nature medicine

Perspective

https://doi.org/10.1038/s41591-026-04438-y

AI-induced never-skilling in medical education

Received: 30 January 2026

Accepted: 30 April 2026

Published online: 22 May 2026

Check for updates

Yuhe Ke[1,2,3], Liyuan Jin[1,4], Jasmine Chiat Ling Ong[1,5], Arun J. Thirunavukarasu[6], Josip Car[7,8], Carol Y. Cheung[9], Yih Chung Tham[4,10,11], Daniel Shu Wei Ting[1,4], Marcus Eng Hock Ong[12,13], Scott Compton[14], Aditee Narayan[15], Pearse A. Keane[16,17], Tien Yin Wong[4,18,19], David W. Bates[20,21,22], Patrick Tan[23,24,25,26] & Nan Liu[1,12,27,28,29]

The integration of artificial intelligence (AI) into medical training is accelerating faster than the educational frameworks designed to govern it. This Perspective identifies a risk that has received insufficient attention: that trainees who rely on AI during the early formative years of clinical education may fail to develop the foundational reasoning skills that safe, independent practice requires. We refer to this as 'never-skilling', distinguishing it from deskilling in experienced clinicians and from mis-skilling, in which uncritical acceptance of AI errors leads trainees to internalize flawed clinical knowledge as fact. Although direct evidence from medical training is absent, the concern is grounded in established learning theory and supported by early empirical signaling from nonclinical settings. AI is not inherently harmful to learning; its educational impact depends on how and when it is introduced. We propose a three-phase competency-protective framework: establishing AI-independent baseline competency, building critical calibration through structured pedagogy, and integrating AI under supervision in medical training. This is a pedagogy research agenda that requires further empirical investigation to ultimately inform future policy recommendations.

Large language models (LLMs), computer vision systems and AI-based clinical decision-support systems have demonstrated substantial clinical value across medical specialties[1–4]. They have improved diagnostic accuracy, medication safety, workflow efficiency and access to specialist care[5,6]. These developments indicate that working effectively with AI will become a key competency for future physicians. Yet an important question is: how can AI be incorporated into medical education and early residency training without weakening the independent clinical reasoning needed for safe and resilient practice?

The implications extend beyond educational standards to patient safety, licensing and workforce stability. If trainees are unable to demonstrate clinical reasoning without AI assistance, licensing and credentialing processes may become misaligned. This could produce clinicians who function well with AI support but struggle when systems fail or when such tools are unavailable. At scale, such dynamics could increase supervision and liability burdens, contribute to a bifurcation of the workforce between AI-dependent and AI-independent clinicians, and widen disparities between healthcare systems with differing levels of AI infrastructure.

AI integration in medical training may introduce three distinct risks to competency development: deskilling, mis-skilling and never-skilling. Deskilling refers to the erosion of established competencies in experienced clinicians who increasingly rely on AI support. Mis-skilling represents a related but distinct phenomenon in which incorrect or biased AI outputs are adopted uncritically, reinforcing flawed patterns of clinical reasoning[7]. In contrast to these two phenomena, never-skilling refers to the failure to develop foundational competencies during formative training when AI substitutes a majority of the cognitive effort required to build independent clinical reasoning. For medical students and trainees, the central risk is therefore not the loss of previously acquired skills, but the failure to develop the capacity for independent clinical reasoning in the first place (Table 1).

A full list of affiliations appears at the end of the paper. e-mail: liu.nan@duke-nus.edu.sg

Table 1 | Core concepts and definitions

Concept	Definition	Example manifestations
Never-skilling	Failure to develop foundational clinical reasoning competencies during formative training due to excessive AI substitution for cognitive effort. This produces an inability to demonstrate AI-independent competence required for progression, graduation or licensure.	• Unable to generate differential diagnoses without AI prompting • Cannot formulate management plans in AI-restricted settings • Failure to meet AI-independent progression requirements • Primarily affects learners during initial skill acquisition
Deskilling	Degradation of established competencies in already trained clinicians due to prolonged AI reliance. Existing neural pathways weaken from disuse, but foundational architecture remains intact.	• Measurable performance decline in previously mastered tasks • Slower processing and reduced accuracy when unassisted • Baseline ability to manage cases through independent reasoning • Affects experienced clinicians with established baseline competency
Mis-skilling	Acquisition of incorrect clinical reasoning patterns through uncritical adoption of erroneous or biased AI outputs. Clinical schemas become contaminated by AI errors or bias, leading to systematic distortion of diagnostic or management reasoning.	• Confidently incorrect differential diagnoses aligned with AI output • Internalization of biased diagnostic associations (for example, race or gender bias) • Repeated reproduction of AI-generated errors in independent reasoning • Occurs across all training stages when AI outputs are accepted without verification
Metacognitive calibration	Accurate internal modeling of one's own competence boundaries and the alignment between subjective confidence and objective performance that enables appropriate help-seeking behavior.	• Confidence judgments match actual performance • Consulting when approaching competence limits • Detecting reasoning mistakes before implementation • Choosing cases matched to skill level; declining cases beyond capability
Automation bias	The tendency to over-rely on automated decision-support systems, accepting their outputs uncritically even when contradictory clinical evidence exists; susceptibility is heightened among novices who lack the foundational schemas needed to evaluate AI recommendations.	• Accepting an AI-generated diagnosis despite conflicting history or examination findings • Failing to independently verify AI recommendations before implementation • Reduced vigilance in monitoring AI outputs over repeated exposures • Disproportionate impact on early trainees versus experienced clinicians
False proficiency	A transient performance state in which AI-assisted clinical reasoning appears adequate during training, masking the absence of underlying independent competency that becomes evident only when AI support is withdrawn or unavailable.	• Strong performance on AI-enabled workplace assessments but poor results on AI-independent licensing examinations • Confidence in clinical decisions that collapses rapidly without AI support • Discrepancy between supervised (AI-rich) and unsupervised (AI-free) clinical performance • Difficulty articulating the reasoning pathway behind an AI-supported conclusion

These risks manifest differently across training stages. In medical school education, never-skilling may threaten progression if students fail to develop the independent reasoning required for advancement. In residency training, the concern is more subtle. Formal licensing examinations across many jurisdictions, typically high-stakes and standardized assessments, remain explicitly designed to assess unaided clinical reasoning. Yet clinical rotations and workplace-based assessments increasingly occur in environments where AI tools are readily available and commonly used. As a result, trainees may perform adequately in AI-enabled clinical environments while failing to demonstrate the independent competence required in AI-restricted, high-stakes assessments. Direct causal evidence linking AI exposure to erosion of clinical competency is still limited, although early signals are beginning to appear[8,9].

