Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Graded Autonomy Framework for Governing Agentic AI in Health Care

Sing Chee Tan

Sing.tan@nh.org.au

Northern Health, Epping, Victoria, Australia; Melbourne Medical School, The University of Melbourne, Melbourne, Victoria, Australia

Vlada Rozova

Centre for Digital Transformation of Health, The University of Melbourne, Melbourne, Victoria, Australia

Rebecca Jessup

School of Allied Health, Human Services and Sport, La Trobe University, Melbourne, Victoria, Australia; Allied Health Research, Northern Health, Epping, Victoria, Australia

Daniel Capurro

School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia; Centre for Digital Transformation of Health, The University of Melbourne, Melbourne, Victoria, Australia

Research Article

Keywords: Artificial intelligence, Autonomous systems, Clinical governance, Patient safety, Health care delivery

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10297334/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A Graded Autonomy Framework for Governing Agentic AI in Health Care

Authors

Sing Chee Tan, MBBS, FCICM[1,2*] · Vlada Rozova, PhD[3] · Rebecca Jessup, PhD[4,5] · Daniel Capurro, MD, PhD, FAIDH, FAMIA[6,7]

[1] ICU Consultant, Director of Clinical Digital Innovation, Northern Health, Epping, Victoria, Australia

[2] PhD Candidate, Melbourne Medical School, The University of Melbourne, Melbourne, Victoria, Australia

[3] Senior Research Fellow, Centre for Digital Transformation of Health, The University of Melbourne, Melbourne, Victoria, Australia

[4] Associate Professor, School of Allied Health, Human Services and Sport, La Trobe University, Melbourne, Victoria, Australia

[5] Allied Health Research, Northern Health, Epping, Victoria, Australia

[6] Professor of Digital Health, School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia

[7] Deputy Director, Centre for Digital Transformation of Health, The University of Melbourne, Melbourne, Victoria, Australia

***** Corresponding author: Sing Chee Tan — Sing.tan@nh.org.au

Abstract

Agentic AI can autonomously execute multi-step tasks, yet current evaluation metrics overlook the operating conditions and fallback mechanisms that determine patient safety. We propose a graded-autonomy framework, adapted from the automotive J3016 standard, that classifies systems across three domains and six levels and separates capability from authorisation. Using autonomous sepsis management as a worked example, we examine implications for regulation, workforce readiness, and the safe governance of autonomous agentic AI.

Keywords : Artificial intelligence, Autonomous systems, Clinical governance, Patient safety, Health care delivery

Competing interests

The authors have no conflicts of interest to declare

Writing Assistance and AI Use Statement

Writing assistance: This manuscript was developed with the assistance of Claude (Anthropic, claude.ai), which was used for editorial review, and iterative refinement across multiple sessions including stylistic review.

AI content statement: All AI-assisted content was reviewed, edited, and verified by the authors, who take full responsibility for the accuracy, integrity, and originality of the work. No AI-generated text has been reproduced without substantive author revision. No AI-generated material is cited as a primary source. The authors confirm there is no plagiarism of text or images in AI-assisted material included in this submission.

Author Contributions (CRediT Taxonomy)

SCT: Conceptualization; Writing – original draft; Writing – review & editing
VR: Conceptualization; Writing – review & editing
DC: Writing – review & editing
RJ: Writing – review & editing

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Grant number: Not applicable.

Introduction

“The beginning of wisdom is the definition of terms” - Socrates

Until recently, AI in healthcare have operated on an assistive model where the AI makes a recommendation, and the human decides or acts. AI systems are now moving toward a more independent operating model via agentic AI systems, which can autonomously execute complex tasks with reduced human oversight(1).

We recognise that autonomous technologies are not new and already exist in healthcare. For example, volume-guaranteed modes on ventilators self-titrate driving pressures to achieve a target tidal volume. However, these rule-based algorithms have been constrained to a narrow range of low complexity tasks. The modern integration of AI technology with tooling-harnesses now expands the range of tasks agentic systems are able manage, allowing them to reason through and execute multi-stage tasks(2). This is a fundamental shift in human-AI interaction, moving from assistive tools to machines that execute complex tasks autonomously (3).

