

PERSPECTIVE ARTICLE

From P(doom) to P(harm) in healthcare:
An operational surveillance framework for
accumulative clinical riskArturo Loaiza-Bonilla^{1,2*}  and Nikhil G. Thaker^{3,4} ¹Massive Bio, Boca Raton, Florida, United States of America²St. Luke's University Health Network, Lewis Katz School of Medicine, Temple University, Bethlehem, Pennsylvania, United States of America³Capital Health Medical Center, Pennington, New Jersey, United States of America⁴Bayta Systems, Newtown, Pennsylvania, United States of America

Abstract

Debates about artificial intelligence (AI) safety often reach for P(doom), a shorthand in AI-safety discussions for the probability of catastrophic AI risk. Healthcare needs a nearer and more testable companion: P(harm), a proposed surveillance framework for cumulative clinical risk. We define accumulative risk as harm that arises from repeated small shifts in diagnosis, triage, clinician behavior, training, or workflow rather than a single catastrophic failure. We operationalize P(harm) as a family of workflow-specific, severity-weighted conditional estimates: in a declared workflow and model version, it is the estimated probability that an independently reviewed AI-exposed encounter has an adverse outcome of at least a prespecified severity within a prespecified time horizon, after stratifying or adjusting for acuity, case mix, and local workflow covariates. In pilot use, the estimator should be prespecified, for example, a risk-adjusted empirical incidence with confidence or credible intervals or a hierarchical logistic/Bayesian model; P(harm) is not a causal-attributable fraction unless paired with a valid comparison design. The framework is organized around four local pathways—AI–clinician discordance in complex patients, automation bias, skill erosion and never-skilling, and consumer AI triage failure—plus a fifth, less mature pathway for correlated vendor or foundation-model failures across institutions. The motivating evidence is early and heterogeneous, often retrospective or vignette-based, so each pathway is framed as a surveillance hypothesis rather than proof of generalized harm. The purpose is not to slow beneficial AI, which may reduce errors and administrative burden, but to make clinical drift visible early enough for health systems to adjudicate, recalibrate, roll back, or retrain before repeated small failures accumulate.

Keywords: Artificial intelligence; Patient safety; Clinical decision support; Automation bias; Causal inference; Accumulative risk; Health-system governance; Clinical AI surveillance

***Corresponding author:**Arturo Loaiza-Bonilla
(arturolb@massivebio.com)

Citation: Loaiza-Bonilla A, Thaker NG. From P(doom) to P(harm) in healthcare: An operational surveillance framework for accumulative clinical risk. *Artif Intell Health*. doi: 10.36922/AIH026250061

Received: June 15, 2026**Revised:** July 7, 2026**Accepted:** July 7, 2026**Published online:** July 16, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In artificial intelligence (AI) safety debates, P(doom) refers to the probability that advanced AI could cause human extinction or an irreversible civilizational catastrophe.¹ That debate is important, but it sits at a different analytical level from the safety problems already facing hospitals and clinics, healthcare workers, patients, and their caregivers. Medical AI is becoming routine in documentation, diagnostics, decision support, image interpretation, and self-triage. The immediate question for health systems is therefore less whether a single future event can be assigned a global probability, and more whether smaller AI-related changes in everyday care can be identified, reviewed, audited, and corrected before they accumulate.

Kasirzadeh's distinction between decisive and accumulative AI risk clarifies this gap.² Decisive scenarios imagine an abrupt takeover by a highly capable system; accumulative scenarios describe smaller, interconnected disruptions that compound over time. In healthcare, drift should be understood broadly. It includes model drift, as well as workflow drift, clinician behavior drift, trainee learning drift, and patient behavior drift. These forms of drift are not as dramatic as apocalypse, but they are more observable and more governable.

This is not a technopessimistic argument. AI systems can reduce human error, improve adherence to guidelines, surface red flags, support image review, and reduce documentation or administrative burden.³ Those benefits are precisely why the safety question should be operational rather than rhetorical. A useful framework should help health systems preserve the gains of AI while detecting when a tool, workflow, or reliance pattern begins to degrade clinical safety. We use P(harm) for that purpose: not as a universal score, but as a proposed family of monitored, workflow-specific safety estimates.