These concerns should be considered alongside the genuine benefits that AI offers in both medical education and clinical care. In educational settings, well-structured AI learning tools have been shown to improve examination performance[10] and may provide faster feedback cycles, broader case exposure and personalized practice at scale[11]. In clinical practice, AI systems have expanded access to diabetic retinopathy screening[12] and improved medication safety through decision-support tools[6,13]. Importantly, the success of these systems depends on clinicians who possess sufficient foundational expertise to critically evaluate AI outputs and override them when necessary. Yet adoption is advancing faster than educational preparation: two-thirds of physicians in the United States reported using AI in 2024 (ref. 14), while fewer than 15% of students and faculty report formal expertise[15]. Although AI is often framed as a benevolent 'copilot'[16,17], a copilot is only useful if trainees first learn how to be pilots. The conditions under which AI supports or undermines the development of clinical expertise therefore remain poorly defined. This is precisely the gap that our framework, outlined below, seeks to address.

This Perspective focuses specifically on the use of LLMs for point-of-care diagnostic reasoning during formative stages of medical training, particularly among medical students and early trainees. This group may be most vulnerable to the effects of AI substitution on the development of foundational clinical skills. We do not address the use of AI for administrative tasks, documentation or scientific discovery, as these domains involve different cognitive demands and learning processes.

Within this context, we present never-skilling as a conceptual risk model rather than an established empirical phenomenon. Under appropriate conditions, AI integration may redirect cognitive effort toward higher-order reasoning or support the development of new competencies, such as evaluating machine-generated outputs. However, longitudinal evidence tracking independent clinical competency in learners trained in AI-rich environments remains largely absent. This Perspective therefore proposes a precautionary framework aimed at preserving foundational clinical competence while supporting the safe and effective integration of AI into medical training.

Evidence base: theoretical grounding and early signals

AI and prior medical technologies

New technologies have repeatedly prompted concern about cognitive dependency in medicine. The introduction of imaging was said to displace physical examination skills. Electronic health records were criticized for eroding clinical memory. Calculators were believed to weaken numeracy. In most cases, these anxieties were not validated by evidence of lasting harm. Why should AI be different?

The concern is that AI differs from prior technologies in two respects relevant to skill acquisition. First, prior technologies shifted the type of cognitive work required but preserved its presence. A CT scan provides anatomical images that require expert anatomical

knowledge to interpret. Laboratory values demand integration with the clinical presentation. Electronic health records organize data but do not generate diagnostic conclusions. AI can, by contrast, execute the entire diagnostic chain autonomously. This is not a shift in cognitive work. It is a substitution for it.

Second, the timing of exposure differs[18]. Meshaka and Arthurs[19] note that imaging overreliance has been documented primarily in trained clinicians who already possess foundational skills. AI, however, is available from the first day of medical school, before any clinical reasoning architecture has formed. In this context, the question is not whether trainees who acquire skills then lose them, it is whether they acquire them at all.

This distinction matters for how we frame the problem. Some technology-induced dependencies in medicine are not harmful. Surgeons trained after the widespread adoption of laparoscopy were never taught to operate without it. They do not need to. The relevant question is not whether AI creates dependency in trainees—it likely will. The question is which dependencies are harmful and which represent rational delegation.

A provisional answer can be proposed. Competencies foundational for patient safety across all practice settings, including resource-limited ones, require independent mastery. These include clinical history taking, physical examination, diagnostic hypothesis generation and management planning under uncertainty. Competencies routinely supported by technology in all modern practice environments may be candidates for delegation. This distinction requires empirical investigation and should not be assumed.

Importantly, declining proficiency in some traditional skills may not constitute harm. For example, the precordial stethoscope was once routinely used to monitor heart and breath sounds during anesthesia, before capnography and pulse oximetry became widely available. Today the stethoscope is rarely used in routine anesthesia practice, having been largely replaced by continuous electronic monitoring[20]. Its decline reflects technological progress rather than a loss of clinical competence. The same logic applies to AI. Some forms of AI dependency may represent appropriate adaptation to new tools that improve efficiency and safety. Others may represent a genuine threat to patient safety when AI systems are unavailable or incorrect. Distinguishing between these possibilities is therefore critical and requires domain-specific empirical study.

Theoretical grounding

Educational science provides a principled basis for concern about AI substitution during formative training, independent of direct clinical evidence.

Desirable difficulties theory holds that the conditions that make learning harder in the short term tend to produce more durable long-term retention and transfer[21]. Making correct answers readily available removes the cognitive effort through which knowledge becomes consolidated. This predicts that AI answer delivery may produce apparent performance gains during training that do not translate to independent competency.

Deliberate practice theory identifies effortful, feedback-rich problem-solving as the mechanism through which expert reasoning develops[22]. When trainees use AI to generate clinical diagnoses rather than constructing them independently, they may accumulate experience hours without the cognitive investment those hours are designed to produce.

Cognitive load theory proposes that schema construction, the foundation of clinical reasoning, depends on effortful processing in working memory[23]. AI tools that bypass this processing may prevent schema construction rather than supporting it. The result may be trainees who are adept at interacting with AI outputs but have not built the underlying cognitive structures needed to reason without them.

The expertise reversal effect is particularly relevant[24]. Instructional support that benefits novice learners can be neutral or harmful for experts, because experts bring existing schemas that make additional guidance redundant. This predicts that AI assistance may affect novices and experienced clinicians differently, with potential for both augmentation and degradation depending on patterns of use. For novices, AI may substitute for the process of building foundational schemas. For experienced clinicians, AI may augment already established reasoning but also risk deskilling if over-relied upon. This provides theoretical grounding for a developmental stage-specific framework.

Early empirical signals

Direct evidence for never-skilling in clinical trainees is currently lacking, as the necessary longitudinal studies have not yet been conducted. The available evidence is indirect but suggestive. Budzyń et al.[8] reported that experienced endoscopists who routinely used AI-assisted colonoscopy demonstrated a 6% lower adenoma detection rate during subsequent unassisted procedures, consistent with potential deskilling in trained clinicians. Outside the clinical context, Kosmyna et al.[9] observed that sustained use of LLM-assisted writing was associated with weaker neural connectivity and poorer recall of self-generated content compared with unaided writing[9]. Taken together, these findings should be interpreted as preliminary signals rather than definitive evidence.

Experimental work in education offers more direct support for a plausible mechanism. In a randomized study of high school students, Bastani et al.[25] reported that unrestricted access to an AI tutor improved performance during assisted practice but was associated with a 17% relative reduction in normalized scores on a subsequent unaided, closed-book mathematics examination (mean difference: −0.054 on a 0–1 scale), with the greatest detriment observed among students with lower baseline achievement. Although this study is conducted outside medical training, it provides the clearest experimental illustration of how AI assistance may mask deficits in underlying competence while impairing durable independent skill development. Additional correlational evidence comes from Gerlich et al.[26], who surveyed 666 adults and reported a negative association between AI tool use and critical thinking ability, mediated by cognitive offloading and most pronounced in younger participants. Overall, these findings in non-clinical populations suggest that the concern for clinical education is plausible, but not yet demonstrated. Direct evidence for never-skilling in clinical trainees remains absent.

Together, these strands of evidence suggest that the timing of AI exposure during training may be critical. Clinicians who develop core competencies before the introduction of AI tools may experience occasional deskilling, whereas trainees exposed to AI during formative learning may risk never fully acquiring those competencies. The central prediction of the never-skilling hypothesis is that AI-assisted performance during training may create a period of false proficiency: apparent competency that depends on AI availability and does not persist when that support is withdrawn or becomes unavailable.

