ABOUTOpen Artificial Intelligence

AboutOpen | 2026; 13: 47-53
ISSN 2465-2628 | DOI: 10.33393/ao.2026.3940

POINT OF VIEW

Open Access

Clinical artificial intelligence as a cognitive intervention: cognitive algovigilance, adverse events, and post-market surveillance

Daniele Angioni[1-3], **Fabio Di Bello**[4-6]

[1]General Practitioner, Novara, Piemonte - Italy; SIEMG-certified physician in clinical ultrasound, Italian Society of Ultrasound in General Practice and Clinical Ultrasound, Parma - Italy
[2]Member of the National Commission for Digital Innovation and Technology of AME with delegation for AI, Udine - Italy
[3]Council member of the Medical Order of Novara, Novara - Italy
[4]Senior Education & AI Development Manager, Wiley, Hoboken, NJ - USA
[5]Faculty Member, Master in Humanization of Care, University of Udine, Udine - Italy
[6]Faculty Member, University Bicocca Academy of Milan, Milan - Italy

ABSTRACT

Artificial intelligence (AI) safety in medicine is still evaluated largely through model-centered metrics: accuracy, robustness, bias, drift, transparency, and regulatory compliance. These dimensions remain essential, but they do not capture the full safety problem created when AI enters clinical reasoning. The literature has already described automation bias, overtrust, alert fatigue, trust miscalibration, deskilling, and failure modes in human–AI collaboration. The unresolved problem is that these effects are often treated as scattered human-factors concerns rather than as observable safety events. This Point of View proposes clinical AI as a cognitive intervention. Its primary safety unit is the human-machine cognitive system, not the isolated model. We define cognitive algovigilance as the systematic surveillance of cognitive effects arising from human–AI interaction, and Cognitive Adverse Events as measurable AI-mediated changes in clinical decision-making that increase the probability of suboptimal decisions, including when the algorithmic output is technically correct. We distinguish technical/software surveillance as an enabling layer from cognitive algovigilance proper and define cognitive red flags as pre-event signals, including those elicited by risk-triggered or randomly sampled pre-output reasoning probes. We identify AI-induced diagnostic narrowing as a sentinel phenotype and propose Cognitive Post-Market Surveillance as a complementary post-deployment layer for collecting cognitive near misses, comparing pre-AI and post-AI differentials, auditing overrides and high-risk non-overrides, and monitoring verification decay, cognitive dependence, and loss of the diagnostic tail. The aim is not to slow AI adoption, but to distinguish cognitive augmentation from cognitive toxicity and make clinical AI safety a verifiable cognitive discipline.

Keywords: Artificial intelligence, Decision support systems, Clinical, Diagnostic errors, Human–AI interaction, Patient safety, Post-market surveillance

Introduction: the blind spot of clinical safety

Clinical artificial intelligence (AI) is commonly assessed through categories that are necessary but incomplete: predictive performance, fairness, explainability, cybersecurity, privacy, external validation, professional liability, and regulatory compliance. Current scientific and regulatory frameworks already require human oversight, transparency, lifecycle risk management, post-market monitoring, and attention to the performance of the human–AI team (1-10). Yet the cognitive behavior of clinicians exposed to AI outputs remains less systematically observed than the software itself.

This does not mean that the cognitive dimension has been ignored. Empirical studies and reviews have described automation bias, acceptance of erroneous advice, human factors in clinical decision support, the modulation of trust by explainable AI, uncertainty communication, alert fatigue, and deskilling (11-20). Diagnostic error has also long been understood as the product of cognitive, systemic, and contextual factors (21-24). The gap is more precise: these risks are recognized, but they are rarely organized as a class of safety events that can be defined, measured, reported, and surveilled.

Clinical AI adds a new epistemic artifact to the diagnostic encounter. Recent work on LLM-mediated judgment has described epistemia: the illusion of knowledge that arises when surface plausibility substitutes for epistemic verification (25). This mechanism is directly relevant to clinical use.

Received: June 21, 2026
Accepted: July 15, 2026
Published online: July 29, 2026

Corresponding author:
Daniele Angioni
email: studiomedicoangioni@gmail.com

AI output does not merely provide information; it can redirect attention, compress uncertainty, modify the threshold for verification, alter trust, narrow the diagnostic differential, and weaken the willingness to dissent. The key safety question is therefore not only whether the AI answer is correct, but whether exposure to that answer changes how the clinician searches, doubts, verifies, and decides.