P(harm) is not intended to replace ordinary patient-safety metrics such as adverse-event rates, diagnostic-delay reviews, mortality reviews, readmissions, or incident reports. It differs by making the AI-exposed workflow the denominator, linking model and version exposure to subsequent clinician or patient action, stratifying by local context and severity, and requiring independent validation of triggers. In short, existing metrics ask whether harm occurred; P(harm) asks whether harm is changing in a defined AI-exposed workflow in a way that warrants review.

A note on terminology is necessary. We use "AI" only as an umbrella term. The relevant systems differ: large language models (LLMs) used for diagnostic reasoning or consumer triage; computer-aided detection and image-

classification systems, such as mammographic Breast Imaging Reporting and Data System (BI-RADS) support or colonoscopy computer-aided detection (CADe); and predictive or automation tools embedded in clinical workflows. Some features of P(harm), such as prospective exposure logging, outcome adjudication, and recalibration after workflow change, are general. The exposure definition, outcome window, ground truth, and attribution strategy must be system-specific.

2. Four local pathways of drift

The following pathways are not a systematic review and should not be read as settled effect estimates. They are motivated by a small, heterogeneous, and rapidly changing evidence base. Several studies use retrospective cases, standardized vignettes, or single procedures rather than live clinical workflows. Those limitations are central to the argument: they justify surveillance hypotheses and pilot testing, not broad causal claims.

2.1. Artificial intelligence–clinician discordance in complex patients

Discordance between AI output and clinician judgment is most informative when cases are complex and the stakes are high. A 2025 *BJS Open* retrospective study comparing ChatGPT-4 (GPT-4) with colorectal cancer multidisciplinary meeting decisions reported 72.5% overall concordance, with discordance clustering among older patients and those with American Society of Anesthesiologists class III or higher; excluding high-risk groups increased concordance to 85.1%.⁴ A virtual urgent care study similarly found that initial AI recommendations and final physician recommendations were not always aligned.⁵

These studies are useful but limited. The colorectal cancer study was single-centered and retrospective, used a general-purpose LLM outside an integrated clinical workflow, and did not test prospective patient outcomes. It therefore cannot support a general claim that LLM-multidisciplinary team discordance predicts harm. It does, however, identify a plausible surveillance target: discordance in high-complexity or high-acuity strata, where standardized recommendations may collide with comorbidity, frailty, patient preference, and uncertainty. Discordance should not be read automatically as an AI error or a clinician error. It is a review signal.

2.2. Automation bias

Automation bias arises when a recommendation changes human judgment before the human has formed an independent view. In a randomized clinical trial, GPT-4

alone outperformed physicians on diagnostic reasoning cases, but physicians given GPT-4 did not significantly outperform those using conventional resources.⁶ The same behavioral problem appears outside LLMs. Incorrect AI BI-RADS suggestions in mammography reduced reader performance across experience levels.⁷ These findings do not imply that AI assistance is futile. They show that the performance of the human-AI team is not guaranteed by model accuracy alone and must be measured as a workflow property.

The clinical trial reported by Goh *et al.*⁶ is one of the strongest motivating studies, but it used standardized case vignettes rather than live clinical encounters. Therefore, ecological validity matters. In live care, automation bias may be amplified or attenuated by time pressure, interruptions, clinical uncertainty, patient interaction, local norms, and workload. P(harm) surveillance should therefore measure automation bias prospectively in the deployed workflow rather than extrapolating directly from vignette performance.

2.3. Skill erosion and never-skilling

Skill erosion is harder to prove than short-term automation bias, and the claim should be kept narrow. A multicenter observational study of colonoscopy reported a lower adenoma detection rate during unassisted procedures after routine exposure to CAdE, suggesting that continuous AI support may change endoscopist behavior when the tool is absent.⁸ This is plausible and important, but it is one observational study of a specific, procedurally anchored perceptual skill. It should not be generalized without qualification to all diagnostic reasoning, problem synthesis, knowledge retrieval, or judgment.