The case for AI as a learning accelerant

A plausible alternative view deserves substantive engagement. AI tools, when deliberately designed for learning rather than answer delivery, may accelerate foundational competency development. Adaptive AI tutors can expose learners to a wider range of clinical cases than any single training site can provide. They can deliver immediate feedback on reasoning processes and adjust case difficulty to the learner's current level. These are conditions that, the field of educational science, are associated with effective skill development.

Recent evidence supports this. Leong et al.[10] demonstrated that a specialized educational chatbot was perceived by residents as more useful and efficient than a standard chatbot for postgraduate examination preparation. Wang et al.[27] compared a Socratic AI tutor, which asked questions rather than delivered answers, to standard AI and found greater clinical reasoning engagement in the Socratic condition. These findings show that AI design shapes educational outcomes.

The risk of never-skilling is not a risk of AI per se, but a risk of using AI in answer-delivery mode during periods when cognitive architecture is still developing.

We propose a distinction that clarifies when AI may harm versus help (Table 2). Most AI tools deployed in clinical environments operate in answer-delivery mode. This is appropriate for their intended purpose. The concern is that trainees who use these tools during formative training may primarily encounter answer-delivery mode AI at precisely the developmental stage when learning-mode engagement would be most beneficial.

The mechanistic pathway: from AI substitution to competency risk

If never-skilling occurs, it likely operates through three interrelated mechanisms: competency acquisition failure, calibration deficit and metacognitive erosion. Each is a hypothesis, and we present them as a framework for investigation here. The degree of risk likely varies substantially across clinical domains and foundational clinical skills. Pattern-recognition specialties such as radiology, pathology and dermatology may face different AI substitution dynamics than procedural fields such as surgery or emergency medicine, where embodied skills and real-time decision-making are central. Even within clinical reasoning, some cognitive functions may be more susceptible to substitution than others. This variation should be incorporated into domain-specific research designs.

Core mechanisms of competency erosion

Competency acquisition failure. When AI systems supply answers during formative training, they may prevent the construction of clinical reasoning schemas. Schema formation requires effortful processing[23]. A student who correctly diagnoses diabetic ketoacidosis with AI assistance may not have built the pattern-recognition and pathophysiological reasoning networks allowing independent recognition of atypical presentations.

Productive failure theory predicts that struggling with problems before receiving solutions produces better long-term conceptual understanding than receiving solutions first[28]. Applied to clinical training, this suggests that AI systems delivering diagnoses before trainees have attempted them may undermine the cognitive conditions for durable learning, even when the delivered diagnosis is correct.

If cognitive architecture forms incompletely during early training, the resulting deficit may be difficult to detect. AI-assisted performance can appear adequate even when unaided performance is substantially impaired. The threshold at which such deficits become difficult to reverse is unknown. Neural plasticity may allow late remediation. Whether remediation is equally effective at later training stages is an open empirical question.

Calibration paradox. Effective AI oversight requires calibration and the capacity to know when to trust, question or override an AI output. This capacity does not arise from AI exposure alone. It develops through cycles of independent problem-solving, error recognition and feedback at the limits of one's knowledge. This is the very cycle disrupted by uncritical AI use. Trainees who encounter AI before building clinical reasoning schemas may develop familiarity with AI outputs without the foundation needed to evaluate them. The result is a cognitive moral hazard: epistemic responsibility is abdicated not because delegation is appropriate, but because the trainee cannot distinguish appropriate delegation from uncritical acceptance[29].

This vulnerability is amplified by a structural property of LLMs, which may express high confidence in incorrect conclusions, producing an illusion of expertise that expert clinicians can often detect and novices cannot[30]. Evidence across aviation, nuclear power and anesthesiology consistently demonstrates that automated monitoring reduces human vigilance, with novices more susceptible than experts[31].

Table 2 | Answer-delivery mode versus learning-mode AI

Answer-delivery mode (never-skilling risk)	Learning mode (potentially beneficial)
Provides diagnosis or differential diagnosis directly without explanations	Asks Socratic questions to prompt trainee reasoning
Delivers treatment recommendations without requiring trainee justification	Presents clinical scenarios for trainee to work through, with feedback on reasoning process
Removes cognitive effort from the diagnostic task	Adds deliberate cognitive friction to build durable schemas
Performance during AI use looks adequate; independent performance may be impaired	Short-term performance may be lower; long-term retention and transfer tend to be stronger
Deployed in clinical workflow contexts (for example, LLM as ward round assistant)	Deployed in deliberate practice and simulation contexts (for example, AI-facilitated problem-based learning)

In medical education, nonspecialists who stand to benefit most from AI support are simultaneously those most susceptible to overreliance[32,33].

Abdulnour et al.[7] crystallize the structural nature of this problem through a 'verify and trust' paradigm: clinicians must verify AI outputs against their own independent clinical reasoning and external evidence before placing trust in them. Verification demands a cognitive standard to verify against. A trainee without an independent clinical architecture cannot verify, only accept. Every AI interaction becomes a leap of faith that foundational competency is supposed to eliminate. Therefore, the calibration deficit is not merely a risk of trusting AI too much, but the absence of the infrastructure that verification requires (Fig. 1). We do not contend that human–AI collaboration is inherently detrimental. Evidence on collaborative performance is variable[34,35], and AI augmentation can meaningfully extend clinical capability in well-defined tasks. The essential skill lies in distinguishing contexts in which AI adds value from those in which it does not; and this too is a product of the independent reasoning that uncritical AI use undermines.

Metacognitive and professional identity deficit. Medical training aims to develop not just technical competency but metacognitive maturity: the capacity to monitor one's own reasoning, recognize uncertainty, tolerate ambiguity and maintain intellectual humility[36]. These attributes may develop through sustained engagement with difficult problems in which easy answers are unavailable[37].

When AI systems routinely supply diagnostic formulations and management decisions, they may reshape what trainees understand their professional role to be. Rather than developing as autonomous reasoners, trainees may come to view themselves as intermediaries who interpret and relay AI outputs[38]. Whether this shift in professional identity occurs, and whether it has measurable consequences for clinical performance, is an empirical question.

