Revista de Estudos Constitucionais, Hermenêutica e Teoria do Direito (RECHTD)
18(1):02-26, janeiro-abril 2026
Unisinos - doi: 10.4013/rechtd.2026.181.01

From Guidelines to Algorithms: How AI is Rewriting the *Leges Artis* and Medical Liability in Europe*

Das Diretrizes aos Algoritmos: Como a IA está a Reescrever as *Leges Artis* e a Responsabilidade Médica na Europa

Mario Caterini[1]
University of Calabria (Italy)
mario.caterini@unical.it

Antonella Guzzo[2]
University of Calabria (Italy)
antonella.guzzo@unical.it

Marianna Rocca[3]
University of Calabria (Italy)
marianna.rocca@unical.it

Abstract

The integration of artificial intelligence (AI) systems into clinical practice requires a heavy reconsideration of legal categories related to medical liability, not only nationally but also within the increasingly interconnected European framework. This article examines the transition from the traditional *lex artis*, rooted in national jurisdictions, to a new paradigm shaped by EU governance. Starting from the Italian regulatory framework established by Law No. 24/2017 ("Gelli-Bianco"), the analysis then turns to Regulation (EU) 2024/1689 (the AI Act), which classifies AI systems in healthcare as "high-risk" and imposes obligations of transparency and human oversight. A comparative perspective is then developed, focusing on Italy, Germany, France, and Spain, and revealing different conceptions of the standard of care: from the German *fachärztlicher Standard* to the Spanish *lex*

* The authorship of the paragraphs is as follows: Mario Caterini, paragraphs 1, 4, 6, and 8; Antonella Guzzo, paragraph 5; Marianna Rocca, paragraphs 2, 3, and 7.

[1] Full Professor of Criminal Law and Director of the Institute of Criminal Studies 'Alimena'", University of Calabria. Department of Culture, Education and Society. Pietro Bucci street, Building 18/B, 5th floor, 87036, Arcavacata di Rende (CS), Italy.
[2] Associate Professor of Information Processing Systems, University of Calabria. Department of Computer Engineering, Modeling, Electronics and Systems Engineering (DIMES). Pietro Bucci street, Building 42/B, 6th floor, 87036, Arcavacata di Rende (CS), Italy.
[3] PhD candidate in Criminal Law, University of Calabria. Department of Business and Legal Sciences. Pietro Bucci Street, Building 18/B, 5th floor, 87036, Arcavacata di Rende (CS), Italy.

artis ad hoc, through the French dualism between fault-based liability and "national solidarity". Special attention is devoted to explainable artificial intelligence (XAI) as a means to mitigate the opacity of black-box models, together with the implications of the new directive on liability for defective products. The article further explores emerging scenarios from human-machine interaction, such as concordant error and dissent from algorithmic recommendations, assessing them against national approaches. The study reveals a legal landscape in profound transformation, where technological innovation acts as a catalyst for change in diverse national legal cultures, and concludes by sketching a roadmap for governing the possible convergence toward a European model of "augmented medical liability", in which harmonized EU principles are interwoven with national legal traditions, seeking equilibrium between technological innovation, patient safety, and legal certainty.

Keywords: artificial intelligence; medical liability; clinical guidelines; explainable artificial intelligence (XAI).

Resumo

A integração de sistemas de inteligência artificial (IA) na prática clínica exige uma profunda reconsideração das categorias jurídicas relacionadas à responsabilidade médica, não apenas em âmbito nacional, mas também no contexto europeu cada vez mais interconectado. Este artigo examina a transição da tradicional *lex artis*, enraizada nas jurisdições nacionais, para um novo paradigma moldado pela governança da UE. Partindo do quadro regulatório italiano estabelecido pela Lei nº 24/2017 ("*Gelli-Bianco*"), a análise se volta para o Regulamento (UE) 2024/1689 (*AI Act*), que classifica os sistemas de IA na área da saúde como de "alto risco" e impõe obrigações de transparência e supervisão humana. Em seguida, desenvolve-se uma perspectiva comparativa, com foco na Itália, Alemanha, França e Espanha, revelando diferentes concepções do padrão de cuidado: do padrão médico alemão (*fachärztlicher Standard*) à *lex artis ad hoc* espanhola, passando pelo dualismo francês entre responsabilidade baseada em culpa e "solidariedade nacional". É dada especial atenção à inteligência artificial explicável (XAI) como meio de mitigar a opacidade dos modelos de caixa preta, juntamente com as implicações da nova Diretiva sobre a responsabilidade por produtos defeituosos. O artigo explora ainda cenários emergentes da interação humano-máquina, como erros concordantes e discordâncias em relação às recomendações algorítmicas, avaliando-os à luz das abordagens nacionais. O estudo revela um panorama jurídico em profunda transformação, onde a inovação tecnológica atua como catalisador de mudanças em diversas culturas jurídicas nacionais, e conclui esboçando um roteiro para governar a possível convergência rumo a um modelo europeu de "responsabilidade médica aumentada", no qual princípios harmonizados da UE se entrelaçam com as tradições jurídicas nacionais, buscando o equilíbrio entre inovação tecnológica, segurança do paciente e segurança jurídica.

Palavras-chave: inteligência artificial; responsabilidade médica; diretrizes clínicas; inteligência artificial explicável (XAI).

Introduction

Law and medical practice are currently undergoing a paradigmatic transition, marked by the apparent tension between two approaches often deemed irreconcilable (Nuti and Panero 2011). On one side lies personalization, the dominant principle of contemporary medicine, grounded in patient autonomy and centrality. It demands an art of care responsive to the irreducible biological, biographical, and psychological singularities of each individual. On the other side, 'standardization', embodied most prominently in the rapid rise of artificial intelligence, ushers in a revolution that, while promising unprecedented levels of therapeutic individualization, does not rely on the physician's intuition or empathy but rather on inferential patterns extracted from vast repositories of *big data* (Johnson et al. 2021). This represents a new form of uniformity, no longer anchored in evidence-based medicine protocols, but in statistical modeling and machine self-learning.

If personalization and standardization are often conceived as polar opposites, it may be because the complexity of clinical practice – nourished by both – is overlooked (Mannion and Exworthy 2017). In fact, clinical reasoning depends simultaneously on adherence to validated protocols and on their adaptation to the concrete case. Their relationship is not one of frontal opposition but of uneasy coexistence. It is at this conceptual threshold that one of the most decisive challenges of contemporary health law emerges: redefining the classical categories of professional liability in a context where medicine oscillates between the humanity of the therapeutic relationship and the efficiency of automated computation, oriented toward an "individualized standardization of care" (Pfaff et al. 2010). The task is not to choose one model over the other, but to conceptualize an innovative synthesis of human and artificial intelligence. This is not a mere compromise but a more advanced form of care that requires an ethics capable of capturing complexity without reducing it to formulas or surrendering it to technocracy.

What was once a debate confined within national jurisdictions now unfolds on a supranational scale. Regulation (EU) 2024/1689 (AI Act) acts as a catalyst, compelling a pan-European reconsideration of medical liability paradigms. This raise pressing questions about whether a harmonized European framework can effectively accommodate the heterogeneous legal traditions of Member States regarding professional fault. The critical issue is whether Europe's new governance of AI will foster convergence in liability regimes, or whether national specificities will continue to fragment the legal landscape in a domain so central to the protection of fundamental rights.

To address the above issue, this study adopts a legal-comparative and interdisciplinary approach, aimed at linking the Italian regulatory framework with recent European developments and the experiences of other national legal systems. In Section 2, the analysis begins with an examination of Regulation (EU) 2024/1689 (AI Act), which introduces specific obligations for the use of artificial intelligence systems in healthcare and how it will inevitably influence the liability of healthcare professionals. A comparative discussion then follows in Section 3, extending the focus to Italy, Germany, France, and Spain, and highlighting the

differences between the national legal frameworks and consequently, resulting implications for medical liability. The study subsequently turns to conceptual and epistemological aspects, such as the opacity of black-box models and the potential role of explainable AI (XAI) in reducing informational asymmetry in sections 4 and 5. Finally, novel scenarios of human-machine interaction, such as concordant error or dissent from algorithmic recommendations, are explored in sections 6 and 7, in light of diverse national legal traditions. This methodological path guides the reader from the regulatory dimension to comparative insights, culminating in a forward-looking reflection on the possible convergence toward a European model of "augmented medical liability".