The educational concern is related but distinct. Recent medical-education literature has described deskilling, mis-skilling, and never-skilling, with never-skilling referring to the risk that learners may fail to acquire foundational independent skills when AI supplies the first-pass work too early.⁹⁻¹¹ Direct evidence from medical training remains limited. We therefore use skill erosion and never-skilling as surveillance hypotheses: they apply most clearly to tasks in which clinicians or trainees must remain a resilient unaided backstop when AI is unavailable, confidently wrong, overfitted, poorly calibrated, or inappropriate for a patient context.

2.4. Consumer artificial intelligence triage failure

Consumer-facing LLM triage creates a different surveillance problem because the first clinical decision may occur outside the health system. In a structured evaluation of ChatGPT Health, 52% of true emergencies

were under-triaged, and triage urgency decreased when bystanders minimized symptoms.¹² Unlike an embedded electronic health record (EHR) tool, consumer AI use may be invisible unless patients are asked about it. A P(harm) approach, therefore, cannot rely on EHR logs alone. It would need validated vignettes, patient-reported AI use at intake, time-to-presentation measures, and careful attention to ascertainment bias.

This evidence also requires caution. The ChatGPT Health evaluation used structured vignettes rather than real-world conversations. Patients and caregivers may prompt iteratively, omit symptoms, exaggerate symptoms, translate across languages, copy model responses into messages, or seek care despite reassurance. Conversely, some patients who are reassured by a tool may never present, leaving no denominator in the health system record. A surveillance pathway for consumer AI should therefore distinguish between vignette performance and real-world care-seeking behavior, and treat patient-reported use as an incomplete signal.

3. From P(doom) to P(harm)

3.1. Defining the estimand and pilot estimator

At this stage, P(harm) should be read as a conceptual and operational research framework, not as a validated governance metric. To be testable, each P(harm) implementation should declare six elements in advance: the workflow boundary, the AI exposure definition, the unit of analysis, the severity scale, the outcome horizon, and the estimator. Without those elements, P(harm) becomes only a general call for AI-outcome logging.

Let i index encounters, and let s and τ denote the prespecified severity threshold and outcome horizon. Define $Y_i = 1$ if an independently review-confirmed adverse outcome of severity at least s occurs within τ , and $Y_i = 0$ otherwise. A schematic surveillance estimand is shown in Equation 1:

$$P_{harm}(w, a, s, \tau, m) = Pr(Y_i = 1 \mid E_i = 1, w_i = w, a_i = a, m_i = m, X_i) \quad (1)$$

where E_i is documented AI exposure before the relevant clinical action, w_i is the workflow, a_i is the acuity or complexity stratum, m_i is the model and version, and X_i is a prespecified set of local covariates. The unit of analysis must be stated: patient, encounter, order, image, procedure, institution, or population. Severity weighting prevents frequent minor events from masking rare serious harms.

A pilot can begin with a transparent empirical estimator: among AI-exposed encounters in the declared

stratum, count review-confirmed adverse outcomes of severity at least s within τ and divide by the number of eligible AI-exposed encounters, with confidence or credible intervals. A risk-adjusted extension could use a prespecified logistic, hierarchical, or Bayesian model, for example, modeling the probability of Y_i as a function of acuity, case mix, clinician role, workload, model/version, site, and exposure pattern. Calibration then means comparing estimated and observed review-confirmed outcome rates within local strata over time, especially after model updates or workflow changes. The selected estimator should be named in the surveillance plan before results are reviewed.

Influence should be recorded separately from exposure. Timestamped AI outputs, view logs, accepted suggestions, edited drafts, draft and signed orders, overrides, and downstream verifications can identify whether AI was plausibly available to influence the decision. They cannot, by themselves, prove that the AI caused the decision. For this reason, P(harm) is best reported as an early-warning estimate and trend, paired with adjudication and causal designs where feasible, rather than as an automatic causal attribution score.