Consequences of never-skilling beyond medical education

Global health equity and the tiered physician systems. Unregulated AI integration could create physicians whose clinical competency is contingent on AI infrastructure. Such physicians may perform well in well-resourced settings where AI is reliable, but may struggle in resource-limited settings, during system failures or in emergency conditions[39,40]. On the other hand, AI technologies also hold considerable potential to improve healthcare equity. AI-based chest radiograph interpretation for tuberculosis diagnosis has expanded diagnostic access in settings without specialist radiologists. The equity argument cuts in two directions. AI may reduce global disparities by extending expertise to underserved settings. It may simultaneously create disparities if graduates from AI-rich training environments cannot practice without it. The net effect depends on implementation. AI integration that includes mandatory foundational competency safeguards could

https://doi.org/10.1038/s41591-026-04438-y

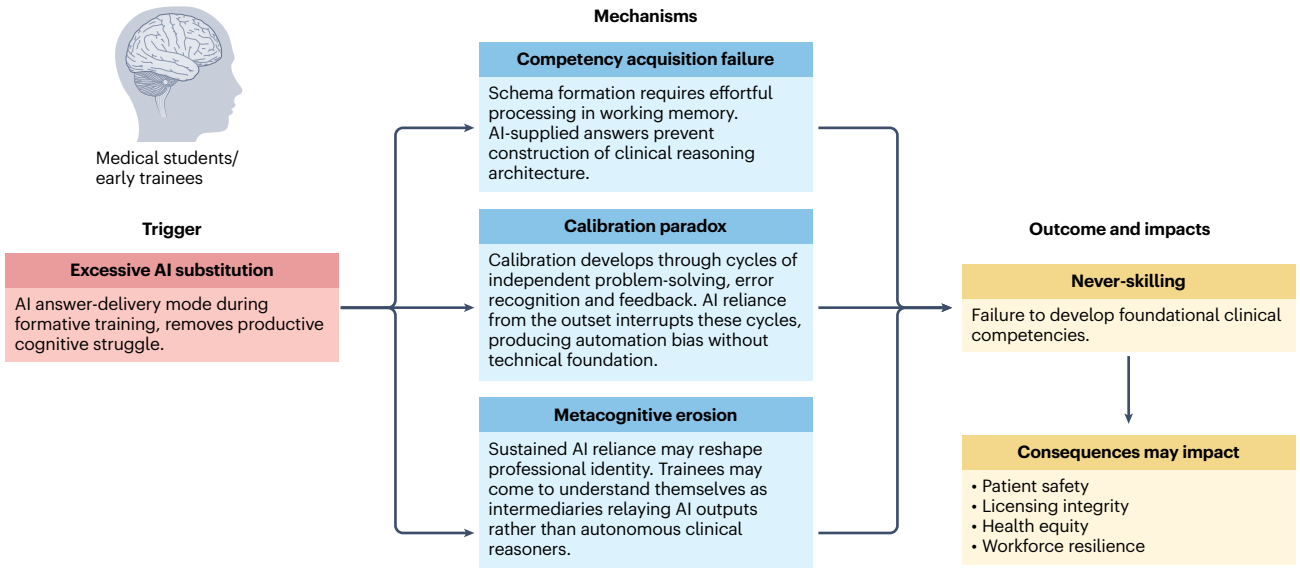

Fig. 1 | The mechanistic pathway from AI substitution to clinical competency risk. A proposed pathway through which excessive AI use during formative medical training may produce never-skilling. Left, the trigger is AI use in answer-delivery mode during the developmental period in which clinical reasoning schemas are being formed. This mode supplies correct outputs without requiring the cognitive effort through which durable competency develops. The population of primary concern is medical students and early trainees, for whom foundational clinical reasoning architecture has not yet been established. Center, three interrelated mechanisms are proposed through which AI substitution may impair competency development. Competency acquisition failure occurs when AI-supplied answers prevent the effortful schema construction that working memory-dependent learning requires. The calibration paradox describes the failure to develop accurate self-assessment of AI reliability, which depends on cycles of independent reasoning and error recognition that persistent AI use disrupts. Metacognitive erosion describes the possible reshaping of professional identity toward relay of AI outputs rather than autonomous clinical judgment. Right, the primary outcome of this pathway, if it occurs, is never-skilling: failure to develop foundational clinical competencies. Downstream consequences include risks to patient safety, licensing integrity, health equity and workforce resilience. All three mechanisms are presented as hypotheses requiring empirical testing, not established phenomena.

allow AI to reduce rather than exacerbate global disparities. AI integration without such safeguards risks producing a tiered workforce.

Governance and accountability gaps. Medical education lacks coherent governance frameworks for AI integration. Regulatory bodies have limited mechanisms to enforce competency safeguards[41]. Commercial AI vendors bear no responsibility for long-term educational outcomes. The US Food and Drug Administration explicitly excludes educational applications from its oversight of clinical decision support[42].

Current competency frameworks do not yet specify how clinical competence should be demonstrated in the presence of AI. The absence of explicit standards for AI-independent competency in clinical training creates a gap that will require proactive attention as AI use in clinical learning environments expands.

Healthcare workforce planning and leadership pipeline. A cohort of AI-dependent physicians would require reconfigured workforce planning. Physicians who need continuous AI infrastructure may require greater supervision where AI availability is less consistent, creating workforce shortfalls that aggregate headcount metrics do not capture.

The implications for academic medicine are speculative and difficult to evidence. The possibility that reduced foundational competency in one generation of trainees could affect the quality of supervision and teaching in subsequent generations warrants acknowledgement as a downstream concern, while recognizing that this causal chain is unproven.

Global physician mobility and credential recognition. Medical licensure assumes transferable competency. If competency becomes contingent on AI infrastructure, credential portability may be affected. Medical licensing bodies may eventually need to distinguish credentials validated under AI-assisted conditions from those validated under AI-independent conditions. This is a future concern the field should begin to anticipate.

A framework for responsible AI integration

AI clinical tools offer genuine, evidence-based benefits. When used by appropriately trained clinicians, AI can augment diagnostic accuracy, reduce errors, improve efficiency and extend specialist expertise to underserved settings. The aim of this framework is not to restrict AI use. It is to ensure clinicians first develop the foundational competencies that make effective, critical AI use possible.

We present this as a hypothesis-driven, precautionary framework, not a policy mandate. Its effectiveness has not been empirically tested. It should be implemented as a series of pilot studies with rigorous evaluation, not as a universal standard.

Proposed three-phase framework

Phase 1: Foundational competency (AI-free mode). Formal licensing examinations in most jurisdictions already require AI-independent clinical reasoning. The concern this phase addresses is not formal examinations but the broader clinical training environment, where AI tools are increasingly present during case-based learning, ward rounds and problem-solving exercises.

Phase 1 proposes that early clinical training should include explicitly structured periods of AI-free problem-based learning. This does not mean prohibiting AI from all educational activities. AI can serve as a study aid, a practice question generator or a simulated patient—without substituting for clinical reasoning. The critical distinction is between AI delivering answers the trainee should be constructing, and AI facilitating the practice of reasoning the trainee is developing independently.