The European Regulatory Framework: the AI Act and New Paradigms of Liability

The European Union has sought to bring order to the development and use of artificial intelligence through pioneering legislation designed to establish clear and harmonized rules (Ruschemeier et al. 2025). Regulation (EU) 2024/1689, commonly referred to as the AI Act, not only governs the placing of technological products on the market but also exerts a direct impact on standards of diligence and expectations of safety in critical sectors such as healthcare, with significant consequences for liability regimes.

The AI Act adopts a risk-based approach, calibrating requirements according to the potential severity of harm an AI system might cause to health, safety, and fundamental rights. Within this taxonomy, healthcare occupies a position of particular prominence. Article 6(1) of the Regulation, read in conjunction with Annex III, classifies most AI systems deployed in medicine as "high-risk". These include software functioning as medical devices, clinical decision-support systems, diagnostic algorithms and applications influencing therapeutic or surgical interventions. This classification is not merely formal: it entails stringent obligations on both developers and users. Importantly, the AI Act operates complementarily and simultaneously with existing sector-specific legislation, notably the Medical Devices Regulation (MDR, Regulation (EU) 2017/745). While the MDR addresses general risks associated with software as a medical device, the AI Act introduces specific rules to tackle issues inherent in AI, such as opacity, bias and lack of reliability (Aboy et al. 2024).

Although the AI Act does not directly regulate civil or criminal liability, it will inevitably influence the liability of healthcare professionals. For high-risk systems, compliance with requirements – accuracy, cybersecurity and human oversight – becomes a precondition for market entry. In judicial proceedings, these same regulatory standards will serve as benchmarks for defining the legal standard of care. Thus, the use of a non-compliant, outdated or insufficiently safeguarded AI system by a physician or healthcare provider may no longer be treated as a mere regulatory breach but as a violation of the expected standard of professional diligence. While not directly regulating civil or criminal liability, it contributes to establishing a new harmonized benchmark for the *lex artis* across Europe, transforming a product regulation into a foundational element of healthcare liability law.

At the core of this framework are obligations ensuring the retention of human control over automated decision-making (van Kolfschooten and van Oirschot 2024). Article 14 ("Human Oversight") requires that high-risk systems be "effectively overseen by natural persons during the period in which they are in use". This principle also serves a preventive function against automation bias, the human tendency to rely uncritically on machine outputs. To safeguard the physician's final judgment, essential to the attribution of professional liability, the Regulation mandates that clinicians understand system functionality, be aware of its limits and monitor its operation, with the capacity to intervene, disregard or discontinue its use.

This European approach is mirrored in national legislation, such as Italy's AI Bill (Bill No. 1146/2024, definitively approved on 18 September 2025), which reiterates that final clinical decisions must always rest with the physician. Yet the notion of "effective oversight" remains legally indeterminate and will inevitably be interpreted through the prism of national judicial traditions. What a German court might regard as adequate supervision, measured against the strict *fachärztlicher Standard*, may differ substantially from a Spanish court's assessment under the more flexible *lex artis ad hoc*. This creates the paradox of "fragmented harmonization": a single European rule may yield divergent liability outcomes across Member States, undermining legal certainty for suppliers and users in the single market.

Another critical issue concerns the evidentiary burden. Recognizing that algorithmic opacity – the so-called black box effect – creates near-insurmountable hurdles for patients seeking redress (Panattoni 2021), the EU has revised the defective product liability regime. Directive (EU) 2024/2853 now explicitly extends its scope to software and AI systems, designating them as "products". Its most significant innovation is the introduction of a "presumption of causality". This alleviates the burden of proof for injured parties: where a claimant can show that the AI provider failed to comply with a duty of diligence (e.g., a requirement under the AI Act), that the product was defective, and that a reasonable probability of causation exists, courts may presume a causal link between the breach and the harm. Access to technical information held by providers is also enhanced through specific disclosure obligations (Fragasso 2024). The burden then shifts to the supplier to rebut the presumption, for example by proving that the harm was caused by an unrelated factor.

This reform opens a new litigation pathway for patients. Instead of bringing a negligence claim against the physician, grounded in the assessment of professional conduct, patients may proceed directly against the AI supplier under defective product liability rules. This path offers advantages to claimants by leveraging the presumption of causality and circumventing the difficulty of deciphering a black-box system. Nonetheless, uncertainty persists: the regime imposes significant burdens on producers, who must demonstrate the absence of defects or causal links in complex technical environments. An increase in claims against AI developers is likely, which in some cases may "shield" physicians but also generate intricate scenarios of joint liability among clinicians, healthcare institutions and technology providers.

Comparative analysis of national systems: codified standards versus case-law principles

While the European regulatory framework establishes harmonized rules for technology, the assessment of physicians' conduct remains firmly rooted in national legal traditions. A comparative analysis of Italy, Germany, France, and Spain reveals significantly different approaches to defining the standard of care and allocating liability (Masieri 2024; Lanzara 2019).

In Italy, the legal framework reflects a persistent tension between legislative efforts to standardize the evaluation of medical fault (Massi 2024) and judicial protection of professional autonomy. The relevant provisions are Articles 589 and 590 of the Penal Code, dealing with negligent homicide and personal injury, complemented by Article 590-sexies, introduced by Law No. 24/2017 (the "Gelli-Bianco Law"), which marked a turning point in medical liability (Carraro 2022; Forti et al. 2010). This provision emphasizes the need for healthcare professionals to comply with recommendations contained in guidelines and good clinical practices, providing, in paragraph 2, a ground for exemption from liability for errors due to inexperience, where guidelines appropriate to the specific case have been followed (Basile and Poli 2022).

Medical performance is generally regarded as an obligation of means, and liability for its breach is assessed according to fault, measured against professional diligence (Art. 1176(2) of the Civil Code) and compliance with guidelines. However, the Supreme Court has interpreted the provision so as to safeguard the primacy of clinical judgment: the lex artis is not reduced to the mere application of protocols, but refers to the physician's ability to adapt them and, if necessary, to deviate from them with justification, based on the characteristics of the individual patient (see, inter alia, Supreme Court, Criminal United Sections, 21 December 2017, No. 8770). In this perspective, guidelines and good clinical practices become flexible instruments that guide, without replacing, professional judgment, ensuring uniformity and certainty in the evaluation of the physician's conduct, while leaving room for the appreciation of the specificities of the individual case and for personalized therapeutic choices (De Francesco 2021).

Consequently, compliance with guidelines does not automatically exclude liability, since the criminal judgment retains a subjective evaluative dimension that takes into account the specificity of the conduct and its context (Di Landro 2012). Where necessary, the physician is required to move beyond the boundaries of standard protocols, embracing the personal history, lifestyle, drug response, and wishes of the patient who has entrusted their health to them.

Compared with the previous legislation (Decree-Law No. 158/2012, converted into Law No. 189/2012, also known as the Balduzzi Law), the new provision introduced two significant changes: it eliminated the distinction between gross and slight fault, focusing on lack of expertise as the only form of fault relevant for the exclusion of punishment, and it conditioned exoneration on the conformity of guidelines to the specificity of the clinical case. The Supreme

Court (Criminal United Sections, 21 December 2017, No. 8770) nonetheless offered a more nuanced interpretation, in some respects reintroducing the distinction between gross and slight negligence and clarifying that criminal liability may arise even in cases of gross inexperience in the execution of appropriate recommendations. Thus, formal adherence to guidelines does not automatically guarantee exoneration from liability, shifting the focus instead to the physician's ability to apply the lex artis flexibly and contextually. The Court identifies, as criteria for evaluating the degree of fault, the divergence between actual and expected conduct, the foreseeability and avoidability of the event, the physician's knowledge and conditions and the reasons for urgency. In this perspective, liability depends on the intrinsic nature of the error, and mere adherence to a recommendation (whether from guidelines or, as later discussed, from algorithms) does not in itself exclude fault, which must be assessed in terms of its significance, recognizability, and inevitability.

The German system grounds medical liability in civil law, particularly in the general clause of tort under § 823 of the Civil Code (BGB). The standard of care is objective and uniform, identified as the *fachärztlicher Standard*, meaning the diligence required of a conscientious and attentive physician, consistent with the scientific knowledge consolidated at the time of treatment (Frahm 2020). Although not legally binding, medical guidelines (*Leitlinien*) play a significant evidentiary role, as they represent the state of science at a given moment and can in any case provide valuable indications for assessing liability (Wienke et al. 2020).