A practical pilot might begin with one workflow, one model class, one severity outcome, and one validation period. For example, an LLM-supported virtual urgent care workflow could be monitored for high-acuity under-triage over 90 days, with flagged cases reviewed by clinicians blinded to whether the initial recommendation was AI-assisted. A useful null result would also be prespecified: no excess validated under-triage among AI-exposed encounters compared with a matched or concurrent control, no severe case in which blinded reviewers judge AI influence plausible, and no trigger that adds actionable information beyond existing quality monitoring. Such a null would suggest that this particular P(harm) implementation does not add safety value and should be revised or discontinued.

3.2. Attribution and causal identification

The causal contribution of AI is the difficult part and should be given more weight than a dashboard metaphor implies. Clinicians who override AI recommendations are likely to differ systematically from those who accept them in terms of experience, specialty, risk tolerance, workload, institutional culture, and case mix. Patients who consult consumer AI before presenting may differ in health literacy, symptom severity, access barriers, language, trust, and care-seeking behavior. These differences can create misleading concordance, override, delay, and outcome patterns even when AI has no causal effect.

The causal estimand is distinct from the surveillance estimand. A causal question asks how outcomes would have differed if the same encounter had occurred with versus without AI exposure. That counterfactual cannot be read from exposure logs alone. Interrupted time-series analysis can help when there is a stable pre-intervention trend and no major co-interventions. Difference-in-differences designs require a comparable control workflow or site and a plausible parallel-trends assumption. Randomized or stepped-wedge rollouts are stronger but often difficult, because health systems deploy tools for operational, contractual, or political reasons rather than experimental balance. Propensity adjustment, negative-control outcomes, and sensitivity analyses may reduce but not remove bias.

Accordingly, early P(harm) implementations should be honest about what they can deliver. They can provide high-sensitivity non-causal signals for review, identify strata where the human-AI workflow deviates from baseline, and prioritize chart review, recalibration, training, or rollback. They should not claim that AI caused harm unless the evidence supports that inference. The framework's value may be attenuated in real-world settings by confounding, incomplete exposure logging, and unobserved patient behavior; those limitations should be part of the governance decision.

All triggers in [Table 1](#) are illustrative pilot triggers rather than proposed standards. Each would require feasibility testing, reliability assessment, calibration, bias assessment, and evidence that responding to the trigger improves patient outcomes.

3.3. Beyond local surveillance: correlated and systemic risk

Workflow-indexed surveillance is necessary but not sufficient. Local dashboards may miss synchronized failures when many institutions depend on the same foundation model, vendor infrastructure, data pipeline, or silent update. An erroneous AI-exposed result can also propagate through shared records, registries, or downstream training data. This correlated-risk pathway is qualitatively different from local surveillance because the relevant denominator may be a vendor network, model family, or update window rather than a single institution.

A detectable, synchronized shift does not imply that raw rates are identical across institutions, because baseline rates, case mix, coding, acuity, and workflow differ. A more defensible signal would compare each site with its own recent baseline and then look for concurrent standardized shifts, such as multiple sites crossing prespecified control-chart limits or showing elevated risk-adjusted rate

Table 1. Proposed P(harm) surveillance framework: illustrative operational metrics, validators, and time horizons for monitoring AI-exposed clinical risk