Positioning of the proposal

This manuscript does not claim that automation bias, overtrust, or deskilling are new. They are established risks in automation research, human–AI interaction, and clinical decision support (12-20,26,27). The proposed shift concerns the unit of analysis. Instead of treating these phenomena as isolated biases or individual psychological failures, we treat the coupled human–AI cognitive system as an object of patient safety.

Three distinctions are central. First, an adverse cognitive effect does not require the algorithm to be wrong. A correct output may still become unsafe if it induces premature closure, reduces independent verification, or makes rare but dangerous diagnoses disappear from active consideration. Second, the event is defined by a measurable change in the decision-making process, not only by final patient harm. Third, surveillance must include traces of interaction: the pre-AI and post-AI differential, quality of dissent, high-risk non-overrides, verification behavior, and persistence of autonomous clinical competence. The contribution is therefore an operational taxonomy, not another item in a list of biases.

Relative to existing literature, the framework makes four operational moves: it defines cognitive adverse events even when the AI output is technically correct; treats trust as a safety variable to be calibrated through verification and dissent; places dependency, verification decay, and deskilling within longitudinal surveillance; and links diagnostic narrowing and preservation of the diagnostic tail to a taxonomy suitable for consensus and validation.

Clinical AI as a cognitive intervention

A clinical AI system should be evaluated not only as a predictive or generative technology, but also as a cognitive intervention. Any tool capable of changing attention, working memory, trust, verification thresholds, diagnostic breadth, or willingness to dissent enters the domain of patient safety. The primary safety unit is consequently not the isolated model. It is the human-machine cognitive system in a concrete workflow.

This position does not replace model audit, external validation, or device surveillance. It extends them. A model may perform well on technical benchmarks and still produce an unsafe decisional effect in practice. Such risk is not located entirely in the software or entirely in the clinician. It emerges from the coupling of model output, interface, user expertise, time pressure, organizational culture, and clinical context.

Cognitive algovigilance and cognitive adverse events

By cognitive algovigilance we mean the systematic surveillance of the cognitive effects produced by interaction among an AI system, the clinical user, and the shared decision-making environment. The analogy with pharmacovigilance is useful but limited. Drugs mainly generate biological adverse events; clinical AI may generate cognitive adverse events, affecting how professionals reason, verify, calibrate trust, and act. Figure 1 depicts the overall surveillance loop, and Table 1 distinguishes its technical enabling and cognitive surveillance domains.

We define a Cognitive Adverse Event (CAE) as a measurable modification of the clinical decision-making process induced or facilitated by interaction with an AI system, when that modification increases the probability of a suboptimal decision, independently of the intrinsic accuracy of the algorithmic output. Five elements are required: exposure to an AI output; a documentable change in the decision process compared with pre-AI reasoning or a plausible baseline; clinical relevance of the change; systemic attribution to the human–AI interaction; and reversibility, near miss, potential harm, or actual harm.

The definition avoids two simplifications. It does not reduce every event to clinician error, and it does not attribute every deviation to the model. Attribution must include interface design, workflow, training, time pressure, local governance, and the cognitive affordances of the output itself.

Before a CAE is established, cognitive algovigilance may detect cognitive red flags: proximal, observable signals of increased cognitive risk that do not yet satisfy the event definition. Red flags may arise from routine interaction traces or be elicited by a software-embedded cognitive probe. Examples include absence of an independent pre-AI hypothesis; immediate adoption of a fluent output; unexplained contraction of the differential; loss of a must-not-miss diagnosis; omitted source verification; repeated high-risk non-overrides; repeated bypass of a manual reasoning prompt; or progressive reliance on AI for routine synthesis. A red flag may indicate that a safeguard was absent, bypassed, or ineffective. It should trigger proportionate verification or workflow escalation, not automatic attribution of harm. The proposed escalation from red flags to CAE-I through CAE-IV is summarized in Table 2; the six principal terms and their conceptual boundaries are summarized in Table 3.

Sentinel phenotype: AI-induced diagnostic narrowing

The most useful sentinel phenotype is AI-induced diagnostic narrowing: the premature, inappropriate, or insufficiently verified reduction of the diagnostic differential after exposure to AI output. The phenomenon includes automation bias and algorithmic anchoring, but it is more specific. Its distinguishing feature is loss of the diagnostic tail: low-frequency but high-impact hypotheses become non-operative, not because they were actively excluded, but because the AI interaction narrowed the cognitive space being explored.