Case law, including that of the Federal Court of Justice (*Bundesgerichtshof – BGH*), has clarified that guidelines cannot automatically be equated with the standard of diligence required to assess medical error, nor can they substitute for expert opinion (*Sachverständigengutachten*). Nonetheless, they may acquire evidentiary value (*Indizwirkung*), as they reflect available scientific knowledge: an unjustified deviation from high-quality guidelines, such as S3-Leitlinien, may constitute a strong indicator of fault (Wienke 2008). In general, under the German system, the burden of proof lies with the patient; however, pursuant to § 630h of the Civil Code, the burden of proof (*Beweislastumkehr*) may be reversed in cases of gross medical error (*grober Behandlungsfehler*) or deficiencies in clinical documentation, where a causal link between the error and the damage is presumed.

This framework was later integrated by the pioneering *Digitale-Versorgung-Gesetz* (DVG) of 2019, which institutionalized the integration of digital technologies into the healthcare system (Sauermann 2021). The DVG created a fast-track pathway for the approval and reimbursement of digital health applications (*Digitale Gesundheitsanwendungen – DiGA*) by the Federal Institute for Drugs and Medical Devices (BfArM). This process not only ensures reimbursement but also establishes a system of digital tools evaluated and approved by the State. Consequently, if a DiGA is prescribed and widely used for a specific pathology, it may, over time, be considered by law as part of the standard of care for that condition. Should a physician choose not to use such a tool or to ignore its indications without valid clinical justification, a court may interpret this conduct as a deviation from the standard of care, giving rise to a new form of liability based on a "digital standard of care".

The French legal system is characterized by a dual approach, introduced by Law No. 2002-303 of 4 March 2002, known as the *Loi Kouchner* (Anzanello 2017). This system combines the traditional fault-based liability regime with a no-fault compensation mechanism designed to address serious, non-negligent medical accidents. On the one hand, the fault-based regime (*faute*) remains in force, requiring patients to prove an error by the physician or healthcare institution. The standard of care is defined by the *données acquises de la science* (established scientific knowledge), and the clinical practice recommendations (*recommandations de bonne pratique*) issued by the Haute Autorité de Santé (HAS) serve as an authoritative reference, though they are not strictly binding. On the other hand, in cases of serious harm resulting from a medical accident not attributable to fault (the so-called *aléa thérapeutique*), the law provides compensation funded by *solidarité nationale* and managed by the National Office for the Compensation of Medical Accidents (ONIAM). The no-fault scheme applies when the damage exceeds a certain severity threshold and is not the result of professional error but rather the inevitable occurrence of therapeutic risk. Access to this mechanism is guaranteed, in particular, in cases of permanent disability exceeding 24% (as set by decree) or temporary disability lasting at least six consecutive months. These requirements make access to ONIAM compensation relatively selective, as they require both the absence of fault and the presence of serious harm (over 24% disability or equivalent temporary incapacity) (Scarchillo 2017).

Notably, this dual track offers a unique solution to the black-box problem in AI. Where a patient suffers serious and unexpected harm following a procedure assisted by an opaque algorithm, compensation may be granted by ONIAM even if it is impossible to prove a specific fault of the physician or a technical defect in the AI. In such cases, the French system prioritizes victim compensation for severe and unforeseeable harm, decoupling redress from the complex and sometimes impossible attribution of fault. In this way, the creation of a national solidarity fund bridges the responsibility gap generated by technological opacity, introducing a form of mandatory insured liability.

The Spanish system is characterized by the centrality of the concept of *lex artis ad hoc*, a principle of jurisprudential origin that defines the standard of care (Vázquez López 2010). Unlike the German *fachärztlicher Standard*, the *lex artis ad hoc* is not an objective, predefined standard but rather a flexible criterion determined by the judge on a case-by-case basis, taking into account all the specific circumstances of the treatment: the physician's specialization, the complexity of the case, available resources, and the state of medical science at the time. It is an obligation of means, not of result, and its assessment is entrusted to judicial discretion, supported by expert evidence.

Clinical practice guidelines (*Guías de Práctica Clínica – GPC*), while important tools for guiding medical practice, are not considered by the case law of the Tribunal Supremo to constitute the *lex artis* itself. They serve as an authoritative reference, an element used by judges and experts to give substance to the *lex artis* in the specific case, but they neither exhaust nor bind it. Physicians may – and must – depart from them if the patient's circumstances so require (Feliu 2022). This approach, based on a fluid and judicially defined standard, provides maximum flexibility to adapt to technological innovation. A Spanish judge

may weigh the novelty of an AI system, the physician's experience with it, and the specific clinical context of the patient in determining whether the conduct was diligent. However, this flexibility comes at the cost of legal certainty. Unlike the more structured German approach, the outcome of litigation concerning AI-related medical liability in Spain will depend heavily on the individual court and appointed experts, leading to less predictable case law. This uncertainty may hinder the uniform and safe adoption of new technologies, making judicial discretion a double-edged sword.

Artificial intelligence as an emerging stakeholder in clinical decision-making.

Once the legal perimeter has been briefly outlined, attention must turn to the new protagonist entering the clinical stage: artificial intelligence. This is not merely a technological upgrade, but the introduction of a form of non-human "reasoning" at the very heart of medical practice. Yet what is the true nature of this silent consultant? Should it be regarded as nothing more than a sharper scalpel or a more powerful diagnostic tool, or does it represent an ontologically distinct category? Does the opinion it provides speak the same language as evidence-based medicine? or does it confront us with an alien logic, grounded in inscrutable statistical correlations? And what becomes of the physician's liability when this powerful digital oracle delivers its verdicts from behind the impenetrable veil of a "black box"? Analyzing these dimensions is a necessary step in uncovering the profound implications that AI projects onto the future of medical fault.

(a) Functional distinction: from "Technical Aid" to "Intellectual Aid"

The expression "artificial intelligence" evokes a profound tension between the imitation of human faculties and their transcendence. In medicine, such technologies are predominantly manifested in the form of weak AI: devices designed to perform well-defined functions, with high performance but without generalizing capacity. In these cases, the physician's role remains essential. The auxiliary model, however, to become bifurcated. To fully analyze the impact of artificial intelligence on professional liability in healthcare, it is useful to draw a functional distinction between AI as a technical aid and AI as a cognitive aid.

The first operates as a tool designed to enhance the physician's executive capabilities, functioning as a prosthesis of perceptual or operational skills, for example, a surgical robot or post-processing software for radiological images. In such instances, the algorithm does not intervene in the clinical decision-making process but merely amplifies human action. Responsibility, therefore, continues to follow traditional pathways: the physician is liable for improper use of the instrument, while the manufacturer is liable for defects in design or performance.

Far more problematic, however, is AI as an intellectual aid, where the system intervenes directly in diagnostic, prognostic, or therapeutic assessment, offering physicians solutions

generated by predictive models or classifiers. In this context, the medical act no longer represents the exclusive domain of clinical rationality, but the outcome of an interaction – sometimes symbiotic, sometimes conflictual – with an artificial entity. The physician does not merely "use" AI, but must engage with outputs that may appear opaque. He or she is then required to justify adherence to, or dissent from, algorithmic recommendations, bearing the weight of the final decision while facing increasing difficulty in understanding its logical foundations (Amore et al. 2024).

This complexity becomes even more evident in continuous learning systems, which evolve unpredictably, often without being able to explain how conclusions are reached. This phenomenon, known as the black box effect, compromises the transparency of the clinical process, undermining the ability of both physician and patient to reconstruct the logical sequence leading to a therapeutic recommendation. The algorithm thus escapes not only control but also comprehension. And since liability requires the explainability of actions, a doctrinal and regulatory void emerges – the so-called responsibility gap – which destabilizes the classical structure of medical fault (Matthias 2004; Giubilini 2025).

In such scenarios, AI does not merely support clinical activity; it risks altering its very essence. The physician becomes an executor of externally computed decisions, losing the ability to tailor care to the singularity of human experience. The very notion of negligence falters suspended between what can be calculated and what resists rationalization. Medical knowledge – traditionally intertwined with experience, intuition, and empathy – is now compelled to interact with an artificial knowledge, non-experiential, non-empathetic, but extraordinarily performant. This raises profound legal questions: to what extent is it legitimate – or appropriate – to entrust therapeutic choice to an algorithm? What legal weight should be attributed to AI-generated recommendations? Should they, or could they, become a new standard in the assessment of medical expertise?