P(harm) pathway	System class and exposure definition	Observable signal	Minimum data sources	Hypothetical trigger, validator, and time horizon
AI–clinician discordance	LLM recommendation generated and viewed before plan in a high-acuity or complex workflow; model/version logged.	Override or discordance rate and adverse outcome rate stratified by complexity, acuity, and severity.	Timestamped AI output, view logs, draft and signed orders, acuity flags, clinician role, model/version, and chart review.	Trigger: Sustained shift > 2 SD from rolling local baseline over four weeks or 90-day pilot; Validator: Independent safety team with blinded specialty review; adjust for case mix and clinician role.
Automation bias	AI recommendation displayed before independent clinician assessment, or clinician action follows AI recommendation, later judged incorrect.	Commission-error rate when AI recommendation is incorrect; time to override in flagged cases.	AI output and confidence, clinician action logs, comparator or adjudicated ground truth, workload/alert data, outcome review.	Trigger: Increase above predeployment or concurrent-control benchmark; Validator: Blinded adjudication of AI correctness; review workload, alert fatigue, and staffing over monthly cycles.
Skill erosion/never-skilling	Routine AI support in a task that clinicians or trainees must still perform unaided.	Unassisted diagnostic or perceptual accuracy during AI-off audits; trainee milestones; recovery after AI removal.	Periodic AI-off drills, audit cases, training records, case difficulty measures, and supervisor assessments.	Trigger: Prespecified decline below competency threshold during quarterly AI-off audits; Validator: Education/QI reviewers blinded to exposure phase; adjust for case difficulty and learning curves.
Consumer AI triage failure	Patient reports using consumer AI before deciding whether or when to seek care.	Under-triage of clinician-rated emergencies; delay to presentation; ED bounce-back or escalation after reassurance.	Validated vignette testing, ED intake questions, timestamps, patient-reported AI use, and acuity adjudication.	Trigger: Rising delayed-presentation signal over 90 days or high vignette under-triage; Validator: Blinded acuity panel; assess ascertainment and unknown-denominator bias.
Correlated/systemic drift	Multiple sites share a model, vendor, version, retrieval layer, prompt template, or update pathway.	Cross-site standardized rate ratios or control-chart breaches around model updates; failed post-update revalidation.	Model/version metadata, vendor change logs, local aggregate outcome summaries, revalidation results, privacy-preserving reporting.	Trigger: Synchronized standardized shift in ≥ 2 sites within an update window; Validator: Trusted safety network; exclude concurrent policy, population, staffing, seasonal, or coding changes; suppress small cells.

Abbreviations: AI: Artificial intelligence; ED: Emergency department; LLM: Large language model; QI: Quality improvement; SD: Standard deviation.

ratios within a shared model version or update window. Validation would need to exclude concurrent changes in policy, staffing, referral patterns, seasonality, case definitions, or coding.

Cross-institutional monitoring also requires privacy-preserving aggregation. Sites should not exchange patient-level data or free-text cases by default. A trusted safety network could share aggregate counts, denominators, severity strata, model and version identifiers, update windows, and revalidation pass/fail results, with minimum cell-size rules, date coarsening, and suppression of rare combinations to prevent reverse identification. Federated queries or secure distributed analyses may be preferable when even aggregate data are sensitive.

Model and version tracking is a central operational obstacle. Many health systems may not know whether a

vendor has silently changed prompts, routing logic, fine-tuning, retrieval corpora, or the underlying foundation model. P(harm) governance should therefore require stable model-family and version identifiers, update logs, rollback paths, local gateway logs, and contractual notice of material changes. Where those conditions cannot be met, systemic P(harm) should be positioned as a research agenda and procurement requirement rather than a fully operational safety program.

4. Governance before scale

Governance should assign responsibility before deployment and should fund surveillance as part of the cost of AI implementation. Monitoring P(harm) should not fall solely to the vendor, the individual clinician, or an informal quality-improvement group. A clinical AI safety committee—including patient-safety leadership, frontline

clinicians, informaticians, data scientists, compliance or risk staff, educators when trainees are affected, and when appropriate patient representatives—should own the surveillance plan, review cadence, and escalation pathway.

The committee should sit within, or report directly to, existing patient-safety and quality infrastructure so that AI triggers become part of the institution's ordinary safety architecture rather than a parallel technology project. It should have authority to require a second review, pause a workflow, request vendor information, mandate post-update revalidation, and escalate serious events to the board quality committee or equivalent leadership body. Resourcing should include data-engineering support, clinical reviewer time, analyst time, and protected time for frontline domain experts.

Accountability also requires conflict management. The group that sponsors or purchases an AI product should not be the only group judging its safety. Institutions can separate the business owner from the safety owner, require conflict-of-interest declarations, include independent patient-safety leadership, document decisions in minutes, and use vendor contracts that preserve access to audit logs, model/version metadata, update notices, incident reports, and rollback procedures. Safety oversight should be able to recommend de-implementation even when the tool is financially or operationally attractive.