Consider a patient with atypical chest pain, dyspepsia, and mild dyspnea. Before AI, the clinician keeps reflux, anxiety, and atypical acute coronary syndrome in the differential. After a fluent AI output focused on reflux and anxiety, the cardiac hypothesis remains formally possible but no longer drives action. The electrocardiogram is deferred, the verification threshold falls, and the error is recognized later.

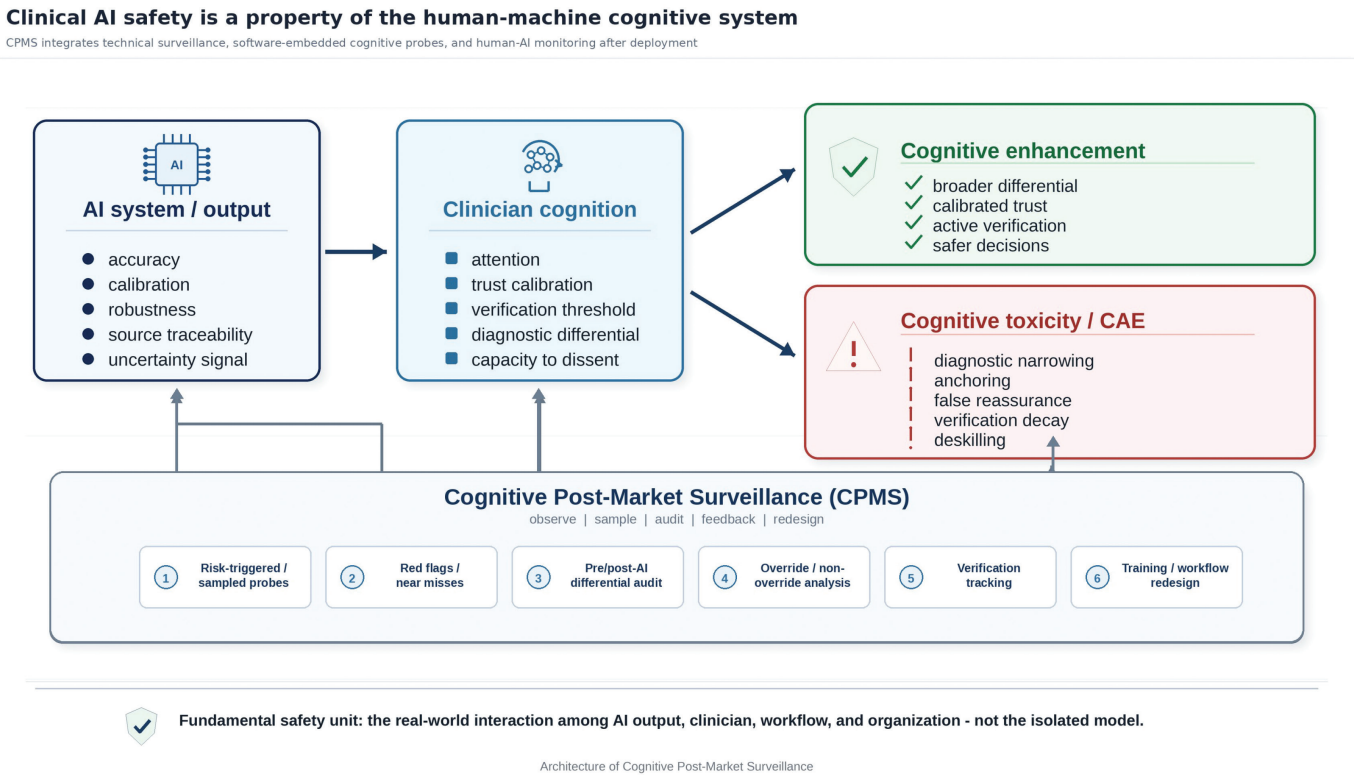

FIGURE 1 - Architecture of Cognitive Post-Market Surveillance. The AI output enters clinician cognition and may produce cognitive enhancement or cognitive toxicity/Cognitive Adverse Events. CPMS integrates technical surveillance, risk-triggered and randomly sampled pre-output reasoning probes, user- and interaction-level cognitive surveillance, and feedback through red flags and near misses, pre/post-AI differential audit, override and non-override analysis, verification tracking, training, and workflow redesign.

