EDITORIAL

Rethinking AI in Medical Imaging: Assumptions, Reality, and Reframing; Towards Physician-Centered Agentic Reasoning

Arman Rahmim [1,2,3*], Nourhan Bayasi [1,2], Xiaoxiao Li [4,5], Babak Saboury [1,6], Fereshteh Yousefirizi [1,2]

[1] *Department of Basic and Translational Research, BC Cancer Research Institute, Vancouver, BC, Canada*

[2] *Department of Radiology, University of British Columbia, Vancouver, BC, Canada*

[3] *Departments of Physics and Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada*

[4] *Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada*

[5] *Vector Institute, Canada*

[6] *Institute of Nuclear Medicine, Bethesda, MD, USA*

*Corresponding Author: Arman Rahmim

Email: arman.rahmim@ubc.ca

Received: 20 September 2026 / Accepted: 27 September 2026

PROOF

Abstract

Medical imaging has served as primary proving ground for clinical Artificial Intelligence (AI), yet a decade of intense research has not translated into proportionate bedside impact. We argue that this gap is not primarily a product of insufficient algorithmic performance, inadequate regulation, or lim¬ited explainability. Rather, it reflects a structural misalignment, between how AI systems are designed and evaluated, and how clinical decisions are made. This work identifies 8 interconnected dimensions of this misalignment: 1) the dominance of pixel-only models in a multimodal clinical world; 2) the erosion of physician trust through opaque and inflexible systems; 3) the unfulfilled promise of founda¬tion models in data-sparse medical domains; 4) the persistent bottleneck of non-shareable under-curated datasets; 5) the gap between validated algorithms and deployable clinical platforms; 6) the slow, unmonitored translation of evidence into routine practice; 7) the risk of AI displacing rather than extending physician judgment; and 8) the failure of prediction-centric AI to generate actionable, agentic clinical guidance. For each dimension, we reframe the problem and propose a path forward, culminating in a vision of physician-centered AI solutions with agentic reasoning that extend, rather than replace, clinical judgment.

Keywords: Artificial Intelligence; Medical Imaging; Physician-In-The-Loop Artificial Intelligence; Agentic Artificial Intelligence; Foundation Models; Clinical Trust; Implementation Science; Data Sharing.

Copyright © 2026 Tehran University of Medical Sciences.
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.
DOI:

1. Introduction

Artificial Intelligence (AI) in medical imaging has undergone rapid expansion, with proposed solutions proliferating across nearly every domain of clinical practice. The field, however, is replete with exaggerated claims (whether of the present or of the future) [1]. Less than a decade ago, convolutional neural networks were confidently predicted to replace radiologists, a prediction that, in hindsight, failed due to misunderstanding what radiologists do and underestimating technical constraints; and today, foundation models are promoted as universal engines of medical reasoning [2]. Clinical reality, however, remains sobering: adoption is patchy, impact modest, and trust fragile [3]. The widening gap between technical breakthroughs and bedside benefit is not due to regulatory inertia or clinician caution alone; it reflects deeper misalignments between how AI is developed and how medicine is practiced. Medical imaging exposes these tensions most clearly. Here, clinicians do not reason from pixels alone. They integrate laboratory trends, comorbidities, therapies received, prior imaging, and evolving disease trajectories, particularly in complex or ambiguous cases. AI algorithms designed around images alone, however elegant, may fall short of clinical relevance in complex, context-dependent scenarios. Broader challenges persist in models that collapse patients into voxel grids, tools that fail to earn physician trust, foundation models that scale without context, and a persistent scarcity of curated multimodal datasets. This paper does not claim to introduce an entirely new problem; rather, it offers an integrated perspective across these interconnected dimensions, oriented toward embedding AI as an extension of the physician and grounded in the realities of clinical practice. We argue that these are not independent technical shortcomings but expressions of a single structural problem: medical imaging AI has been optimized for benchmark performance rather than clinical utility. Solving this requires not incremental algorithmic refinements, but fundamental reorientation, from AI as a product to AI as a physician-aligned capability embedded in the realities of clinical practice.

These tensions between technological assumptions and clinical reality occur across nearly every dimension of medical imaging AI [4]. They are not isolated problems but manifestations of a deeper structural mismatch in how AI is conceived, evaluated, and deployed. This broader trajectory is cap-turned in Figure 1, which maps 8 major themes in the field, spanning clinical context, trust, foundation models, data, deployment platforms, implementation science, and agentic AI, across three stages: early optimism, practical constraints, and clinically grounded reframing; illustrating the shift from model-centric innovation toward clinically integrated, context-aware, physician-aligned AI. A growing body of literature has examined aspects of this misalignment, including issues of workflow integration, model generalizability, and clinical trust. The present work seeks to contribute a synthesis across these interdependent aspects, spanning clinical trust, multimodal data integration, implementation science, and agentic AI, brought together within a single clinical alignment framework. We acknowledge that several themes discussed here, particularly data scarcity, harmonized sharing, and governance, are themselves technical and policy challenges; the intent is not to avoid those domains but to situate them within a broader clinical alignment framework.