Law once again finds itself chasing technology. Yet in doing so, it cannot relinquish its highest function: preserving responsibility as a human space of choice, judgment, and understanding.

(b) The epistemological difference between AI output and traditional guidelines

An AI recommendation is epistemologically distinct from a guideline. Guidelines distill human scientific knowledge, grounded in biological causality, validated through systematic review of the literature (evidence-based), and articulated in a symbolic and comprehensible language (Caputo 2012; Valbonesi 2013). They operate deductively: from general principle to indications for patient classes.

By contrast, AI output (particularly in deep learning) is the result of large-scale inductive processes. The algorithm does not "understand" biology but learns statistical correlations from millions of data points (data-driven), often operating on sub-symbolic representations (numerical vectors in multidimensional spaces) that are incomprehensible to humans (Peluso

2021). Its recommendation is not a general rule but a specific, probabilistic conclusion for the individual case.

This distinction is fundamental: the guideline offers a rational and comprehensible orientation; AI provides an expert opinion of alien nature, powerful yet whose logical foundation is often inaccessible.

(c) The black box problem: opacity, brittleness, and the new horizon of clinical risk

The "black box" problem extends beyond mere lack of transparency. Opacity prevents physicians from assessing the robustness and limits of the algorithm. A model may achieve extremely high average accuracy but fail in unpredictable and catastrophic ways in "out-of-distribution" cases (patients with rare diseases, anomalous data, or populations underrepresented in the training set). This brittleness introduces a new dimension of clinical risk. Physicians, unable to understand why AI suggests a given output, are equally unable to anticipate when or how it might err, leaving them to manage a risk they cannot measure.

A new dimension of clinical risk thus emerges, one that transcends established categories of human error and organizational dysfunction, traditionally embedded in the paradigm of clinical risk management and characteristic of high-reliability contexts (Caputo 2020). It is a dark, invisible, imperceptible risk, eluding conventional tools of monitoring and mitigation. Consequently, transparency can no longer be regarded as a merely technical or functional requirement; it must be considered an indispensable precondition for the implementation and governance of AI-based systems.

Transparency as an Essential Requirement: Explainable Artificial Intelligence (XAI)

Faced with the intrinsic opacity of the "black box models", the medical and legal communities are calling for a non-negotiable requirement: transparency. To transform the algorithm from an inscrutable oracle into a reliable partner, it is necessary to pierce the veil of opacity and give it a voice. Yet is it truly possible to teach a machine to "explain" its decisions in a language comprehensible to humans? How might this new capacity for dialogue restore full control of the decision-making process to the physician and, at the same time, provide more solid tools for legal defense? And, ultimately, is explainability always a guarantee of truth, or does it also conceal pitfalls by creating an illusion of understanding? Exploring explainable artificial intelligence (XAI) is not a mere technical exercise, but rather the fundamental step toward building an alliance between humans and machines grounded in trust and accountability (Mienye et al. 2024).

(a) Techniques of XAI in the Medical Domain

The aim of explainable artificial intelligence (XAI) is to render the internal mechanisms of decision-making models comprehensible and thus transparent, and to make the results they produce justifiable. Current research efforts are directed toward two categories of "explainability": *ex post* techniques and *ex ante* techniques.

The first category encompasses all methods whereby, after an AI output (e.g., diagnosis of a tumor or prediction of a response to chemotherapy), the user can trace how the input is mathematically linked to the output through an interpretive function. Among the most established approaches in this category are LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al. 2016) and SHAP (SHapley Additive exPlanations) (Lundberg and Lee 2017).

Specifically, LIME explains a single prediction (e.g., on a radiograph or medical report) by making small variations to the original input (such as altering the image or modifying clinical values), obtaining the black-box model's responses to these perturbed data, and training a simple model – such as linear regression – on them. The weights of this interpretable model provide a local explanation of the original decision.

SHAP, by contrast, is based on Shapley value theory and evaluates the marginal contribution of each decision variable by comparing the output of the full model with that obtained when features are excluded or combined systematically. The result is a fair and mathematically grounded distribution of weights assigned to the features, i.e., a quantification of how much each variable contributed to the outcome. For instance, in a predictive model of cardiovascular risk, SHAP may reveal that a patient's age increased the likelihood of disease by 12%, while physical activity reduced it by 7% (Sun et al. 2025).

Despite the transparency and explainability ensured by these techniques, they remain insufficient in contexts such as medicine, since they often lack interpretability, defined as the ease with which a human can understand the decision-making process. It is not uncommon, for example, that the interpretable surrogate models (e.g., linear regression) fail to correspond mathematically to the complexity of the actual model (e.g., numerical variables that lack clinical meaning). Another limitation lies in their restricted generalizability: LIME models are sensitive to the initial sampling, while SHAP models rely on associations rather than causality, meaning that statistically relevant variables may not necessarily be clinically significant.

As a practical example, we train a MobileNetV2 (Gulzar Y. 2023) with two added Dense layers (256 and 64), two Dropout layers (0.2 and 0.1), softmax activation, and GlobalAveragePooling, for the recognition of meningioma tumors from MRI images, followed by the application of LIME and SHAP respectively. The experiment used the public "BRAIN-Tumor" dataset available on Kaggle (https://www.kaggle.com/datasets/deeppythonist/brain-tumor-mri-dataset/data?select=test). Moreover, the model was trained using radiomic features automatically extracted from the images (including shape, volumetric, and texture characteristics, commonly referred to as Radiomic Features). Results are shown in Figure I illustrating the impact of these features on tumor diagnosis.

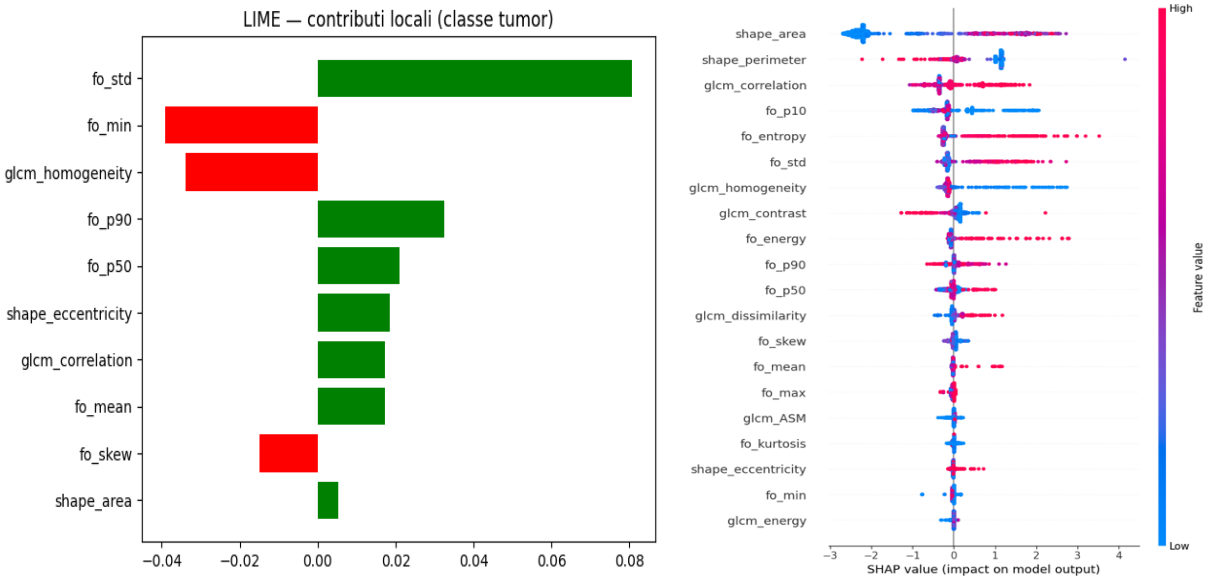

Fig. I an example of explanation obtained by LIME (left) and SHAP (right)

Another approach to explainability that is gaining attraction in the medical domain is the use of heatmaps generated by the Gradient-weighted Class Activation Mapping (Grad-CAM) technique. The core idea of this method is to exploit the gradients flowing into the last layer of a neural networks to produce a color map that highlights the regions (in terms of pixels) that contribute most to the model's output (Selvaraju et al. 2017). Unlike the previous methods, which require a certain level of expertise to interpret the results, these techniques are widely used thanks to their highly intuitive visual approach. On the other hand, they remain association-based rather than causal, and they are not model-agnostic (as SHAP and LIME are), but can only be applied to convolutional neural networks.