TABLE 1 - Integrated surveillance architecture for the human–AI cognitive system

Surveillance domain	Role and primary question	Exemplary indicators or safeguards	Risk intercepted
Technical/software surveillance (enabling layer)	Does the system function adequately, and do cognitive safeguards operate, in the local context of use?	Local accuracy; calibration; drift; hallucination rate; source traceability; uncertainty communication; interaction logging; risk-triggered or sampled pre-output probes; completion or bypass.	Technical error; performance decay; non-applicable output; absent, bypassed, or ineffective cognitive safeguard.
User-level cognitive surveillance (cognitive algovigilance)	Does the clinician use the system in a cognitively safe way?	Pre-AI manual reasoning; verification rate; override quality; calibration gap; quality of dissent; operational dependence.	Overtrust; undertrust; reduced verification; cognitive dependence.
Interaction-level cognitive surveillance (cognitive algovigilance)	What emerges from the coupled system-clinician interaction?	Pre/post-AI differential shift; rare-but-dangerous retention; tail preservation; time-to-second-thought; cognitive red flags; cognitive near misses.	Diagnostic narrowing; anchoring; premature closure; false reassurance; epistemic deskilling.

Traditional analysis asks whether the AI was wrong. Cognitive algovigilance asks what the interaction did to the differential, trust, and verification behavior.

Diagnostic narrowing is operationally tractable. It can be measured by comparing pre-AI and post-AI diagnostic lists, tracking must-not-miss diagnoses, and determining whether rare but dangerous hypotheses remained linked to clinical action rather than only appearing as nominal possibilities.

Cognitive Post-Market Surveillance: a complementary post-deployment layer

Cognitive Post-Market Surveillance (CPMS) is the operational layer of cognitive algovigilance: continuous monitoring of cognitive, behavioral, and decisional signals generated by real-world clinical AI use after a system is placed on the market, put into service, or locally deployed. We retain the

TABLE 2 - Escalation from cognitive red flags to Cognitive Adverse Events

Level	Operational definition	Clinical example	Suggested response
Cognitive red flag (pre-CAE signal)	Observable signal of elevated cognitive risk, arising from routine traces or a software probe, without a documented decision-process modification meeting CAE criteria.	A clinician accepts a fluent output without an independent hypothesis, a must-not-miss diagnosis disappears, or a prompted manual response is repeatedly bypassed.	Trigger or repeat a brief pre-output probe, independent verification, or escalation; log activation, completion, and bypass.
CAE-I	Mild cognitive distortion, recognized or rapidly corrected, without substantial modification of the clinical decision.	The clinician notices excessive focus on the AI hypothesis and reopens the differential.	Local annotation, formative feedback, prompt or interface review.
CAE-II	AI-mediated and reversible decisional modification, intercepted before clinical execution.	A therapy is initially selected after false reassurance and then corrected after independent verification.	Internal reporting, case audit, analysis of the cognitive factor.
CAE-III	Relevant AI-mediated clinical error or omission, with potential or actual impact on the patient.	An alternative diagnosis is not explored after anchoring to the AI output, with diagnostic delay.	Incident reporting, multidisciplinary review, workflow correction.
CAE-IV	Systemic cognitive dependence or observable metacognitive deterioration in a team or process.	Stable reduction in autonomous generation of the differential or an increase in unverified decisions.	Institutional surveillance, retraining, restrictions of use, system redesign.