1) Beyond Pixels. Much of imaging AI has inherited a paradigm from natural image analysis, reducing patients to image voxel arrays and diagnostic problems to classification tasks. Well-validated image-only AI systems, such as large vessel occlusion triage tools and breast cancer screening algorithms, already demonstrate meaningful clinical value. Radiologists often function effectively even with limited clinical history, and in many settings, this remains the norm. Nevertheless, for complex or ambiguous cases, the integration of laboratory values, comorbidities, prior therapies, and treatment histories has potential to substantially enhance diagnostic accuracy and clinical utility. Models designed with no ability to access or incorporate such context may be technically sophisticated yet constrained in their applicability to nuanced clinical scenarios. As an example, variations in patient hemoglobin levels during prostate cancer therapies are significant considerations for subsequent therapy plans [5] , and an AI-based Clinical Decision Support (CDS) tool that only relies on radiological images will be extremely limited in such contexts. As such, AI models that assess tumor response without access to concurrent

PROOF

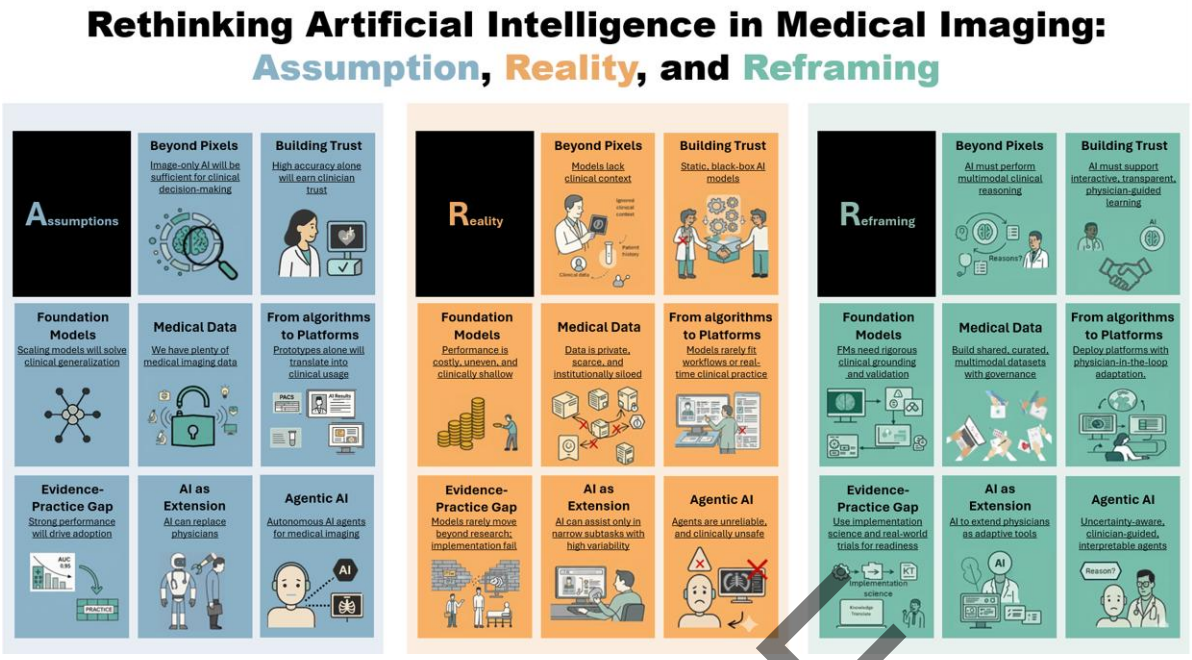


Figure 1. Overview of the evolving landscape of AI in medical imaging, spanning 8 interconnected dimensions: models that go beyond pixels, physician trust, foundation models, medical data, the path from algorithms to platforms, the evidence–practice gap, AI as an extension of the physician, and agentic AI. Each dimension is shown across three stages: Assumptions (what the field has assumed or believed), Reality (what is actually observed or what has failed), and Reframing (what is needed for meaningful clinical impact). Together, these trajectories illustrate the broader shift from model-centric innovation to clinically integrated, context-aware, physician-aligned AI

PROOF

laboratory trends or systemic therapy records may mischaracterize disease trajectory in ways that image data alone cannot resolve [6].