The images in Figure II illustrate a Grad-CAM heatmap generated on the same CNN model and the same dataset. On the left is the image used for glioma diagnosis, and on the right the heatmap produced by Grad-CAM following a positive prediction.

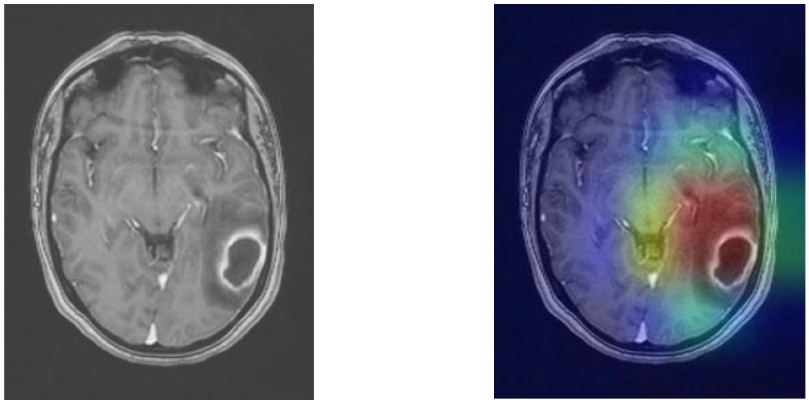

Fig. II an example of explanation obtained by GRAD-CAM (right) over a diagnosis of glioma (left)

Of a different nature are *ex ante* methods. Here, the designer's effort is to ensure explainability *by design*, i.e., intrinsically embedded in the model architecture from the algorithm's conception. This is the case with rule-based systems and decision trees, whose logical structure allows one to follow the system's reasoning step by step.

Such models are particularly suited to the clinical context, as their decision sequence can be built on diagnostic rules similar to those applied by healthcare professionals.

Practical applications include automated systems for the early diagnosis of type 2 diabetes, built on established clinical criteria, or triage systems in prevention centers. The advantage of these models is twofold: on the one hand, the transparency and reproducibility of the decision-making process; on the other, the ability to communicate in a comprehensible manner to both physician and patient.

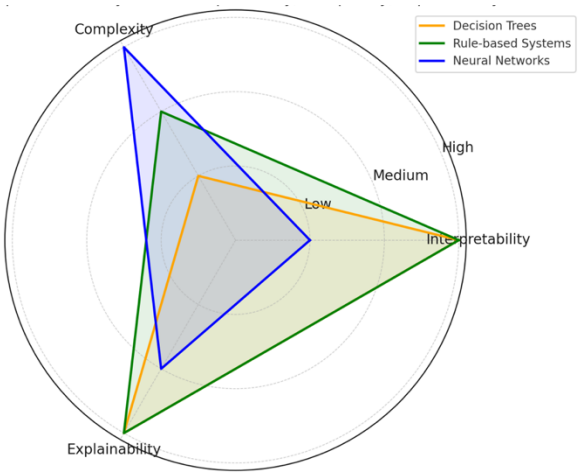

Fig. III Comparison of ex-post vs ex-ante methods based on complexity, explainability and interpretability

(b) Explainability versus Accuracy

One may ask why these explainable and easily interpretable models are increasingly marginalized in the field of healthcare innovation. The main reason lies in the trade-off between explainability and accuracy.

Empirical evidence demonstrates that the level of system complexity tends to progress in parallel with accuracy, at the expense of explainability, as schematized in the figure. Indeed, black-box systems, being trained on large amounts of data and considering highly varied scenarios, exploit the ability of deep networks to capture complex and non-linear relationships that are often imperceptible even to experts (e.g., volumetric features extracted from 2D images that cannot be detected by the naked eye).

In light of this, the challenge for artificial intelligence in medicine is to govern the inferential system as much as possible by introducing explainable steps within black-box architectures. The most promising examples include the integration of *Attention mechanisms*,

which highlight the portions of the input on which the model focuses when making decisions; *Activation techniques*, which visualize the inputs that maximize the response of a specific neuron; and *Weight visualization*, which helps to understand input–output relationships within the neural network.

(c) The Role of the Human User: Towards Human-in-the-Loop Models

An increasingly relevant approach in the context of AI is Human-in-the-Loop, in which human judgment is integrated into the AI decision-making process (Mosqueira-Rey et al. 2023). This paradigm, characterized by its interactive and collaborative nature, aims not only to improve system accuracy and fairness but also to generate outcomes that are more transparent, interpretable, and acceptable, particularly in healthcare, where trust, responsibility, and explainability are indispensable.

The first concrete results obtained in this field concern, in particular, the contribution of humans in the data-labeling phase, known as Active Learning. In oncological image segmentation systems, or in Interactive Machine Learning (IML), interaction with the human user is more frequent, incremental, and focused: it does not merely consist of providing labels on request, but involves continuous and active participation, such as offering suggestions, correcting errors, and guiding the model's evolution.

The current challenge is one in which the human expert designs the model's learning experience, determining which knowledge should be transferred and with which examples. This challenge anticipates new scenarios of responsibility in the use of AI.

Human–Machine Interaction Scenarios in a Comparative Perspective

Once the technological contours of artificial intelligence have been defined, the analysis must now turn to its most concrete and delicate dimension: criminal liability for fault. When physician and algorithm "decide" together, who is responsible if the choice proves fatal? A complex "dance" of liability is set in motion, where each step – trust, doubt, dissent – acquires unprecedented legal weight. Can a software recommendation produced by private entities provide the same legal "shield" as a public guideline? What happens when doctor and machine converge on the same error, and conversely, when the clinician, relying on instinct, disregards a correct AI insight? And how does evidentiary litigation change in court when the prosecution can wield the "digital counterproof" of a correct alternative that was available in real time? Exploring these scenarios means mapping the new frontiers of criminal risk in 21st-century medicine, adopting a legal assessment that is intrinsically tied to the national legal system of reference.

(a) The Non-Equivalence of Algorithmic Recommendations and Established Clinical Guidelines

We have already highlighted the sharp epistemological distinction between traditional guidelines and AI-generated outputs. It is therefore unsurprising that algorithmic recommendations – despite their promise of high predictive accuracy – cannot at present be considered either legally or epistemologically equivalent to guidelines. The lex artis is based on tools that physicians can understand, critically assess, and responsibly adapt to the individual case. Algorithmic recommendations, by contrast, are often intrinsically opaque, unverifiable, and structurally vulnerable to distortions caused by training data biases, heterogeneity of sources, and lack of model transparency.

This is not merely a practical limitation: the so-called black box effect epitomizes a profound epistemic barrier, breaking the chain of rational justification and depriving physicians of the ability to motivate – clinically, scientifically, and legally – the adoption or rejection of algorithmic advice. An algorithm's output therefore cannot be assimilated to a guideline. It lacks formal requirements, since it evades validation and publication mechanisms, and substantive ones, as it represents a specific, probabilistic conclusion rather than a general evidence-based recommendation. Moreover, guidelines result from a public and transparent process, while algorithms are often proprietary, shielded by industrial secrecy, intrinsically opaque and resistant to independent scrutiny.

Elevating such outputs to standards of diligence would impose on professionals a duty of conformity without epistemic foundation, shifting liability onto them while depriving them of intelligibility (Salvadori 2021). But liability in medicine must never be blind: it is an exercise of cognitive and decision-making autonomy that presupposes the ability to understand, evaluate, and justify the content of the decision. Stripping the physician of this faculty would reduce them to a mere executor of automated choices, sacrificing the ethical core of clinical practice on the altar of computational efficiency.

(b) The "Erroneous-Conformity" Scenario: When Doctor and Machine Agree in Error

We now turn to liability scenarios that may arise from human-machine interaction. Despite the potential of AI, not even the most advanced systems can guarantee flawless results. Two mirror-image figures of professional negligence emerge, both marked by structural ambiguity: concordant error and proactive dissent. Focusing on the former, this scenario arises when the physician embraces an AI recommendation, making it their own and sharing in its erroneous outcome (Amore et al. 2024).