TABLE 3 - Glossary of key terms and conceptual boundaries

Term	Operational definition	Conceptual boundary	Primary surveillance implication
Cognitive intervention	AI exposure capable of changing attention, trust, verification thresholds, diagnostic breadth, or willingness to dissent.	Describes an effect on reasoning, not therapeutic intent or model class.	Evaluate the human-machine cognitive system, not output accuracy alone.
Cognitive algovigilance	Systematic surveillance of cognitive effects arising from AI-clinician-workflow interaction.	Strictly concerns user and interaction effects; technical surveillance is an enabling layer.	Monitor verification, dissent, dependence, and changes in reasoning.
Cognitive red flag	Observable precursor of elevated cognitive risk, detected in routine traces or elicited by a software-embedded probe, that does not yet meet CAE criteria.	Triggers review or safeguards; a bypass is a signal, not automatic evidence of harm.	Use risk-triggered and randomly sampled probes; track recurrence, completion, and bypass.
Cognitive Adverse Event (CAE)	Measurable AI-mediated change in clinical decision-making that increases the probability of a suboptimal decision, regardless of model correctness.	Requires a documentable process change and clinical relevance.	Classify severity, investigate systemic attribution, and mitigate.
AI-induced diagnostic narrowing	Premature, inappropriate, or insufficiently verified contraction of the differential after AI exposure, including loss of the diagnostic tail.	Sentinel CAE phenotype; not every clinically appropriate narrowing.	Compare pre-AI and post-AI differentials and retain must-not-miss diagnoses.
Cognitive Post-Market Surveillance (CPMS)	Continuous post-deployment collection and analysis of cognitive, behavioral, and decisional signals, including sampled pre-output reasoning, in real-world AI use.	Complements formal regulatory surveillance and may apply beyond medical-device classification.	Feed findings into software, sampling rules, workflow, training, governance, and regulation.

term post-market to locate the activity in the lifecycle phase in which use-related effects emerge and accumulate. CPMS is neither a substitute for nor a redefinition of formal regulatory post-market surveillance or post-market monitoring (2,3). It is a complementary clinical-cognitive layer that can inform institutional governance, developer feedback, and regulatory safety processes, including for systems not legally classified as medical devices.

To avoid a category error, technical/software surveillance should be distinguished from cognitive algovigilance proper. Technical surveillance asks whether the system functions adequately; cognitive algovigilance asks what the system does to clinical reasoning. The domains nevertheless require an integrated architecture. CPMS therefore combines: (i) a technical enabling layer monitoring local performance, drift, failure modes, source quality, uncertainty communication, interaction logging, and the operation of risk-triggered and randomly sampled cognitive probes and other safeguards; (ii) user-level cognitive surveillance of independent verification, quality of dissent, calibration gap, and operational dependence; and (iii) interaction-level cognitive surveillance of changes in the differential, diagnostic narrowing, trust shifts, cognitive near misses, and final decisions. User- and interaction-level surveillance constitute cognitive algovigilance in the strict sense; technical/software surveillance enables, triggers, and responds to it.

Software and interface design are therefore not external to cognitive safety. One implementable design is a software-embedded cognitive probe. At risk-triggered encounters and at a randomly sampled fraction of otherwise routine encounters, the interface briefly stages disclosure of the AI output and asks the clinician to enter a concise unaided response: a leading hypothesis, short differential, must-not-miss diagnosis, confidence estimate, or intended action. The output is then released, allowing CPMS to compare pre-AI and post-AI reasoning and to log completion, bypass, verification, and downstream red flags. Risk-triggered probes increase sensitivity to known hazards; random sampling provides a less biased surveillance denominator and makes latent dependence observable. In live high-stakes care, probes must be brief, non-punitive, time-bounded, and immediately bypassable with a documented reason; they must never delay urgent action or withhold clinically necessary support.

A crucial methodological point is that CPMS must not be limited to overrides. Failure to override may be safe when AI is correct and critically integrated. It may also indicate overtrust, loss of dissent, or verification decay. High-risk non-overrides therefore require review, especially when the output is fluent, assertive, or aligned with the clinician's initial dominant hypothesis.

Metrics and operational tools

Cognitive safety cannot remain a qualitative impression. It must become measurable without becoming a documentation burden. Candidate indicators include Differential Breadth Index, Rare-but-Dangerous Retention Rate, Tail Preservation Index, Verification Rate, Override Quality, Calibration Gap, Dependency Index, Verification Decay, Time-to-second-thought, Cognitive Red-Flag Rate, Cognitive Probe Completion Rate, and Safeguard Activation/Bypass Rate. These are not validated standards; they are candidate endpoints for prospective studies, audit programs, and consensus development.

Two tools can make the framework testable. The first is the Chain-of-Verification: formulation of pre-AI reasoning; exposure to the AI output; independent verification of recommendations; explicit search for a serious alternative or counterfactual; final decision; and documentation of what changed and why.

The second is the GENERATE-Dx protocol: Generate the pre-AI differential; Elicit AI alternatives; Name one diagnosis not to be missed; Examine data for and against; Refute the dominant hypothesis with a counterfactual prompt; Audit sources and applicability; Track trust and breadth of the differential; and Escalate when residual uncertainty remains clinically relevant. The protocol is proposed as an experimental scaffold, not as a validated standard.