Implementing AI-free learning environments in today's world involves practical challenges. Medical students carry smartphones with access to LLMs, and monitoring AI use is not feasible. What is feasible is designing assessment tasks that reveal AI-independent competency: oral examinations, observed clinical encounters and structured cases requiring explanation of reasoning rather than provision of final answers. These approaches already exist in medical education. This

https://doi.org/10.1038/s41591-026-04438-y

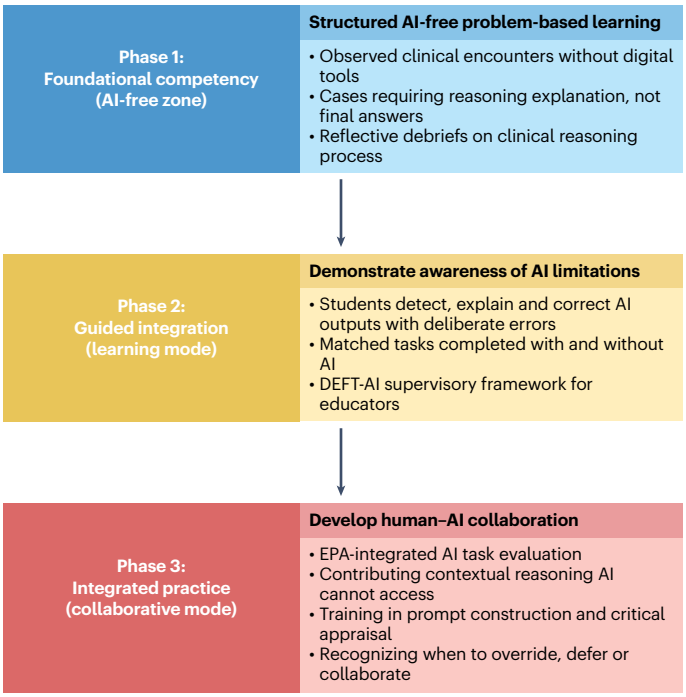

Fig. 2 | A three-phase framework for the responsible integration of AI into medical education. A proposed sequential framework in which progression between phases requires documented competency at the preceding level. The key metaphor reflects the prerequisite logic: each phase unlocks the next only when a defined competency threshold is met. Phase 1 establishes AI-independent clinical reasoning through structured problem-based learning, observed clinical encounters and assessment tasks requiring explicit reasoning rather than final answers. No AI assistance is permitted in answer-delivery mode during this phase. Phase 2 introduces AI through adversarial pedagogy, in which students encounter AI-generated clinical reasoning containing deliberate errors and are assessed on their ability to detect, explain and correct them. Matched tasks completed with and without AI develop calibrated reliance. The DEFT-AI supervisory framework provides educators with a structured tool for converting incidental AI use into deliberate teaching moments[7]. Phase 3 applies primarily to residency training and develops the specific competencies required for effective human–AI collaboration, including EPA-integrated task evaluation, contributing contextual clinical reasoning that AI cannot access, prompt construction and adaptive strategy selection. Each phase is a hypothesis requiring prospective evaluation rather than an established intervention.

framework proposes making AI-independent competency demonstration an explicit and documented milestone before progression to AI-integrated training.

An analogy can be drawn with aviation training, where pilots must demonstrate manual flight proficiency alongside the use of autopilot systems[31]. The comparison is not exact, but it illustrates a broader principle: foundational competence should precede reliance on technological assistance.

Phase 2: Guided integration (learning mode). With baseline competency established, phase 2 introduces AI through pedagogy structured around calibration: knowing when to trust, question or override AI outputs. The core approach is adversarial. Students encounter AI-generated clinical reasoning containing deliberate errors and are assessed on their ability to detect, explain, and correct them. Multi-agentic architectures simulating competing clinical perspectives can broaden the range of error types students encounter[43]. Embedded errors should reflect documented AI failure modes specific to the clinical domain, including anchoring bias, failure to flag contraindications and misinterpretation of laboratory values in context. Validation by clinical experts and iterative pilot testing are prerequisites before deployment as curriculum.

A recognized limitation is that training on clearly labeled, simulated errors may not transfer to real clinical settings, where AI failures are often subtle, unlabeled and embedded within otherwise plausible reasoning. Addressing this limitation requires the deliberate design of realistic adversarial cases. AI developers and clinical experts should collaborate to construct scenarios in which errors are distributed across otherwise coherent clinical arguments, mimicking the ambiguity encountered in practice. The objective is not simply to teach trainees to reject incorrect outputs, but to cultivate the ability to interrogate AI reasoning, identify points of uncertainty and justify when its recommendations should be accepted, modified or overridden.

Adversarial sessions should therefore be paired with calibration exercises in which students complete matched clinical tasks with and without AI assistance and then compare outcomes. These reflective cycles are the proposed mechanism through which calibrated reliance develops, rather than reflexive trust or reflexive skepticism. Learning objectives include identifying where AI adds genuine diagnostic value, recognizing where it introduces error and building an accurate model of one's own performance relative to AI across clinical domains.

At the supervisory level, the DEFT-AI framework (Diagnosis, Evidence, Feedback, Teaching, and recommendation for AI use) provides educators with a practical structure for converting incidental AI encounters into deliberate teaching moments[7]. The framework guides a sequential conversation when an educator observes a learner using AI: establishing what AI was used and how (diagnosis/discussion/discourse), probing how outputs were verified (evidence), prompting the learner to reflect on their own AI engagement (feedback), delivering focused instruction on appropriate use (Teaching) and providing learner-specific guidance for subsequent interactions (recommendation for AI engagement).

Phase 3: Integrated practice (collaboration mode). Phase 3 applies primarily to residency training, where entrustable professional activities (EPAs) provide the existing competency-based framework into which AI-integrated task evaluation can be incorporated. Progression should require documented competency at both the AI-independent level established in phase 1 and the calibration level developed in phase 2 (Fig. 2).

The goal of phase 3 is not to supervise AI. The clinical value that a trainee adds when working alongside AI may not lie in checking AI outputs. It lies in contributing contextual reasoning that AI cannot access: patient-reported history, physical examination findings and individual patient values. Phase 3 should therefore develop the specific skills through which clinicians contribute value beyond what AI provides alone, including recognizing when to override AI, when to defer to it, and when collaboration adds versus subtracts value relative to independent judgment.

Phase 3 should also include training in prompt construction. The quality of an LLM output depends substantially on the quality of the prompt that generates it[7]. Vague or leading prompts produce responses that are less accurate and potentially sycophantic, mirroring the questioner's prior beliefs rather than providing independent clinical evaluation. Asking AI to explain its reasoning before accepting an output is a teachable skill; it transforms passive acceptance into active appraisal and is directly applicable to safe clinical AI use.

Framework implementation challenges

The framework faces four practical barriers that any implementation program must address.

Barrier 1: Fragmentation. Medical schools, licensing boards, accreditation bodies and specialty certification organizations would each need to recognize AI-independent competency as a formal standard. No such coordination currently exists. Establishing it would require multilateral agreement across bodies with different governance structures, timelines and incentive frameworks.