In Italian law, the physician's liability is not automatically excluded by the mere fact of having followed an algorithmic suggestion, even if it is validated, widespread, and apparently reliable. Uncritical adherence to an algorithmic output – especially when expressed opaquely – risks becoming a form of negligence where critical professional scrutiny is absent. From an evidentiary standpoint, the difficulty lies in assessing the degree of transparency and comprehensibility of the algorithmic decision, often hidden within proprietary and inaccessible logics.

If the error were intelligible and recognizable, the physician could not shield themselves behind the machine's technicality. If, instead, the output was structurally opaque and no clinical alerts existed, the error might be deemed unavoidable and liability consequently excluded. Where the distortion stems from an intrinsic system flaw (such as training data bias or systemic diagnostic error), liability might shift to the software producer or, potentially, the healthcare institution for *culpa in eligendo* (Fragasso 2024; Salito 2022).

Ultimately, the law cannot demand superhuman performance from physicians, but neither does it permit them to abdicate their decision-making autonomy in favor of presumed technical infallibility. The machine is never neutral: it embeds the choices, biases, and approximations of its designers, offering knowledge that is always modeled, never an objective truth. In this context, an algorithm that confirms an error does not absolve the physician of responsibility but compels them to justify their adherence.

The nature of the error will be decisive. If it is subtle, complex to detect, and statistically rare, concurrence between doctor and AI may excuse the conduct. But if the error derives from a known system limitation (e.g., a systemic bias against certain patient groups) and the physician failed to assess the system's reliability, adherence may constitute culpable ignorance or excessive delegation. Shared error does not eliminate liability but reframes it, prompting reflection on the degree of human judgment autonomy in digitally assisted environments and the extent to which medical expertise must remain vigilant against the seduction of the machine (Mosqueira-Rey et al. 2023).

Comparatively, different legal systems approach "erroneous conformity" differently. In Germany, courts may assess whether adherence to algorithmic output deviates from the *fachärztlichen Standard*: a modest, non-detectable error is unlikely to constitute a *grober Behandlungsfehler* (gross error), though minor fault may exist if additional checks required by the standard were omitted. In Spain, the analysis would turn on the lex artis ad hoc: reliance on a certified system could be reasonable unless clinical warning signs that a diligent physician should have recognized were ignored. In France, the dual system applies: for serious harm, patients may claim no-fault compensation through ONIAM as an *aléa thérapeutique*, while fault-based liability would require proving negligence in uncritically accepting algorithmic advice. At the supranational level, the EU Regulation 2024/1689 (AI Act) imposes strict obligations for high-risk AI systems, including effective human oversight throughout the system's lifecycle. This requirement – part of a broader framework of transparency, reliability, and ethics – seeks to ensure that automated decisions never escape human control, but remain subject to continuous monitoring proportionate to risk, enabling intervention to prevent or correct harmful or discriminatory effects.

(c) The "Erroneous-Divergence" Scenario: When the Doctor Disagrees with a Correct AI Recommendation

Equally problematic is the hypothesis of proactive dissent (Amore et al. 2024). Here, the physician consciously chooses not to follow an algorithmic recommendation that proves ex post correct, basing the decision on parameters beyond the machine's scope (e.g., complex

patient history, recent data, or extra-clinical elements). Dissent is not an omission but a positive medical act requiring strengthened justification. Liability does not stem from "disobeying" the machine, but from potentially depriving the patient of a correct diagnostic or therapeutic opportunity.

To defend themselves, physicians must show that they had strong clinical reasons to deem the output unreliable and, crucially, that they pursued alternative actions to rule out the AI's suggestion. Dissent should mark the beginning of further clinical inquiry, not its end (Verdicchio and Perin 2022). Legal assessment thus focuses on the rationality of divergence: if adequately justified and based on scientific evidence or clinical context, the conduct may fall within legitimate professional discretion. If, however, dissent is arbitrary, reckless, or unjustified, it may ground liability.

A paradoxical risk arises whereby fear of negative consequences compels physicians to conform passively to algorithmic recommendations even when they wish to depart from them. Documentation of algorithmic suggestions becomes vital: a transparent and complete record would provide indispensable evidence to assess (and potentially excuse) the margin of human error.

In comparative perspective, solutions vary by legal system. In Germany, physicians may need to demonstrate that their dissent aligned with the *fachärztlichen Standard*, supported by detailed clinical records; absent this, divergence risks being seen as diagnostic or therapeutic error. In Spain, courts may emphasize the plausibility and rationality of clinical justification, verifying whether the *lex artis ad hoc* required additional scrutiny: discretion would be recognized if exercised with reasonableness and traceability. In France, motivated dissent grounded in case-specific factors could be valued as an expression of professional freedom, though patients suffering severe harm from non-fault adverse events might still access compensation through the national solidarity fund, provided statutory thresholds are met.

(d) Algorithmic Evidence in Medicine and the Transnational Implications of Digital Counterproof

The output of an AI system, recorded and timestamped, becomes a form of "digital counterproof" that reshapes evidentiary dynamics. In traditional litigation, evaluation of conduct involves an *ex post* comparison between the physician's decision and a reconstructed standard of care. With clinical AI systems, however, the paradigm shifts: the algorithm produces not only support but a concrete, documented alternative, available hic et nunc at the very moment the physician makes their decision.

Consequently, evaluation of medical conduct is no longer based solely on a reconstructed standard in court, but on a certified and consultable alternative. The burden of argument thus shifts: physicians will no longer suffice by showing that their conduct was reasonable against an abstract standard; they must also justify why they rejected an algorithmic recommendation that later proved correct. Discretion is not abolished but constrained: divergence requires explicit justification, documented awareness, and reinforced adherence to scientific method (Giannelli 2024).

AI output is thus not evidence, but a strengthened indicator of what the physician could (and perhaps should) have done. A silent yet incisive presence in litigation, it retrospectively rewrites the narrative of fault. The weight of such "digital counterproof" will likely vary by jurisdiction: in Germany, where clinical documentation is paramount, an output ignored without annotation could weigh decisively against the physician; in Spain, it would become an indispensable element in expert and judicial evaluation of the lex artis ad hoc. In all cases, however, the presence of digital counterproof shifts the argumentative burden, requiring physicians not only to demonstrate the reasonableness of their conduct against an abstract standard, but also to explain concretely why they disregarded a plausible, documented alternative offered hic et nunc by the algorithm.

Strategies for Clinical Governance of AI in Europe

Having explored the complex fault lines of liability, the analysis must now turn to building a system capable of governing innovation, harnessing the power of AI within a safe perimeter for both patient and physician. Safe and effective integration of AI into healthcare requires governance strategies that move beyond mere regulatory compliance, translating European principles into concrete clinical practices at national and local levels. The goal is to govern innovation without stifling it, and without creating new forms of risk for patients and professionals.

Should AI itself become the new guideline, or should guidelines instead be adapted to govern AI, defining its rules of engagement? What happens when validated AI use becomes the new standard of care, turning the choice not to employ it into potential fault? Yet defining such strategies risks generating a new paradox: a more rigid form of defensive medicine, where fear of contradicting the machine replaces clinical judgment, undermining the very aim of personalized care.

(a) Guidelines as a Tool to Govern AI Use

The most effective integration does not lie in equating AI with a guideline, but in ensuring that guidelines regulate AI use. Scientific societies should develop recommendations specifying: in which clinical scenarios validated AI is recommended; what standards of accuracy and robustness an algorithm must meet to be considered reliable; which software versions have been validated (to address model drift); and which protocols should govern documentation of interaction and dissent.

This approach already finds concrete examples in some European contexts: in Germany, the BfArM certifies DiGA, providing an evaluation mechanism for guideline development; in France, the HAS may issue recommendations integrating AI into good clinical practice; at the international level, the World Health Organization, in January 2024, published Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models (WHO 2024), containing over forty recommendations for ethically sustainable use of AI. Its Italian

version (Causio et al. 2024), enriched with an original chapter contextualizing international guidelines within the national regulatory framework, has yet to receive formal validation by the Ministry of Health or integration into SSN protocols.

This regulatory inertia, in the face of accelerating technological change, generates a normative and axiological vacuum in which uncertainties and liability conflicts proliferate. Without timely, methodologically sound updates of clinical guidelines, innovation risks exceeding the boundaries of legality, destabilizing the balance between technical standardization and decision-making autonomy. Today more than ever, it is necessary to reaffirm the centrality of medical judgment, not as subjective arbitrariness, but as applied rationality, capable of integrating technical innovation without surrendering to it. AI must participate in clinical decision-making as an active but not exclusive component of it, dialoguing with medical expertise rather than dictating its limits.