Regulatory transparency is moving in this direction, but should not be overstated. The Office of the National Coordinator for Health Information Technology's Health Data, Technology, and Interoperability-1 final rule establishes transparency requirements for AI and other predictive algorithms as part of certified health information technology, but it is not a universal nutrition-label mandate for every health AI system.¹³ The Food and Drug Administration clinical decision support guidance clarifies when clinical decision support software is treated as a medical device and emphasizes, among other criteria, whether a health care professional can independently review the basis for a recommendation.¹⁴ These rules and guidance can support P(harm) surveillance, but the health system still has to translate transparency into local obligations: an AI inventory, source attributes or equivalent documentation, predeployment safety cases, model/version tracking, prespecified monitoring metrics, post-update validation, incident reporting, and rollback authority.

Mitigation should be matched to the pathway. For automation bias, cognitive forcing can require a preliminary clinician assessment before AI output is displayed. For skill erosion and never-skilling, periodic AI-off drills and AI-independent trainee milestones can provide

longitudinal competence data. For consumer triage, hard-endpoint validation and postmarket monitoring should precede broad deployment; a registered trial is already studying behavioral nudges to reduce automation bias in physician-LLM diagnostic reasoning.¹⁵ The common principle is that safety depends on the human-AI workflow over time, not on a one-time model card or predeployment test alone.

5. Limitations

This proposal has practical limits. It may add measurement burden, depend on imperfect EHR data, miss unobserved consumer AI use, and confuse association with causation if not paired with review and causal-inference methods. It could create alert fatigue or a new documentation burden if implemented poorly. Local thresholds may be manipulated, may fail to generalize across sites, and may reflect changes in case mix, staffing, workflow, or coding rather than AI performance. The recent studies motivating several pathways are not yet extensively replicated, and vignette-based or retrospective results may not translate to live care.

The framework may also advantage institutions with mature data infrastructure, large informatics teams, and established safety analytics. Community hospitals, rural sites, and safety-net systems may have less complete logs, fewer reviewers, less leverage over vendors, and fewer resources for causal evaluation. If P(harm) becomes an expected safety practice without shared tools, funding, and lightweight implementation pathways, it could widen safety disparities. Equity-conscious implementation should therefore include minimal viable metrics, shared registries, regional safety collaboratives, and procurement standards that do not presume academic medical-center infrastructure.

These limitations argue for staged pilots, prespecified metrics, independent adjudication, transparent reporting of uncertainty, and periodic evaluation of whether surveillance itself improves safety. A P(harm) dashboard that detects many signals but produces no useful action would be a measurement theater rather than patient safety.

6. Conclusion

The greatest contribution of P(harm) is not the contrast with P(doom), but the shift from abstract concern to prospective surveillance of accumulative clinical harm. For current narrow and predictive systems, many important pathways to harm are observable enough to monitor, even if they are not simple enough to attribute automatically. More autonomous or adaptive systems may introduce nonlinear failure modes that require additional safeguards,

including red-teaming, staged deployment, override authority, and correlated-failure monitoring.

Healthcare does not need to choose between optimism about the potential benefits of AI and vigilance about safety. It requires robust approaches to detect when incremental workflow changes accumulate into meaningful effects. P(harm) is proposed to address this need: sufficiently localized to audit and verify, flexible enough to accommodate the rapidly evolving landscape of diverse AI systems, sufficiently cautious to treat causal claims as hypotheses requiring evidence, and specific enough to support implementation, pilot testing, and falsification. The future of AI in medicine will depend less on model performance in isolation than on whether health systems preserve accountable, auditable, resilient, proactive, and meaningful human judgment within the human–AI dyad.¹⁶

Acknowledgments

None.

Funding

None.