https://doi.org/10.1038/s41591-026-04438-y

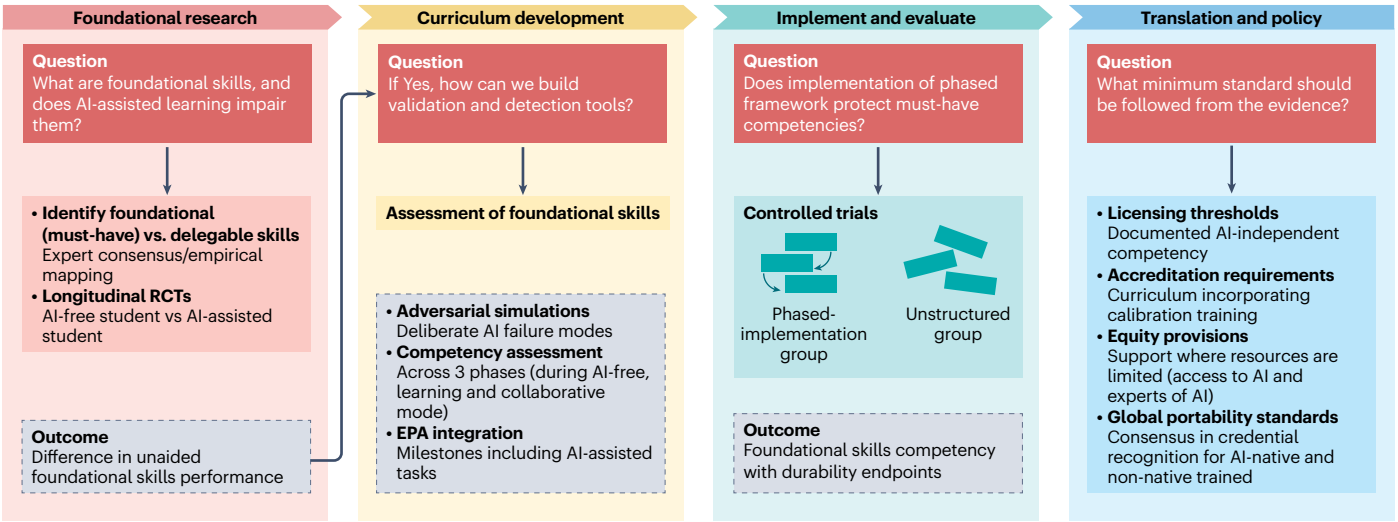

Fig. 3 | A translational research framework for testing the hypothesis that AI use during formative training impairs independent clinical competency. The figure traces a single investigative thread across four sequential research agendas, each building on the evidence generated by the preceding stage. Agenda 1 (foundational research) poses the primary empirical question: how does AI-assisted learning impair foundational skills? The study design should be first to get expert consensus on what are considered foundational/must-have skills, followed by a longitudinal randomized controlled trial (RCT) comparing AI-assisted and AI-free students in these skills. If the answer is yes, then agenda 2 (curriculum development) asks how detection and prevention tools can be built and validated if the foundational hypothesis is supported. Agenda 3 (implementation and evaluation) tests whether the phased framework, when implemented, protects must-have competencies through controlled trials comparing a phased-implementation group against an unstructured group, with competency durability as the primary endpoint. Agenda 4 (translation and policy) converts trial evidence into actionable standards across four domains.

Barrier 2: Faculty capacity. Approximately 9% of medical faculty report AI expertise[15]. Adversarial pedagogy and calibration training require educators who can design clinically plausible AI error cases, facilitate reasoning-focused debriefs and apply supervisory frameworks such as DEFT-AI at the point of care. These are skills that require dedicated development programs, which are themselves resource intensive.

Barrier 3: Resource inequity. Competency profiling across four AI-access levels requires technical infrastructure that may be disproportionately difficult to implement in institutions with limited IT support. If mandatory requirements are introduced without equity provisions, they risk disadvantaging the resource-limited institutions and student populations that would benefit most from competency safeguards.

Barrier 4: Competency detection. AI-assisted training may produce graduates who satisfy initial competency thresholds under AI-enabled conditions but show steeper performance decline when AI is withdrawn. Standard assessments taken at a single time point cannot detect this fragility. Identifying it requires longitudinal tracking under varying AI availability conditions throughout and after training.

These barriers are substantial but not insurmountable, and they are not unique to this framework. Competency-based medical education has confronted comparable implementation challenges across decades of reform[44]. The appropriate response is not to defer action pending resolution. It is to pilot the framework as a research program with embedded evaluation, allowing iterative refinement as evidence accumulates.

Research agenda

The framework rests on hypotheses that require empirical testing before any component can be recommended as standard practice. Four research priorities follow from this, roughly translational in sequence (Fig. 3).

Agenda 1: Foundational research. The foundational questions concern developmental timing and domain specificity. What minimum duration of AI-free problem-based learning is required for durable unaided competency? Which components of clinical reasoning are most susceptible to AI substitution, and which traditionally taught skills genuinely require independent mastery rather than representing candidates for rational delegation? These questions require prospective cohort studies comparing AI-native and traditionally trained learners on AI-independent competency measures at multiple time points.

A rigorous test of the never-skilling hypothesis would require a longitudinal randomized study comparing a structured AI-free curriculum against defined levels of AI-assisted learning, with clearly specified curricula in each arm. Students could be assigned to structured curricula with differing degrees of AI assistance in problem-based learning and clinical reasoning tasks. Outcomes should include both AI-independent reasoning performance and performance in AI-enabled environments, assessed at graduation and during follow-up after training. Given the educational and ethical implications, such a study would be most appropriately designed as an equivalence or non-inferiority trial with prespecified margins, longitudinal assessment of competency retention and interim monitoring safeguards.

Agenda 2: Curriculum development. The curriculum development agenda translates foundational findings into validated tools: assessments spanning all four competency levels in Table 3, a library of clinically representative adversarial cases reflecting documented AI failure modes and curriculum frameworks integrating AI training into existing evidence-based medicine courses. Each requires validation by clinical experts before deployment.

Agenda 3: Implement and evaluate. Implementation trials should compare phased versus unstructured AI integration using primary endpoints of AI-independent performance at training completion and competency durability at 6 to 12 months. Secondary endpoints should include calibration metrics, transfer to novel cases and workplace performance indicators in settings with varying AI availability.

Agenda 4: Translation and policy. The policy translation agenda converts evidence into actionable standards. Research findings should

Table 3 | Four-level competency framework[a]

Level	AI access	What is measured	Example tasks	Considerations
Level 0: Full independence	None	Unaided clinical reasoning capacity; establishes individual performance baseline.	Generate a differential diagnosis for an assigned case using no digital tools or references.	Scores at this level should be benchmarked against validated assessments (for example, USMLE-equivalent licensing items, OSCE scores) and treated as a baseline requiring longitudinal follow-up, not a certification of durable competency.
Level 1: Reference-augmented reasoning	Information retrieval only	Ability to integrate reference knowledge while maintaining independent reasoning architecture.	Generate a differential diagnosis for the same case with access to reference databases only; no AI-generated conclusions permitted.	Distinguishes the trainee's reasoning process from simple knowledge retrieval. Does not yet test collaboration with systems that generate conclusions.
Level 2: Adversarial calibration	Diagnostic AI with embedded errors	Error detection capacity; metacognitive calibration; ability to override AI with independent reasoning. Embedded errors must be validated by domain experts to ensure they reflect domain-specific AI failure modes.	Review an AI-generated differential for the same case; identify, explain and correct the embedded errors.	Transfer validity is a recognized limitation. Performance detecting clearly labeled errors in simulated AI outputs may not transfer to real clinical settings where AI failures are unlabeled, subtle and intermixed with plausible reasoning under time pressure. This level should be treated as an experimental intervention pending evaluation of transfer effects, not an established assessment modality.
Level 3: Supervised AI collaboration	Full AI access with documentation of decision attribution	Human–AI collaborative performance; ability to identify what the clinician contributes beyond AI output; calibrated reliance. Mapped to EPAs requiring documented independent competency before AI integration.	Complete the same case with unrestricted AI access; document each recommendation accepted, modified or overridden with explicit reasoning.	EPA frameworks provide a theoretically grounded structure for progression, but face documented implementation challenges including variable assessor calibration, inconsistent entrustment decisions, and trainee anxiety about milestone tracking. EPA integration at this level should be treated as a structural hypothesis requiring an institution-level pilot study and should not assume EPA implementation fidelity.