(b) Divergence from Validated AI Output as a Potential Violation of Leges Artis

When an AI system is rigorously validated and incorporated into guidelines as part of the standard of care for a specific activity, its use becomes part of the leges artis. Consequently, deciding not to use it or ignoring its alerts without a documented and overriding clinical justification may constitute a breach of good practice, with potential liability for fault defined as failure to apply the state of the art.

The introduction of AI as an integral component of clinical practice thus alters the very notion of professional diligence: physicians are required not only to know and apply traditional knowledge, but also to integrate and critically evaluate algorithmic recommendations. Ignoring or unjustifiably rejecting validated AI output could therefore compromise care quality and result in stricter liability, since in this context the algorithm represents an extension of codified medical science.

At the same time, it must be emphasized that the final clinical decision remains a human act, and deviation from AI output is justifiable only in the presence of well-founded, documented, and clinically relevant reasons. Medical assessment must continue to rest on integrated critical judgment, with AI serving as support, not an automatic constraint.

(c) The Paradox of "Algorithmic Defensive Medicine": A New Risk for Medical Professions

Evidentiary asymmetry and legal pressure may foster a new and pernicious form of defensive medicine. This phenomenon emerges from the convergence of evidentiary imbalances and legal constraints, inducing physicians to an excessively cautious posture aimed at minimizing exposure to litigation or disciplinary proceedings. Physicians may develop a tendency to adhere passively and uncritically to algorithmic suggestions, even when their independent clinical judgment points elsewhere.

The outcome would be a progressive flattening of clinical reasoning, with physicians reduced to executing algorithmic indications mechanically. This not only impoverishes the quality of personalized care but also triggers a process of deskilling: the erosion of medical competence, experience, and decision-making capacity (Oliva et al. 2022). The technology's original aim – promoting personalization, precision, and innovation – would thus be inverted, imposing a new conformism in which the algorithm becomes a new dogma, adhered to without scrutiny, at the expense of critical thinking and professional autonomy long protected by jurisprudence and medical ethics.

This paradox represents a central challenge for healthcare and law: the task is to strike a balance that values AI's contribution without subordinating physicians entirely, preserving the human, ethical, and scientific essence of clinical decision-making.

Conclusions

At the end of this journey, which has navigated courtrooms and artificial intelligence laboratories, it is time to take stock and look ahead. The analysis reveals a legal landscape in profound transformation, where technological innovation acts as a powerful reagent on diverse national legal cultures. The encounter between medical malpractice and binary code is no longer a futuristic scenario but a pressing reality. What paradigm, then, emerges from this complex interaction? And, more importantly, what course should legislators, healthcare institutions, and the scientific community chart to ensure that the promise of AI-augmented medicine translates into safe, ethical, and equitable progress? The following conclusions do not represent a definitive endpoint but rather an attempt to sketch a roadmap for governing one of the most profound revolutions in the history of care.

Synthesizing the Paradigm: Artificial Intelligence Enhancing Personalization Rather Than Replacing Clinical Judgment: Artificial intelligence should not be conceived as a new form of rigid guideline, but rather as a powerful instrument to realize the principle of personalized care demanded by both law and medical ethics. AI does not provide the general rule; it offers a formidable aid in tailoring that rule to the unique singularity of each patient. Within this vision, the machine does not replace the physician's judgment but strengthens it, delivering complex analyses that, if properly understood and governed, can lead to safer and more effective decisions. Responsibility, therefore, does not vanish nor migrate to the machine. Instead, it is reconfigured as responsibility for managing a complex, collaborative decision-making process.

Comparative analysis has shown that, while the AI Act harmonizes rules for placing technology on the market, liability for its clinical use will remain deeply shaped by the distinct legal traditions of Member States. Germany's objective and standardized approach, Spain's flexible and case-based jurisprudence and France's dual system of fault and solidarity will all interpret European principles – such as "human oversight" – through their own hermeneutical lenses. This heterogeneity is not necessarily a limitation; rather, it may serve as a laboratory of diverse legal solutions to a shared challenge.

Recommendations for Legislators, Healthcare Institutions, and Scientific Societies:

Governing this revolution requires coordinated, multi-level action that goes beyond individual national frameworks and embraces a European systems perspective.

At the legislative level, both European and national, while the AI Act and the Product Liability Directive provide a solid framework, national lawmakers should consider introducing coordinating provisions. In Italy, for example, the current rigidity of Article 590-sexies of the Penal Code could be revisited by introducing specific provisions on medical fault in the use of certified AI systems, centered on the "reasonableness" of the physician's overall conduct when interacting with the machine. Similar adjustments may be required in other Member States to clarify the interplay between new EU obligations and national liability regimes.

At the organizational level, healthcare institutions must ensure robust clinical governance of AI. This entails establishing clear protocols for the adoption, validation, and monitoring of algorithms, standardized procedures for handling and documenting reasoned dissent, and the creation of ethical and technical review committees. Equally crucial is investment in continuous training programs enabling staff to use technology critically and responsibly – never as an act of blind faith – in alignment with the AI Act's requirement for human oversight.

At the scientific and professional level, medical societies, in cooperation with national regulatory bodies such as Germany's BfArM or France's HAS, should play a leading role in developing independent frameworks for clinical validation and post-market surveillance of algorithms. Their responsibility includes updating clinical guidelines to incorporate specific recommendations on AI use and actively promoting research into explainable AI (XAI), ensuring that transparency becomes a non-negotiable requirement for all AI-based medical devices.

In summary, the goal is not to construct a fully uniform European law of medical liability an aim both unrealistic and not necessarily desirable. Rather, it is to outline a "*European model of AI-augmented medical liability*", that is a framework of shared principles (human oversight, transparency, reliability, explainability), rooted in EU law, that can be integrated and applied coherently within diverse national legal systems. Only through this dialogue between European harmonization and national traditions can the promise of AI-enhanced medicine be realized as safe, ethical, and equitable progress for all European citizens.

References

ABOY, M.; MINSSEN, T.; VAYENA, E. 2024. Navigating the EU AI Act: implications for regulated digital medical products. *NPJ digital medicine*, **7**(1):237. https://doi.org/10.1038/s41746-024-01232-3

AMORE, N.; CIONI, A.; CORTI, D.; LIPPI, ME.; SESTIERI, M.; VALLINI, A. 2024. Policy Paper. Buone pratiche e modelli di regolamentazione per l'impiego della IA nella diagnostica per immagini. *La legislazione penale*, 3-4:11-13

ANZANELLO, L. 2017. La responsabilità professionale sanitaria dall'Arrêt Mercier alla Loi Kouchner. *Assicurazioni*, 255-266