Agentic systems are a new technology category whose capabilities and failure modes are distinct from classic decision-support tools. However, current frameworks for AI classification and governance generally focus on single-task model performance characteristics such as sensitivity and specificity, in contrast to the multi-task, goal-directed execution that underpins agentic AI design (4–7). Traditional metrics characterise the accuracy of an input-to-output algorithm, but an agentic system chains many such decisions together in a non-deterministic manner; per-task performance therefore says nothing about the overall functions of the system, under what conditions it may operate, or how it behaves when an individual step fails - precisely the properties that determine patient risk.

We suggest that these new capabilities of agentic AI require a dedicated descriptive framework, with the aim of improving clarity and consistency in how such systems are defined and operate. By delineating boundaries around what a system is designed or authorised to do, risk can be clearly identified and appropriate safeguards implemented (8,9).

The purpose of this paper is to present a similar framework for describing graded autonomy in autonomous healthcare AI agents, to build shared understanding across stakeholders, and to support improved governance and implementation.

Lessons from other domains

Sheridan and Verplank (10) introduced the first formal taxonomy of human-computer control in 1978, proposing a scale ranging from assistive to full autonomous operation (3,11). This insight - that a system can automate some domains of a larger task while leaving others to the human - provides the conceptual foundation for graded autonomy frameworks, including the one proposed here. The value of these frameworks was demonstrated by Greenhalgh, who identified that inadequate characterisation of the technology domain is a primary predictor of non-adoption and failure to scale(9). The World Health Organization operationalised this principle through its Classification of Digital Interventions, providing a shared taxonomy to support coherent planning and governance of digital health implementations(8).

A prominent example of the value of a graded taxonomy for autonomous systems is in the automotive industry (3,12), where Society of Automotive Engineers introduced J3016, a standardised taxonomy for autonomous driving systems(13) that has been instrumental in the safe implementation of driver assist features. The J3016 Framework decomposes the task of driving into three distinct domains. First, the dynamic driving task (DDT) describes the actions required to drive and includes control (e.g. steering and braking) and environmental monitoring (e.g. object detection). Importantly, it introduced two additional pillars for autonomous system design – operating conditions and fallback mechanism. The operational design domain (ODD) defines the settings within which an automated system is designed to function, such as the road type and environmental conditions. The fallback domain specifies the behaviour when the system reaches its limits or encounters a failure, and how "human in the loop" fail-safe mechanisms are implemented.

The framework has had a considerable impact on the safe implementation of driver assist features. These features have varying levels of autonomy, from accelerator control to adaptive cruise control with braking capabilities to maintain distance from the car in front, and more recently, complex driving systems that incorporate steering control. Engineers and automotive regulators have adopted the SAE autonomous driving system taxonomy to guide implementation and regulation of these new features(14), providing a common language that enables the precise definitions of system operating boundaries, preventing misleading marketing claims and providing appropriate advice to users to avoid harm(15).

A Framework for Autonomous AI Agents in Healthcare

We propose a framework adapted from the SAE Framework but contextualised to the health care environment (Table 1). First, the clinical task domain (CTD) replaces the dynamic driving task and is defined for each clinical task the system is designed to address. In healthcare, CTD comprises two components: discrete action execution (the specific clinical actions the system performs) and patient monitoring and response - the continuous assessments of patient condition and initiation of appropriate responses, analogous to the Object and Event Detection and Response (OEDR) sub-task in the SAE framework. Second, the operational design domain (ODD) outlines the contextual constraints that the system is allowed to operate within. At a minimum, this should specify (1) the clinical discipline and workflow to which the system applies; (2) the care setting and level of acuity; (3) the patient population, including any exclusion criteria (4) the required input modalities and data-quality conditions. Third, the fallback mechanism needs to describe how the model will function (or escalate) when it encounters an error.

We retain a six-level scale ranging from zero (no automation) to five (full automation). As with the Glasgow Coma Scale, a given system is best characterised not by a single number but by a profile across the three domains - for example, clinical task-3 / operational domain-2 / Fallback-1, as systems are frequently advanced in one domain while immature in another. An agent with strong action capability but a vague "escalate to a human" fallback is the common, and potentially the most dangerous, real-world case. The divergence between the domains is itself a governance signal: a large gap between a high clinical-task level and a low fallback level identifies precisely where safeguards are required.