USMLE, United States Medical Licensing Examination; OSCE, Objective Structured Clinical Examination. [a]This competency schema is a preliminary construct and has not been psychometrically validated. Before operational use in formal assessment, it requires: (a) alignment with established measurement frameworks including competency-based medical education standards[45]; (b) assessor calibration studies; (c) evidence that each level discriminates between trainees at different stages of genuine competency development, not merely familiarity with the task format; and (d) evaluation of transfer validity for level 2 in real clinical environments.

inform minimum competency thresholds for licensing examinations and provide the evidence base for accreditation requirements governing AI literacy in medical training. Implementation science methods are needed to identify adoption barriers across institutional contexts and enable iterative improvement as both AI capabilities and educational evidence evolve.

Limitations

This Perspective has important limitations. Direct causal evidence linking AI exposure during training to competency failure in medical trainees does not exist. The empirical evidence cited is indirect and predominantly nonclinical. Never-skilling is presented as a risk model, not an established phenomenon.

The prevalence, severity and reversibility of any AI-induced competency deficits are unknown. Properly scaffolded AI integration may enhance rather than impair foundational skill development. The conditions under which each outcome occurs remain to be defined empirically. The three-phase framework described above is untested. Its effectiveness, feasibility and unintended consequences have not been evaluated. Pilot implementations with rigorous outcome measurement should precede widespread adoption. Recommendations should be understood as hypotheses for investigation, not established best practices. Lastly, this analysis reflects the current state of AI and medical education at the time of writing. Both are evolving rapidly. Specific recommendations may require revision as evidence accumulates.

Conclusion

The risk of never-skilling cannot yet be confirmed in clinical trainees, but it is theoretically grounded and consistent with early empirical signals from adjacent fields. Importantly, it is structurally difficult to detect and reverse once established at cohort scale. Hence, the risks of never-skilling warrant attention before the evidence of harm is conclusive.

This Perspective does not argue against AI in medical education. AI tools deployed in learning mode, with structured feedback, scaffolded difficulty and explicit reasoning demands, may accelerate foundational skill development. The concern is about sequencing. When AI is introduced in answer-delivery mode during the developmental period in which clinical reasoning schemas are being formed, it may substitute for the cognitive effort that schema formation requires. The goal is not to restrict AI; it is to ensure that trainees first develop the independent competency that makes critical AI use possible.

Declaration of AI usage

During manuscript preparation, we used generative AI tools (ChatGPT-4o, Gemini 3 and Claude 4.5) solely for language refinement. These tools were not used to generate or analyze original data or to draw scientific conclusions. All AI-assisted content was carefully reviewed and validated by the authors, who retain full responsibility for the manuscript.

References

1. Bajwa, J., Munir, U., Nori, A. & Williams, B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc. J.* **8**, e188–e194 (2021).
2. Teo, Z. L. et al. Generative artificial intelligence in medicine. *Nat. Med.* **31**, 3270–3282 (2025).
3. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
4. Wu, J. et al. Vision-language foundation model for 3D medical imaging. *NPJ Artif. Intell.* **1**, 17 (2025).
5. Ong, J. C. L. et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. *Cell Rep. Med.* **6**, 102323 (2025).
6. Ke, Y. H. et al. Clinical and economic impact of a large language model in perioperative medicine: a randomized crossover trial. *NPJ Digit. Med.* **8**, 462 (2025).