BASILE, F.; POLI, P. F. 2022. La responsabilità per "colpa medica" a cinque anni dalla legge Gelli-Bianco. *Sistema penale*, 1-34.
CAPUTO, M. 2012. "Filo d'Arianna" o "Flauto Magico"? Linee guida e checklist nel sistema della responsabilità per colpa medica. *Diritto penale contemporaneo*, 1-39.
_____. 2020. Prevenire è meglio. Uno sguardo interdisciplinare sull'organizzazione sanitaria quale fonte di rischi e garante della sicurezza delle cure. *Rivista italiana di medicina legale e del diritto in campo sanitario*, **4**:1955-1963
CARRARO, L. 2022. *Il medico dinanzi al diritto penale. Alla ricerca di limiti relazionali all'imputazione colposa*. Torino, Giappichelli Editore.
CAUSIO, F. A.; TALIO, A.; DRI, P. (eds.) 2024. *Etica e governance dell'intelligenza artificiale per la salute. Linee guida per i modelli multimodali di grandi dimensioni (LMM).* Società Italiana di Intelligenza Artificiale in Medicina (SIIAM) & Zadig srl Società Benefit. Milano, Zadig srl Società Benefit. https://www.zadig.it/wp-content/uploads/2024/06/LG-AI-IT-def_1.pdf
CRIMINAL UNITED SECTIONS. 2017. No. 8770. 21 December 2017. *Rassegna di diritto farmaceutico e della salute*, **3**:578 ff
DE FRANCESCO, G. 2021. In tema di colpa. Un breve giro d'orizzonte. *La legislazione penale*, 2
DI LANDRO, A. 2012. *Dalle linee guida e dai protocolli all'individualizzazione della colpa penale nel settore sanitario. Misura oggettiva e soggettiva della malpractice*. Torino, Giappichelli Editore.
FELIU, J. S. 2022. Estándar de diligencia médica y valor de los protocolos y guías de práctica clínica. *Revista de derecho civil*, **9**(3):1-52
FRAGASSO, B. 2024. Intelligenza artificiale e crisi del diritto penale d'evento: profili di responsabilità penale del produttore di sistemi di I.A. *Rivista italiana di diritto e procedura penale*, **1**:287-305
FRAHM, W. 2020. Der zivilrechtliche Facharztstandard. *In*: Jansen, C., Katzenmeier, C., Woopen, C. (eds), *Medizin und Standard. Schriften zu Gesundheit und Gesellschaft - Studies on Health and Society*. Vol. 3. Berlin, Heidelberg, Springer.
FORTI, G.; CATINO, M.; D'ALESSANDRO, F.; MAZZUCATO, C.; VARRASO, G. (eds). 2010. *Il problema della medicina difensiva. Una proposta di riforma in materia di responsabilità penale nell'ambito dell'attività sanitaria e gestione del contenzioso legato al rischio clinico.* Pisa, ETS.
GIANNELLI, A. 2024. Nuove declinazioni della decisione motivata. *Diritto amministrativo,* **4**:1020
GIUBILINI, A. 2025. It is not about AI, it's about humans. Responsibility gaps and medical AI. *Journal of bioethical inquiry*. https://doi.org/10.1007/s11673-025-10423-w
GULZAR, Y. 2023. Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique. *Sustainability*, **15**(3):1906. https://doi.org/10.3390/su15031906
JOHNSON, K. B.; WEI, W. Q.; WEERARATNE, D. *ET AL*. 2021. Precision medicine, AI, and the future of personalized health care. *Clinical and translational science*, **14**(1):86-93
LANZARA, O. 2019. *Medical malpractice: uno studio di diritto comparato*. Torino, Giappichelli Editore.
LUNDBERG, S. M.; LEE, S. I. 2017. A unified approach to interpreting model predictions. *Neural information processing systems*, **30**:4768-4777.
MANNION, R.; EXWORTHY, M. 2017 (Re)Making the Procrustean bed? Standardization and customization as competing logics in healthcare. *International journal of health policy and management*, **6**(6):301-304. https://doi.org/10.15171/ijhpm.2017.35
MASIERI, C. M. 2024. *Medical Malpractice Legislation. Reforms in civil law systems*. Routledge, London.

MASSI, S. 2024. Il problema della responsabilità penale del sanitario tra incertezze giurisprudenziali e insufficienze legislative. *Cassazione penale,* **64**(12):4093-4115
MATTHIAS, A. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, **6**:175-187. https://doi.org/10.1007/s10676-004-3422-1
MIENYE, I. D.; OBAIDO, G.; JERE, N.; MIENYE, E.; ARULEBA, K.; EMMANUEL, I. D.; OGBUOKIRI, B. 2024. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Informatics in medicine unlocked*, **51**:101587. https://doi.org/10.1016/j.imu.2024.101587
MOSQUEIRA-REY, E.; HERNÁNDEZ-PEREIRA, E.; ALONSO-RÍOS, D.; BOBES-BASCARÁN, J.; FERNÁNDEZ-LEAL, A. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, **56**:3005-3054. https://doi.org/10.1007/s10462-022-10246-w
NUTI, S.; PANERO, C. 2011. La sfida dei servizi in sanità tra personalizzazione e standardizzazione dei processi. *In*: AAVV. *Nuovi modelli di business e creazione di valore: la Scienza dei Servizi. Sxi – Springer per l'Innovazione / Sxi – Springer for Innovation.* Milano, Springer, 193-213. https://doi.org/10.1007/978-88-470-1845-7_9
OLIVA, A.; GRASSI, S.; VETRUGNO, G.; ROSSI, R.; DELLA MORTE, G.; PINCHI, V.; CAPUTO, M. 2022. Management of medico-legal risks in digital health era: a scoping review. *Frontiers in medicine,* **8**:821756. https://doi.org/10.3389/fmed.2021.821756
PANATTONI, B. 2021. Intelligenza artificiale: le sfide per il diritto penale nel passaggio dall'automazione tecnologica all'autonomia artificiale. *Diritto dell'informazione e dell'informatica*, **II**(2):317-368.
PELUSO, M. G. 2021. Data Driven Innovation in medicina, vantaggi e prospettive critiche. *Responsabilità medica*, **3**:225-248. Available at: http://hdl.handle.net/10446/200692
PFAFF, H.; DRILLER, E.; ERNSTMANN, N.; KARBACH, U.; KOWALSKI, C.; SCHEIBLER, F.; OMMEN, O. 2010. Standardization and individualization in care for the elderly: proactive behavior through individualized standardization. *Open Longevity Science*, **4**:51-57
RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). *Association for computing machinery*, 1135-1144. https://doi.org/10.1145/2939672.2939778
RUSCHEMEIER, H.; BAREIS, J. 2025. Searching for harmonised rules: understanding the paradigms, provisions and pressing issues in the final EU AI Act. *In*: Gsenger, R.; Sekwenz, M. T. (eds) *Digital Decade: How the EU shapes digitalisation research*. Nomos, Baden-Baden, 41-93
SALITO, G. 2022. La responsabilità da algoritmo tra (teoria della) finzione e realtà sanitaria: una nuova declinazione della responsabilità medica? *Rivista italiana di medicina legale (e del diritto in campo sanitario)*, **4**:849-863
SALVADORI, I. 2021. Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale. *Rivista italiana di diritto e procedura penale*, **1**:83–118
SAUERMANN, S.; HERZBERG, J.; BURKERT, S.; HABETHA, S. 2021. DiGA - A Chance for the German Healthcare System. *Journal of european CME*, **11**(1):2014047. https://doi.org/10.1080/21614083.2021.2014047
SCARCHILLO, G. 2017. La responsabilità medica: risarcimento o indennizzo? Riflessioni, evoluzioni e prospettive di diritto comparato. *Responsabilità civile e previdenza,* **82**(5):1490-1520
SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D., GRAD-CAM. 2017. Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE*

international conference on computer vision (ICCV), 618-626. https://doi.org/10.1109/ICCV.2017.74

SUN, Q.; AKMAN, A.; SCHULLER. B. W. 2025. Explainable Artificial Intelligence for Medical Applications: A Review. *ACM transactions on computing for healthcare*, **6**(2):17. https://doi.org/10.1145/3709367

UE. Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence (AI Act). https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=OJ:L_202401689

VALBONESI, C. 2013. Linee guida e protocolli per una nuova tipicità dell'illecito colposo. *Rivista italiana di diritto e procedura penale,* **56**(1):250-301.

VÁZQUEZ LÓPEZ, J. E. 2010. "La Lex Artis ad hoc" como criterio valorativo para calibrar la diligencia exigible en todo acto o tratamiento médico: A propósito de un caso basado en la elección de la técnica empleada en el parto (parto vaginal vs. cesárea). *Cuadernos de medicina forense,* **16**(3):179-182.

VERDICCHIO, M.; PERIN, A. 2022. When doctors and AI interact: on human responsibility for artificial risks. *Philosophy and technology*, **35**(1):11. https://doi.org/10.1007/s13347-022-00506-6

VAN KOLFSCHOOTEN, H.; VAN OIRSCHOT, J. 2024. The EU Artificial Intelligence Act (2024): Implications for healthcare. *Health policy*, **149**:105152. https://doi.org/10.1016/j.healthpol.2024.105152

WIENKE, A.; HÜBNER, L.; GAHN, G. 2020 Facharztstandard und Leitlinien im Arzthaftungsrecht [Specialist standards and guidelines in medical malpractice law]. *DGNeurologie*, **3**:565-567. https://doi.org/10.1007/s42451-020-00188-4

WIENKE, A. 2008. BGH: Leitlinien ersetzen kein Sachverständigengutachten. *GMS Mitt AWMF,* 5: Doc14.

WORLD HEALTH ORGANIZATION. 2024. *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. Geneva, World Health Organization. https://iris.who.int/server/api/core/bitstreams/e9e62c65-6045-481e-bd04-20e206bc5039/content

Submetido: 01/10/2025
Aceito: 07/04/2026