population served. The evaluation of any project should periodically return to the question that guided its creation: which needs were actually met, for whom, and with what consequences for the different groups affected.

LIMITATIONS OF THE MODEL AND CRITERIA FOR APPLICATION

The model presented in this article is theoretical in nature and needs to be adapted to the specific context of each organization. Health services vary substantially in size, institutional mission, availability of resources, regulatory environment, and technological maturity. Mechanically applying the recommendations discussed here, without such adaptation, may create new problems rather than solving them. Before implementing any of the proposed dimensions, it is necessary to map existing processes, identify relevant stakeholders, assess risks specific to the local context, and honestly recognize the capabilities available.

Another limitation stems from the speed of technological change in the field of artificial intelligence. Tools, technical standards, and market configurations evolve rapidly, while institutional rules and regulatory processes usually take longer to consolidate. Governance needs to be stable enough to provide legal and operational certainty, yet flexible enough to incorporate new evidence without requiring a complete overhaul with every technological advance. Scheduled reviews at regular intervals, rather than only in response to crises, help balance these two seemingly contradictory needs.

The availability and quality of data used for evaluation and monitoring also limit, in practice, how the model can be applied. Performance indicators may be incomplete, show inconsistencies across different sources, or be produced directly by the technology vendors themselves, which compromises their independence. Triangulating evidence from different sources, including external audits and qualitative observations of the system's everyday use, reduces dependence on any single performance measure. When the degree of uncertainty about how a system actually functions is high, decisions to expand its use should be gradual and reversible, avoiding institutional commitments that are difficult to undo.

Conflicts of interest deserve specific attention in any practical application of the model. Technology vendors may influence, directly or indirectly, the definition of success criteria and present results selectively, highlighting favorable scenarios while omitting relevant limitations. Contracted consultants, researchers linked to projects, and managers seeking quick results may also have incentives that do not fully align with patients' interests. Formal conflict-of-interest declarations, transparency about contractual terms, and review by independent parties help mitigate, though not entirely eliminate, this type of institutional bias.

Effective application of the model requires committed institutional leadership. Without explicit support from senior management, evaluation committees and oversight protocols tend to become bureaucratic formalities with little practical effect on the organization's actual decisions. Leaders need to make sufficient resources available, respond substantively to the recommendations produced by these processes, and accept, when necessary, the possibility of halting ongoing initiatives, even in the face of investment already made.

Practical criteria for applying the model should include the clinical or organizational relevance of the proposed technology, proportionality

between risk and the level of control required, available institutional capacity, adequate participation of affected professionals and patients, and a structure for continuous monitoring. Each of these criteria can be recorded in a simple decision matrix, which facilitates comparison among competing technological alternatives and keeps a record of the justifications used for each choice, creating a useful history for future evaluations.

It should finally be noted that the decision not to adopt a given technology, or to discontinue its use after a period of evaluation, is a legitimate and sometimes desirable outcome of the governance process. Responsible innovation in healthcare includes the institutional capacity to recognize when the expected benefits of a system do not outweigh its real costs and risks. Well-planned discontinuation processes should preserve relevant clinical data, ensure continuity of the services provided to patients, and keep a record of the knowledge accumulated during the period of use, so that this experience can inform future decisions. These limitations do not diminish the usefulness of the framework proposed in this article; on the contrary, they indicate the conditions necessary for its prudent use. The model's main contribution lies in organizing, in an integrated way, questions that are often handled in a fragmented manner by different areas within the same health organization.

FINAL CONSIDERATIONS

Artificial intelligence can make a significant contribution to health systems, supporting diagnosis, triage, resource management, and care planning. Throughout this article, the argument has been that these benefits depend on governance structures capable of translating broad ethical principles into concrete responsibilities, verifiable procedures, and effective correction mechanisms. Technical performance measured in isolation, through statistical metrics of accuracy or discrimination, does not guarantee equity in care, patient safety, or the strengthening of the institutional capacity of the health organizations involved.

The proposed model, organized around six interrelated dimensions, legitimate clinical purpose, data quality and representativeness, contextual validation, transparency and explainability, effective human oversight, contestation mechanisms, and continuous monitoring, seeks to offer guidelines applicable both to health service managers and to policy makers responsible for regulating these technologies. The article's central contribution lies in treating these requirements as components of a single system, rather than as independent checklists to be met in isolation.