7. Abdulnour, R. -E. E., Gin, B. & Boscardin, C. K. Educational strategies for clinical supervision of artificial intelligence use. *N. Engl. J. Med.* **393**, 786–797 (2025).
8. Budzyń, K. et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol. Hepatol.* **10**, 896–903 (2025).
9. Kosmyna, N. et al. Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. Preprint at https://arxiv.org/abs/2506.08872 (2025).
10. Leong, Y. H., Nambiar, L., Tay, V. Y. J., Lie, S. A. & Yuhe, K. Feasibility of a specialized large language model for postgraduate medical examination preparation: single-center proof-of-concept study. *JMIR Form. Res.* **9**, e77580 (2025).
11. Klimova, B. & Pikhart, M. Exploring the effects of artificial intelligence on student and academic well-being in higher education: a mini-review. *Front. Psychol.* **16**, 1498132 (2025).
12. Diaz, E. A. et al. Diabetic retinopathy screening among federally qualified health center patients using point-of-care AI: DRES-POCAI: a trial protocol: DRES-POCAI: a trial protocol. *JAMA Netw. Open* **8**, e2538114 (2025).
13. Pak, A. et al. Mixed methods evaluation of a clinical decision support system to reduce variation in healthcare. *NPJ Digit. Med.* **8**, 781 (2025).
14. American Medical Association. 2 in 3 physicians are using health AI—up 78% from 2023. https://www.ama-assn.org/practice-management/digital-health/2-3-physicians-are-using-health-ai-78-2023 (2025).
15. Tufts University School of Medicine. How medical faculty and students are using AI today. https://medicine.tufts.edu/news-events/news/how-medical-faculty-and-students-are-using-ai-today (2025).
16. Ke, Y. H. et al. Real-world deployment and evaluation of PEri-operative AI CHatbot (PEACH): a large language model chatbot for peri-operative medicine. *Anaesthesia* **81**, 62–71 (2025).
17. Ong, A. Y. et al. Flight rules for clinical AI: lessons from aviation for human–AI collaboration in medicine. *NPJ Digit. Med.* **9**, 201 (2026).
18. Lea, A. S. Cognitive aids, artificial intelligence, and deskilling in medicine: the history of an enduring anxiety. *NEJM AI* **3**, 1 (2025).
19. Meshaka, R. & Arthurs, O. J. Are we too reliant on medical imaging?. *Br. J. Hosp. Med.* **83**, 1–3 (2022).
20. Litman, R. S. et al. Monitoring. in *Smith's Anesthesia for Infants and Children* 8th edn (eds Davis, P. J., Cladis, F. P. & Motoyama, E. K.) 322–343, https://doi.org/10.1016/b978-0-323-06612-9.00011-0 (Elsevier, 2011).
21. Bjork, E. L. & Bjork, R. A. Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. in *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (ed. Gernsbacher, M. A.) 56–64 (Worth Publishers, 2011).
22. Ericsson, K. A., Krampe, R. T. & Tesch-Römer, C. The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* **100**, 363–406 (1993).
23. Sweller, J. Cognitive load during problem solving: effects on learning. *Cogn. Sci.* **12**, 257–285 (1988).
24. Sweller, J., Ayres, P. & Kalyuga, S. The expertise reversal effect. in *Cognitive Load Theory* Vol. 1, 155–170, https://doi.org/10.1007/978-1-4419-8126-4_12 (Springer, 2011).
25. Bastani, H. et al. Generative AI without guardrails can harm learning: evidence from high school mathematics. *Proc. Natl. Acad. Sci. USA.* **122**, e2422633122 (2025).
26. Gerlich, M. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies* **15**, 6 (2025).
27. Wang, A. et al. Generative AI for medical education: insights from a case study with medical students and an AI tutor for clinical reasoning. in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* 1–8 https://doi.org/10.1145/3706599.3721208 (ACM, 2025).
28. Steenhof, N., Woods, N. N., Van Gerven, P. W. M. & Mylopoulos, M. Productive failure as an instructional approach to promote future learning. *Adv. Health Sci. Educ. Theory Pract.* **24**, 739–749 (2019).
29. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **19**, 121–127 (2012).
30. Griot, M., Hemptinne, C., Vanderdonckt, J. & Yuksel, D. Large language models lack essential metacognition for reliable medical reasoning. *Nat. Commun.* **16**, 642 (2025).
31. Ruskin, K. J., Corvin, C., Rice, S. C. & Winter, S. R. Autopilots in the operating room: safe use of automated medical technology. *Anesthesiology* **133**, 703–716 (2020).
32. Braarud, P. Ø Measuring cognitive workload in the nuclear control room: a review. *Ergonomics* **67**, 849–865 (2024).
33. Fletcher, G. et al. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br. J. Anaesth.* **90**, 580–588 (2003).
34. Vaccaro, M., Almaatouq, A. & Malone, T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* **8**, 2293–2303 (2024).
35. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. Combining human expertise with artificial intelligence: experimental evidence from radiology. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.4505053 (2023).
36. Fletcher, L. & Carruthers, P. Metacognition and reasoning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1366–1378 (2012).
37. Ainge, L. E., Edgar, A. K., Kirkman, J. M. & Armitage, J. A. Developing clinical reasoning along the cognitive continuum: a mixed methods evaluation of a novel Clinical Diagnosis Assessment. *BMC Med. Educ.* **25**, 31 (2025).
38. Morley, J. et al. The ethics of AI in health care: a mapping review. *Soc. Sci. Med.* **260**, 113172 (2020).
39. European Society of Medicine. Bridging global health AI divide with local wisdom. https://esmed.org/bridging-global-health-ai-divide-with-local-wisdom/ (2025).
40. Monteith, S. et al. Artificial intelligence and deskilling in medicine. *Br. J. Psychiatry* https://doi.org/10.1192/bjp.2025.10496 (2026).
41. Wen, X. & Thamotharampillai, T. When is it ethically defensible for a medical practitioner to deviate from clinical practice guidelines? *Ann. Acad. Med. Singap.* https://doi.org/10.47102/annals-acadmedsg.2025189 (2025).
42. US Food and Drug Administration. Software as a medical device (SaMD) https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd (2025).
43. Ke, Y. et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *J. Med. Internet Res.* **26**, e59439 (2024).
44. Ten Cate, O. Competency-based postgraduate medical education: past, present and future. *GMS J. Med. Educ.* **34**, Doc69 (2017).
45. Shah, N., Desai, C., Jorwekar, G., Badyal, D. & Singh, T. Competency-based medical education: an overview and application in pharmacology. *Indian J. Pharmacol.* **48**, S5–S9 (2016).

Author contributions

Y.K. and N.L. conceived the Perspective and designed the framework. Y.K. and J.L. performed the primary literature search and drafted the initial manuscript. J.C.L.O., A.J.T., J.C., C.Y.C., Y.C.T. and D.S.W.T. provided subject matter expertise on AI implementation and clinical integration. M.E.H.O., S.C., A.N. and P.A.K. contributed to the sections

on medical education and health systems policy. T.Y.W., D.W.B., P.T. and N.L. provided critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This work was supported by the Duke-NUS Signature Research Program funded by the Ministry of Health, Singapore. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Health.

Competing interests

P.A.K. is a cofounder of Cascader Ltd. and has acted as a consultant for Skleo Health, insitro, Retina Consultants of America, Roche, Boehringer-Ingelheim and Bitfount and is an equity owner in Big Picture Medical. P.A.K. has received speaker fees from Zeiss, Thea, Apellis and Roche, grant funding from Roche and travel support from Bayer and Roche, and has attended advisory boards for Topcon, Bayer, Boehringer-Ingelheim and Roche. C.Y.C. is a cofounder of i-Cognition Sciences Ltd. and InnoEye Focus Ltd. All other authors declare no competing interests.

Additional information

Correspondence should be addressed to Nan Liu.

Peer review information *Nature Medicine* thanks Jakob Kather, Pranav Rajpurkar and Shinjini Kundu for their contribution to the peer review of this work. Primary Handling Editor: Karen O'Leary, in collaboration with the *Nature Medicine* team.

[1]Duke-NUS AI + Medical Sciences Initiative, Duke-NUS Medical School, Singapore, Singapore. [2]Department of Anaesthesiology, Singapore General Hospital, Singapore, Singapore. [3]Data Science and Artificial Intelligence Lab, Singapore General Hospital, Singapore, Singapore. [4]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [5]Division of Pharmacy, Singapore General Hospital, Singapore, Singapore. [6]International Centre for Eye Health, London School of Hygiene and Tropical Medicine, London, UK. [7]King's Population Health Institute, King's College, London, UK. [8]School of Life Course & Population Health Sciences, King's College London, London, UK. [9]Department of Ophthalmology and Visual Sciences, Chinese University of Hong Kong, Hong Kong SAR, China. [10]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [11]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [12]Pre-hospital & Emergency Research Centre, Health Services Research & Population Health, Duke-NUS Medical School, Singapore, Singapore. [13]Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore. [14]Office of Education, Duke-NUS Medical School, Singapore, Singapore. [15]Department of Pediatrics, Duke University School of Medicine, Durham, NC, USA. [16]Institute of Ophthalmology, University College London, London, UK. [17]NIHR Biomedical Research Centre at Moorfields, Moorfields Eye Hospital NHS Foundation Trust, London, UK. [18]Beijing Visual Science and Translational Eye Research Institute (BERI), Beijing Tsinghua Changgung Hospital, Tsinghua Medicine, Tsinghua University, Beijing, China. [19]School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China. [20]Harvard Medical School, Boston, MA, USA. [21]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [22]Department of Health Policy and Management, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [23]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore. [24]SingHealth Duke-NUS Institute of Precision Medicine, Singapore, Singapore. [25]Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore, Singapore. [26]Precision Health Research, Singapore, Singapore. [27]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. [28]Centre for Biomedical Data Science, Duke-NUS Medical School, Singapore, Singapore. [29]NUS Artificial Intelligence Institute, National University of Singapore, Singapore, Singapore. ✉e-mail: liu.nan@duke-nus.edu.sg