Governance, consensus, and general practice

Cognitive algovigilance changes the governance question. The issue is not only whether an AI tool should be purchased, validated, updated, or restricted. The issue is whether the organization can observe how that tool reshapes diagnostic routines, epistemic habits, and safety culture over time. Training must therefore move beyond prompt literacy. Clinicians need metacognitive training: recognizing when AI expands reasoning and when it narrows it, when uncertainty is clarified and when it is rhetorically compressed, and when trust is calibrated rather than induced by fluency.

The proposed definitions require formal consensus. Delphi methods are widely used in health sciences when evidence is incomplete or heterogeneous, but they require transparent methods and explicit reporting criteria (28,29). A validation program should include clinicians, patients, patient-safety experts, informaticians, regulators, developers, and medical educators. It should also define severity levels, reporting thresholds, and safeguards against punitive use of cognitive surveillance.

General practice is a critical setting. Symptoms are often subtle, uncertainty is high, relational continuity is strong, and workload is substantial. In this context, the diagnostic differential and escalation threshold are central safety instruments. AI-induced diagnostic narrowing should therefore be studied not only in hospital and imaging settings, but also in community-based care. The medico-legal dimension is also relevant because liability in AI-assisted diagnosis remains unsettled and reinforces the need for traceable reasoning and governance of interaction (30).

Research agenda

The framework opens an empirical research program. Priorities include standardized measurement of diagnostic breadth, calibrated trust, quality of dissent, and cognitive

dependence; comparison of pre-AI and post-AI decisions in simulated and real-world scenarios; cognitive stress testing with neutral, assertive, or incomplete outputs; comparison of risk-triggered and randomly sampled cognitive probes for detection yield, workflow burden, and measurement reactivity; multicenter registries of cognitive near misses; and longitudinal study of verification, synthesis, retrieval, and autonomous dissent among clinicians who use AI routinely.

A minimal empirical agenda should include four designs: within-clinician studies comparing pre-AI and post-AI differentials; pragmatic trials with cognitive process endpoints in addition to clinical outcomes; longitudinal cohorts on deskilling, verification decay, and operational dependence; and CPMS registries collecting cognitive red flags and CAE-I to CAE-IV events with sufficient detail on interface, context, training, and workflow. The aim is not better models in the abstract, but clinical environments in which AI expands thought rather than compressing it.

Limitations

This work is conceptual. CAE definitions, severity levels, and proposed indicators are not regulatory standards and require empirical validation. Causal attribution will be difficult because model behavior, interface design, user expertise, organizational pressure, and clinical uncertainty interact. Mixed methods and granular interaction data will be needed.

A second limitation is the risk of punitive surveillance. Cognitive algovigilance should be designed as a learning system, not as a disciplinary instrument. Its purpose is to improve interfaces, workflows, training, and governance. A third limitation is measurability. Excessive documentation could create new workload and new errors. Because a software-embedded probe is itself a cognitive intervention, it may create measurement reactivity, delay, alert fatigue, or ritualized compliance; its sampling rate, timing, and bypass rules therefore require risk-proportionate testing. CPMS should be proportionate to risk: more intensive in high-uncertainty, high-impact, AI-dependent decisions; lighter in low-risk administrative or informational uses.

Finally, the framework is not anti-AI. AI can broaden differentials, improve recall, reduce omissions, and support safer decisions. The purpose is to distinguish cognitive enhancement from cognitive toxicity and to design systems that preserve clinical judgment while augmenting it.

Conclusions

The safety of clinical AI does not coincide with the average correctness of its outputs. It depends on the capacity of health systems to protect clinical judgment while exposing it to a new epistemic artifact. Cognitive algovigilance places the human-machine cognitive system within patient-safety surveillance. CPMS operationalizes that shift by integrating technical surveillance, software-embedded cognitive probes, and user- and interaction-level cognitive surveillance, and by turning cognitive red flags, automation bias, false reassurance, diagnostic narrowing, verification decay, and deskilling into signals that can be observed, audited, and mitigated.

Clinical AI should be evaluated not only for what it answers, but for what it does to the way clinicians think.