It is acknowledged, however, that the model is theoretical in nature and that its practical application requires careful adaptation to each specific institutional context. Future research could test the framework's applicability across different care networks, empirically assessing whether organizations that adopt these six dimensions actually achieve better outcomes in terms of safety, equity, and institutional legitimacy than organizations that treat artificial intelligence predominantly as a technical matter. Comparative studies among health systems with different levels of institutional capacity and different regulatory frameworks could also clarify the extent to which the principles discussed in this article hold up in contexts other than the one that originally motivated their formulation.

Finally, it is worth reaffirming that the responsible incorporation of artificial intelligence into health systems is not merely a technical challenge to be solved by engineers and data scientists, nor merely a regulatory matter to be settled by legislators. It is a continuous process of institutional construction, requiring coordination among different actors, explicit deliberation over competing values, and a willingness to revisit decisions in light of new evidence. The governance discussed throughout this article represents an attempt to organize this complexity into a coherent framework, without claiming to exhaust a debate that, given the speed of the ongoing technological transformations, will continue to develop for many years to come.

An additional question, little explored in the Brazilian literature on the topic, concerns the relationship between artificial intelligence governance and broader reform of health systems. In the context of universal public systems, such as Brazil's Unified Health System (Sistema Único de Saúde), decisions about technology adoption cannot be made in isolation by each care unit, at the risk of producing a mosaic of mutually incompatible practices, with widely varying quality and safety standards from one region to another. Coordination across levels of management, municipal, state, and federal, becomes necessary to ensure that minimum governance principles are observed throughout the network, without, however, precluding local adaptations justified by genuine differences in care context.

This tension between national standardization and local adaptation appears repeatedly in other areas of health policy and is not unique to artificial intelligence. Accumulated experience with clinical protocols, health information systems, and access-regulation policies suggests that successful solutions tend to combine general guidelines set at the central level with room for operational adjustment at the local level. Applied to artificial

intelligence, this logic suggests that regulatory agencies and health-system management bodies should define minimum requirements for safety, transparency, and equity, leaving each organization the task of translating these requirements into processes compatible with its operational reality.

It is also worth considering the role of academic research in sustaining this governance process. Universities and research centers can contribute significantly, not only by developing new algorithms, but also by producing independent evaluations of the real-world performance of systems already in use, something commercial vendors rarely have an incentive to do impartially. Partnerships between research institutions and health services, with adequate safeguards for data protection and conflicts of interest, can function as a complementary quality-control mechanism, particularly in health systems with limited regulatory capacity.

Finally, it is worth noting the temporal dimension of the governance discussed in this article. Many of the failures reported in the literature on artificial intelligence in healthcare did not occur at the moment of initial implementation, but emerged months or years later, as the context of use gradually drifted away from the original validation conditions. This reinforces the idea that governance is not a single event located at the start of a project, but a continuous process that accompanies the system throughout its entire useful life. Organizations that treat initial approval as the endpoint of the evaluation process tend to be caught off guard by problems that could have been identified and corrected in time, had a permanent monitoring structure been in place.

REFERENCES

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517-518.

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

Coiera, E. (2007). Putting the technical back into socio-technical systems research. International Journal of Medical Informatics, 76, S98-S103.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. Journal of Global Health, 9(2), 010318.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37-43.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866-872.

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Revoir, M., Yashar, F., Miller, C., Kester, K., Sandhu, S., Corey, K., Brajer, N., Tan, C., Lin, A., Brown, T., Engelbosch, S., Anstrom, K., ... Balu, S. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. NPJ Digital Medicine, 3(1), 84.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare. Government Information Quarterly, 36(2), 368-383.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. PLOS Medicine, 15(11), e1002689.

Verghese, A., Shah, N. H., & Harrington, R. A. (2018). What this computer needs is a physician: humanism and artificial intelligence. JAMA, 319(1), 19-20.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340.

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. Geneva: World Health Organization.

AUTHOR CONTRIBUTION (CREDIT TAXONOMY)

Antonio Pires Barbosa: Conceptualization; Methodology; Investigation; Formal analysis; Data curation – not applicable, as no empirical data were used; Writing – original draft; Writing – review & editing; Supervision; Funding acquisition – not applicable, no external funding. Simulated entry for technical testing of the OJS platform.

EDITORIAL DECLARATIONS

Funding: no external funding was received.

Conflict of interest: not applicable.

Use of artificial intelligence: not applicable.

DATA AVAILABILITY: NOT APPLICABLE; NO EMPIRICAL DATA WERE USED.