Conflict of interest

Arturo Loaiza-Bonilla is a co-founder and chief medical AI officer of Massive Bio and a co-founder and chief scientific officer of AEQUIS AI, and holds equity in both companies; he is an inventor on four AI-related patents (issued and/or pending); and he holds unpaid leadership roles as a Champion of the ASCO Artificial Intelligence Community of Practice, Program Chair for NeurIPS 2025, and a founding member of CancerX (US Cancer Moonshot). He is employed by St. Luke's University Health Network and holds a faculty appointment at the Lewis Katz School of Medicine at Temple University. Nikhil G. Thaker is affiliated with Bayta Systems and Capital Health. The authors declare these relationships, affiliations, equity holdings, patents, and leadership roles as potential competing interests. The authors declare no other competing interests.

Author contributions

Conceptualization: Arturo Loaiza-Bonilla, Nikhil G. Thaker

Writing—original draft: Arturo Loaiza-Bonilla

Writing—review & editing: Nikhil G. Thaker

Both authors approved the final version and are accountable for all aspects of the work.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable. No datasets were generated or analyzed during the current study.

Further disclosure

During the preparation of this manuscript, the authors used generative AI-assisted tools solely to improve the language, readability, and formatting of the text. No AI or AI-assisted technology was used to produce scientific insights, analyze or interpret data, or draw scientific conclusions. The authors reviewed and edited all content and take full responsibility for the integrity and accuracy of the publication.

References

1. Ord T. *The Precipice: Existential Risk and the Future of Humanity*. Grand Central Publishing; 2020.
2. Kasirzadeh A. Two types of AI existential risk: decisive and accumulative. *Philos Stud*. 2025;182(7):1975–2003.
doi: 10.1007/s11098-025-02301-3
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
doi: 10.1038/s41591-018-0300-7
4. Chatziisaak D, Burri P, Sparn M, Hahnloser D, Steffen T, Bischofberger S. Concordance of ChatGPT artificial intelligence decision-making in colorectal cancer multidisciplinary meetings: retrospective study. *BJS Open*. 2025;9(3):zraf040.
doi: 10.1093/bjsopen/zraf040
5. Zeltzer D, Kugler Z, Hayat L, *et al*. Comparison of initial artificial intelligence and final physician recommendations in AI-assisted virtual urgent care visits. *Ann Intern Med*. 2025;178(4):498–506.
doi: 10.7326/ANNALS-24-03283
6. Goh E, Gallo R, Hom J, *et al*. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969.
doi: 10.1001/jamanetworkopen.2024.40969
7. Dratsch T, Chen X, Mehrizi MR, *et al*. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*. 2023;307(4):e222176.
doi: 10.1148/radiol.222176
8. Budzyń K, Romańczyk M, Kitala D, *et al*. Endoscopist

- deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol.* 2025;10(10):896-903.
doi: 10.1016/S2468-1253(25)00133-5
9. Hough J, Culley N, Erganian C, Alahdab F. Potential risks of GenAI on medical education. *BMJ Evid Based Med.* 2025;30(6):406-408.
doi: 10.1136/bmjebm-2025-114339
10. Abdalnour REE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. *N Engl J Med.* 2025;393(8):786-797.
doi: 10.1056/NEJMr2503232
11. Ke Y, Jin L, Ong JCL, *et al.* AI-induced never-skilling in medical education. *Nat Med.* 2026;32(6):1997-2006.
doi: 10.1038/s41591-026-04438-y
12. Ramaswamy A, Tyagi A, Hugo H, *et al.* ChatGPT Health performance in a structured test of triage recommendations. *Nat Med.* 2026;32(5):1671-1675.
doi: 10.1038/s41591-026-04297-7
13. HTI-1 Final Rule. Office of the National Coordinator for Health Information Technology. Accessed June 24, 2026. <https://healthit.gov/regulations/hti-rules/hti-1-final-rule/>
14. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. US Food and Drug Administration. Published January 2026. Updated January 29, 2026. Accessed June 24, 2026. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>
15. Mitigating automation bias in physician-LLM diagnostic reasoning using behavioral nudges. ClinicalTrials.gov. Accessed June 24, 2026. <https://clinicaltrials.gov/study/NCT07328815>
16. Loaiza-Bonilla A, Penberthy S. Harnessing Moravec's paradox in health care: a new era of collaborative intelligence. *NEJM AI.* 2025;2(5):AIp2500005.
doi: 10.1056/AIp2500005