Every autonomous system deployment should therefore define two levels for each domain: a capability level (what the system can technically do, as validated) and an authorised level (what governance permits it to do in this deployment). The latter authorised level is the operational constraint; the capability level is what the system would do absent that constraint. Recording both, and monitoring the gap, is what operationalises the safeguards described - a system whose capability level has risen above its authorised level (through a model update, for instance) has entered a state that requires re-authorisation before the additional capability is used.

Level		Clinical Task Domain		Fallback mechanism	Operational Setting
		Discrete task execution	*Patient state monitoring & response*		
0	**No automation**	Clinician performs all clinical actions	Clinician monitors patient state and responds to all events	N/A (no automation)	N/A (no automation)
1	**Clinical assistance** AI recommends a specific action, but a clinician decides and acts.	AI recommends actions; Clinician decides and executes	Clinician monitors patient state and responds to all events	N/A - Clinician is in control	Any setting where AI output is advisory only
2	**Partial automation** AI executes defined clinical actions, but a clinician must remain ready to intervene at any moment.	AI executes some actions; others are performed by Clinicians.	Clinician monitors patient state and responds to all events; clinician must be actively available	Clinician actively supervises – when system fails, it halts; Clinician overrides and assumes full task	Defined clinical context; specified patient population with documented exclusion criteria
3	**Conditional automation** AI manages the complete clinical task within defined conditions, but a clinician must be available to take over if needed	AI executes the full defined clinical task, with Clinician available to be notified to take over.	AI detects patient state changes and responds; clinician must be reachable but not required to be immediately present	Clinician monitors passively, and must be researchable – if system fails, escalates and initiates handover to a clinician; holds last safe state pending clinician takeover	Defined clinical context; specified patient population with documented exclusion criteria
4	**High automation** The AI manages the clinical task and its own fail-safe response within defined conditions; no Clinician needed	AI executes the full defined clinical task, with no prerequisite for clinician availability	AI detects patient state changes and responds; no clinician readiness required	AI responsible, with Clinician available but does not need to be engaged - system executes safe-state protocol, including recovery.	Constrained deployment: single site or system type
5	**Full automation** The AI manages the entire clinical task, in any patient, in any condition, without the need for Clinicians	AI executes the full defined clinical task across all patient populations	AI detects patient state changes and responds; no clinician readiness required	Fully AI managed, no Clinician required- system executes safe-state protocol, including recovery, without needing Clinician intervention	**No patient, setting, or environmental restriction**

Table 1: A proposed framework for classifying autonomous agents in healthcare. Green reflect full autonomous operation. Level 5 is included for conceptual completeness only and represents a theoretical limit; no clinical deployment at this level is currently feasible.

Case example: AI-driven Sepsis Management

Sepsis management is a sequential decision problem in which clinicians titrate interventions such as fluids, vasopressors and antibiotics according to changes in a patient's clinical state. Given the non-deterministic nature of sepsis and the need for timely intervention, it is a process that may derive considerable benefit from AI-supported detection and management. Some clinical decision support systems have been described as an "AI clinician"(16), with the implications that these systems may one day be looked upon to increasingly support the delivery of sepsis care(17). With this in mind, we apply our framework to make explicit how Agentic sepsis management systems may look like across the 5 levels of automation (noting more advanced levels outline a hypothetical state) (Table 2).