Acknowledgments

No non-author contributions are reported. OpenAI ChatGPT (GPT-5.5 Pro, OpenAI; June 2026) was used to support language editing, stylistic revision, consistency checking, bibliography formatting, figure drafting, and preparation of the submission file. The authors reviewed and edited all AI-assisted material and remain fully responsible for the final content, interpretation, references, and submitted manuscript.

Disclosures

Conflict of Interest: The authors declare no conflicts of interest related to this manuscript.

Financial Support: This work was conducted independently and received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement: Data sharing is not applicable to this article because it is a Point of View manuscript, and no original datasets, patient-level data, surveys, interviews, or statistical analyses were generated or analyzed.

Author Contributions: DA: conceptualization, methodology, writing - original draft, writing - review and editing. FDB: conceptualization, methodology, writing - review and editing. Both authors approved the final version and agree to be accountable for the work.

References

1. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. Online (Accessed June 2026)
2. European Parliament, Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonized rules on artificial intelligence. Official Journal of the European Union. 2024. Online (Accessed June 2026)
3. European Parliament, Council of the European Union. Regulation (EU) 2017/745 of 5 April 2017 on medical devices. Official Journal of the European Union. 2017. Online (Accessed June 2026)
4. U.S. Food and Drug Administration. Health Canada, Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. Online (Accessed June 2026)
5. U.S. Food and Drug Administration. Health Canada, Medicines and Healthcare products Regulatory Agency. Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles. 2024. Online (Accessed June 2026)
6. International Medical Device Regulators Forum. Good machine learning practice for medical device development: guiding principles. IMDRF/AIML WG/N88 FINAL:2025. 2025. Online (Accessed June 2026)
7. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: NIST; 2023. Online (Accessed June 2026)
8. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial

Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: NIST; 2024. Online (Accessed June 2026)
9. Lekadir K, Frangi AF, Porras AR, et al.; FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554. CrossRef PubMed
10. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024;7(1):258. CrossRef PubMed
11. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med. 2021;4(1):54. CrossRef PubMed
12. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-127. CrossRef PubMed
13. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 2021;4(1):31. CrossRef PubMed
14. Cabitza F, Campagner A, Ronzio L, et al. Rams, hounds and white boxes: investigating human-AI collaboration protocols in medical diagnosis. Artif Intell Med. 2023;138:102506. CrossRef PubMed
15. Reverberi C, Rigon T, Solari A, et al.; GI Genius CADx Study Group. Experimental evidence of effective human-AI collaboration in medical decision-making. Sci Rep. 2022;12(1):14952. CrossRef PubMed
16. Knop M, Weber S, Mueller M, et al. Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence-enabled clinical decision support systems: literature review. JMIR Hum Factors. 2022;9(1):e28639. CrossRef PubMed
17. Rosenbacke R, Melhus Å, McKee M, et al. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. JMIR AI. 2024;3:e53207. CrossRef PubMed
18. Tsai CC, Kim JY, Chen Q, et al. Effect of artificial intelligence helpfulness and uncertainty on cognitive interactions with pharmacists: randomized controlled trial. J Med Internet Res. 2025;27:e59946. CrossRef PubMed
19. Ancker JS, Edwards A, Nosal S, et al; with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2017;17(1):36. CrossRef PubMed
20. Natali C, Marconi L, Dias Duran LD, et al. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. Artif Intell Rev. 2025;58(11):356. CrossRef
21. National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. National Academies Press; 2015. CrossRef
22. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med. 2005;165(13):1493-1499. CrossRef PubMed
23. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003;78(8):775-780. CrossRef PubMed
24. Singh H, Giardina TD, Meyer AND, et al. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med. 2013;173(6):418-425. CrossRef PubMed
25. Loru E, Nudo J, Di Marco N, et al. The simulation of judgment in LLMs. Proc Natl Acad Sci USA. 2025;122(42):e2518443122. CrossRef PubMed
26. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50-80. CrossRef PubMed
27. Parasuraman R, Sheridan TB, Wickens CD; New Collective Author. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum. 2000;30(3):286-297. CrossRef PubMed
28. Gattrell WT, Logullo P, van Zuuren EJ, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med. 2024;21(1):e1004326. CrossRef PubMed
29. Niederberger M, Spranger J. Delphi technique in health sciences: a map. Front Public Health. 2020;8:457. CrossRef PubMed
30. Cestonaro C, Delicati A, Marcante B, et al. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med (Lausanne). 2023;10:1305756. CrossRef PubMed