Level	Descriptor	Clinical Task Domain (task execution & monitoring)	Fallback mechanism	Operational Setting
0	No automation	Clinician assesses the septic patient, decides on and titrates all fluids and vasopressors, and monitors response; no AI involvement.	N/A: no AI involvement	N/A: no AI involvement
1	Clinical assistance	At each timestep the AI recommends a fluid volume and vasopressor dose; the clinician reviews every recommendation, then prescribes and administers.	N/A: AI advisory only	Advice only; adult ICU. No execution authority.
2	Partial automation	AI directly titrates one component within pre-set bounds (e.g. vasopressor infusion to a mean-arterial-pressure target); the clinician sets fluid strategy and can override instantly. Clinician is actively supervising and must be readily available (e.g. bedside nurse).	When operating out of scope, (e.g. sensor dropout, or an out-of-range physiological state) AI halts titration, holds the last infusion rate, and immediately hands control to the Clinician	Defined adult ICU septic-shock cohort (e.g. Sepsis-3 definition); documented exclusions (e.g. no paediatric, no pre-specified comorbidity exclusions).
3	Conditional automation	AI manages the full resuscitation bundle (fluids, antibiotics and titrating vasopressors) to physiological targets Clinician must remain reachable but not immediately present (e.g. elsewhere the ICU)	On detecting out-of-scope condition, AI downgrades to a simpler, protocolised mode (e.g. a guideline-based bundle with safe vasopressor dose range) and continues management until a clinician is notified and take over.	Defined adult Sepsis-3 cohort; constrained deployment: single ICU or site.
4	High automation	AI manages the full resuscitation bundle, continuously monitors patient trajectory, and its operational boundaries. The responsible clinician is not involved in operation.	AI reverts to protocolised care, self-diagnoses the failure mode, and notifies the responsible clinician. Resumption of autonomous management requires clinician acknowledgment before the AI exits the safe state. The clinician need not be present or intervene directly.	Defined adult Sepsis-3 cohort; constrained deployment: single ICU or site.
5	Full automation (theoretical)	AI manages full sepsis resuscitation as per level 4, however can be done in any patient and setting	AI executes safe-state and troubleshooting, and resume autonomous management, without human notification.	No patient-population, acuity, or setting restriction

Table 2. The framework applied to an autonomous sepsis-resuscitation agent. Green reflects full autonomation of that domain.

Discussion

Our framework addresses a critical gap in how we describe autonomous systems in healthcare, and supports a shared understanding of the operational boundaries of this system when deployed in a real-world environment. This impacts how we evaluate, manage and autonomise AI systems for real-world deployments, as two systems with identical performance metrics (such as accuracy) may have significantly different governance and requirements if their specifications and fallback mechanisms differ.

A key contribution of our framework is providing a language to delineate the difference between what an agentic system is authorised to do within governance boundaries, as compared to what it is capable of. This distinction prevents autonomy creep, which we define as the incremental expansion of a system's operational scope without corresponding governance oversight. This occurs when a system is deployed for one purpose but is then used for adjacent purposes. Our framework guards against this risk by ensuring that autonomy levels are an explicit documented decision. If a health service wishes to increase an AI deployment from level 2 to level 3, the framework identifies what additional considerations need to be addressed.

Our framework also draws attention to the specific risks that are introduced by AI-driven autonomous systems. AI agent frameworks are generally structured as observe-decide-act loops(18). As a result, agentic designs will over emphasise specification of the action by the AI agent; the operating envelope or fallback mechanisms are not first-class concerns during development and need to be retrospectively fitted onto the agentic system. An overemphasis on the action domains can contribute to risk and undermine safety, as illustrated by the Uber Advanced technologies group fatality(19), where the self-driving vehicle system was not constrained to exclude unreliable conditions, and had poorly defined fallback mechanisms, which was only specified as “human intervention”.

The framework also implies a shift in workforce competency requirements associated with each level. At Level 3 and above, the human's role shifts from active operator to passive supervisor - a fundamentally different cognitive task that has been demonstrated to have an unanticipated impact on humans performance (11,20). Bainbridge's analysis of automation ironies(20) identified that this transition paradoxically degrades the very human performance it is designed to augment: reduced task engagement erodes procedural skill and situational awareness, such that the clinician who must intervene in a system failure is least prepared to do so at precisely the moment intervention is most consequential. This effect is already observed in AI clinical decision support systems, where clinicians assisted by AI to identify abnormalities on colonoscopy experienced a degree of de-skilling(21).

However, the competency frameworks required to support human supervision of autonomous AI systems does not yet exist in any mature form. Existing frameworks - whether procedural skill assessments, or supervised practice models - were designed to assess a clinician's ability to perform clinical tasks. They do not assess the distinct cognitive capabilities required for effective AI supervision: sustained vigilance under low-event-rate conditions, accurate mental model maintenance of a system whose internal state is not directly observable, and rapid mode-switching from supervisory to operational control under time pressure(21). The development of AI supervisory competency frameworks is a necessary complement to the

governance framework proposed here, and their absence represents a workforce readiness gap that health services should address before authorising deployment at Level 3 or above.

Our proposed taxonomy is intended to complement existing regulatory frameworks rather than replace them. Existing software as a medical device (SaMD) frameworks from regulating bodies globally , broadly govern systems that inform clinical decisions - systems that, in our taxonomy, operate at Level 2 and below, where a human clinician retains decision and action authority. These frameworks were not designed for systems that autonomously execute clinical tasks, monitor their own operating conditions, or manage their own failure responses. Systems where AI assumes primary operational responsibility (particularly at level 3 and above) - are more appropriately governed under higher-risk regulatory provisions, with the explicit specification of ODD constraints and fallback requirements providing the additional governance layer that existing SaMD instruments do not currently require.

Future Directions

Patient Agency

Patient values, preferences, and consent are not currently integrated into the proposed framework. This is a significant limitation for any deployment at Level 3 or above, where the AI takes operational responsibility for decisions may directly affect patients. Additional research is needed to explore the role of patient agency and how patient preferences should constrain or modify the AI's authorised operating scope - including the right of a patient to request human decision-making regardless of the system's technical capabilities.

Non-Determinism and Regulation

Unlike mechanical automation systems, AI agents incorporating large language model components are non-deterministic: the same input may produce different outputs, and performance may shift with model updates, infrastructure changes, or prompt modifications. This creates a fundamental challenge for the certification model inherited from SaMD regulation, which assumes a static, deterministic software package that can be tested once and deployed. The proposed framework does not resolve this challenge but makes it visible: any system operating at Level 3 or above must have continuous monitoring and drift detection capabilities to detecting performance degradation and must have pre-specified triggers for automatic reduction to a lower autonomy level. The development of standards for this continuous validation infrastructure is an urgent research and regulatory priority.

Validation and Generalisability

The framework is presented here as a conceptual proposal, illustrated through case studies constructed by the authors. It requires prospective validation across diverse health systems, clinical contexts, and AI system types. In particular, the level descriptors should be subject to inter-rater reliability testing: if the framework is to serve as governance infrastructure, independent reviewers should be able to assign consistent level ratings to a given system. The development of empirically validated rubrics for each level transition across common clinical AI use cases is a priority for future work.

Conclusion

Autonomous agentic AI systems are already being operationalised across industries and will reach health care soon. The speed of adoption is likely to outstrip the governance frameworks and institutional capability to manage them. The absence of a framework for describing autonomous systems, and specifying clear operating boundaries and fallback mechanisms, creates the conditions for patient harm.

Our proposed framework, adapted from the SAE autonomous driving framework, gives health services the means to make explicit, auditable, and comparable decisions about what autonomous systems are both able to do, as well as authorised to do. It shifts the lens from accuracy of a particular model to questions that determine real-world safety, including clarity around the operating conditions of autonomous systems and what happens when they reach their limit.

Further work is needed to understand generalisability of this framework, monitoring and regulation of autonomous systems, and to incorporate patient agency into decisions. However, these are not reasons to wait – the choice before healthcare leaders is whether the technology arrives into a governance vacuum, or a structure that can address, in some part, the unique risks of these systems. Our framework provides a conceptual foundation for the latter - introducing a shared language for deciding how to manage autonomous systems that may potentially transform healthcare delivery.

References

1. Liu F, Niu Y, Zhang Q, Wang K, Dong Z, Wong IN, et al. A foundational architecture for AI agents in healthcare. Cell Rep Med. 2025 Oct 21;6(10):102374. doi:10.1016/j.xcrm.2025.102374 PubMed PMID: 41015033; PubMed Central PMCID: PMC12629813.

2. Chen J, Wu H, Pang J, Wang Y, Zhang D, Sun C. Tool learning with language models: a comprehensive survey of methods, pipelines, and benchmarks. Vicinagearth. 2025 Nov 27;2(1):16. doi:10.1007/s44336-025-00024-x

3. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern Part Syst Hum. 2000 May;30(3):286–97. doi:10.1109/3468.844354 PubMed PMID: 11760769.

4. Hussein R, Zink A, Ramadan B, Howard FM, Hightower M, Shah S, et al. Advancing healthcare AI governance through a comprehensive maturity model based on systematic review. Npj Digit Med. 2026 Feb 11;9(236). doi:10.1038/s41746-026-02418-7

5. Haidemariam T. From the logic of coordination to goal-directed reasoning: the agentic turn in artificial intelligence. Front Artif Intell. 2025;8:1728738. doi:10.3389/frai.2025.1728738 PubMed PMID: 41602311; PubMed Central PMCID: PMC12833085.

6. Berkhout WEM, van Wijngaarden JJ, Workum JD, van de Sande D, Hilling DE, Jung C, et al. Operationalization of Artificial Intelligence Applications in the Intensive Care Unit: A Systematic Review. JAMA Netw Open. 2025 Jul 23;8(7):e2522866. doi:10.1001/jamanetworkopen.2025.22866

7. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020 Mar 25;368:m689. doi:10.1136/bmj.m689 PubMed PMID: 32213531.

8. World Health Organization. Classification of digital interventions, services and applications in health: a shared language to describe the uses of digital technology for health, 2nd ed [Internet]. Geneva: WHO; 2023 [cited 2026 Mar 20]. Available from: https://www.who.int/publications/i/item/9789240081949

9. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res. 2017 Nov 1;19(11):e8775. doi:10.2196/jmir.8775

10. Sheridan TB, Verplank WL. Human and Computer Control of Undersea Teleoperators [Internet]. Cambridge, MA: MIT Man-Machine Systems Laboratory;

1978 [cited 2026 Mar 19]. Report no.: ADA057655. Available from: https://apps.dtic.mil/sti/html/tr/ADA057655/

11. Wickens CD, Li H, Santamaria A, Sebok A, Sarter NB. Stages and Levels of Automation: An Integrated Meta-analysis. Proc Hum Factors Ergon Soc Annu Meet. 2010 Sep 1;54(4):389–93. doi:10.1177/154193121005400425

12. Beer JM, Fisk AD, Rogers WA. Toward a framework for levels of robot autonomy in human-robot interaction. J Hum-Robot Interact. 2014 Jul;3(2):74–99. doi:10.5898/JHRI.3.2.Beer PubMed PMID: 29082107; PubMed Central PMCID: PMC5656240.

13. SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE Standard J3016_202104. [Internet]. Warrendale, PA: SAE International; 2021 [cited 2026 Mar 19]. Available from: https://www.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles

14. Brooke L. Society of Automotive Engineering International [Internet]. 2016 [cited 2026 May 28]. U.S. DoT chooses SAE J3016 for vehicle-autonomy policy guidance - SAE International. Available from: http://articles.sae.org/15021/

15. Li K. Tesla is ordered to rename “Autopilot” after a California judge ruled that the EV-maker misled consumers. Business Insider [Internet]. [cited 2026 May 28]. Available from: https://www.businessinsider.com/judge-rules-tesla-misled-consumers-dmv-california-2025-10

16. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018 Nov;24(11):1716–20. doi:10.1038/s41591-018-0213-5

17. O’Reilly D, McGrath J, Martin-Loeches I. Optimizing artificial intelligence in sepsis management: Opportunities in the present and looking closely to the future. J Intensive Med. 2023 Nov 29;4(1):34–45. doi:10.1016/j.jointm.2023.10.001 PubMed PMID: 38263963; PubMed Central PMCID: PMC10800769.

18. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: Synergizing reasoning and acting in language models. In: Proceedings of the 11th International Conference on Learning Representations (ICLR 2023) [Internet]. Kigali, Rwanda: OpenReview; 2023 [cited 2026 May 30]. Available from: https://arxiv.org/abs/2210.03629

19. National Transportation Safety Board. Highway Accident Report: Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018 [Internet]. Washington, DC: NTSB; 2019 Nov [cited 2026 May 30]. Report HAR-19/03. Available from: https://trid.trb.org/View/1751168

20. Bainbridge L. Ironies of automation. Automatica. 1983 Nov 1;19(6):775–9. doi:10.1016/0005-1098(83)90046-8

21. Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol Hepatol. 2025 Oct 1;10(10):896–903. doi:10.1016/S2468-1253(25)00133-5 PubMed PMID: 40816301.

22. Abulibdeh R, Agyemang GO, Celi LA, Gorijavolu R, Kalema NL, Kleinlein R, et al. Who's really in the loop? Rethinking oversight in AI-assisted health care. The Lancet. 2026 Apr 30. doi:10.1016/S0140-6736(26)00204-7