2) Building Trust. Trust cannot be legislated into existence; it emerges from a more intricate interplay of utility, reliability, transparency, alignment of values, and professional agency [7]. Current AI systems too often present themselves as static "black box" solutions with middling accuracy, leaving clinicians frustrated [8]. In the physician-in-the-loop AI model, the clinician acts as an active participant who labels data, corrects outputs, and steers model learning, with the primary goal of improving AI performance over time. In the AI-in-the-loop physician model, by contrast, AI serves as a continuous decision-support layer embedded in the clinician's workflow: surfacing relevant information, flagging uncertainty, and offering evidence-grounded recommendations, while the physician retains full authority over the final decision. The first model centers on improving the AI iteratively through human correction; the second centers on augmenting human judgment in real time without requiring the physician to act as a data labeler.

Importantly, neither model can function without addressing the realities of clinical data. In medical imaging, we rarely have the scale, heterogeneity, and longitudinal diversity needed to train models that can internalize the full breadth of clinical variability. By contrast, physicians accumulate thousands of hours of exposure to rare presentations, edge cases, artifacts, and multimodal clinical context, as forms of experiential knowledge that current models do not replicate and that AI systems must increasingly seek to incorporate through active and continual learning frameworks [9–12]. Such frameworks allow AI tools to evolve while avoiding catastrophic forgetting.

Trust also depends on how AI systems present themselves to individual clinicians. Clinicians differ in their tolerance for errors, decision boundaries, and workflow priorities [13, 14]. Rather than mirroring idiosyncratic interpretive tendencies, which would encode inter-reader variability as a feature rather than addressing it as a known diagnostic quality challenge, AI systems should aim to standardize evidence-based pathways while remaining adaptable to different clinical workflow contexts and communication preferences. Interfaces must empower clinicians through intuitive interactions, such as detection

modification and segmentation refinement. Only when AI systems behave as adaptable, transparent, and responsive partners, not distant authorities, can they earn a place as trusted colleagues in clinical practice.

Three concrete gaps limit how current physician-in-the-loop AI interacts with and incorporates physician input. First, repetitive annotation burden: physicians routinely re-draw contours or re-click boxes on every case, even every lesion, adding hands-on time to an already time-pressured read. Second, error repetition: a standard model corrects an error once and forgets it, repeating the same mistake case after case because it lacks any mechanism to carry correction patterns forward. Third, poor uncertainty communication: physicians are rarely told, case by case, how reliable the AI's output is, or when their scrutiny is most likely to add value. Agentic reasoning addresses all three: by carrying correction patterns forward, calibrating uncertainty case by case, and reducing annotation burden through learned physician-specific thresholds. That these gaps matter clinically is underscored by recent evidence: a 2024 meta-analysis of 106 studies found that human–AI pairs perform significantly worse, on average, than the better of the human or AI alone, with the sharpest losses in decision-making tasks, and a randomized trial showed no diagnostic improvement when physicians were supported by a query-only LLM even though the LLM alone outperformed them. Pairing can outperform either party when physicians have something to actively interrogate rather than a pre-computed output to accept or override out of habit, which demands appropriate reliance: the ability to discern when the AI is likely correct and to act on that judgment rather than defer uniformly.

3) Promise and Limits of Foundation Models. Foundation models have been celebrated for their ability to learn general-purpose representations from large datasets. In practice, however, gains in medical imaging have been modest, task-dependent and inconsistent across institutions [15, 16]. Their energy demands are substantial, raising questions about sustainability of scaling as a primary strategy [17,18]. The argument here is not against scale per se, but against scale as a substitute for clinical grounding: without corresponding improvements in data quality, physician-driven validation, and evaluation rigor, larger models are unlikely to close the gap between research performance and real-world impact. Notably, the field has increasingly embraced self-supervised and weakly supervised learning approaches that reduce dependence on curated labels; a shift that partially addresses the labeling bottleneck, though it does not eliminate the need for careful validation and clinical alignment.

In data-sparse domains like medical imaging (and specific modalities within it), genuine progress will depend on carefully curated multimodal datasets paired with rigorous evaluations of clinical benefit [19]. Curation, however, is not synonymous with sanitization: models trained only on clean, carefully selected data may fail to generalize the noise and heterogeneity of routine clinical practice. Curation efforts should therefore preserve representative diversity rather than eliminate it. The goal is not simply more data, but better-structured data aligned with the decisions clinicians face. Without these, foundation models may shine in theory yet deliver only modest returns where it matters most: the bedside.

4) A Central Bottleneck in Data. Medical imaging continues to face profound data scarcity, not only a lack of large, labeled cohorts but, more importantly, a lack of shareable datasets. A 2022 meta-research study found that fewer than 2% (4 of 218) of radiology AI articles published from 2017–2021 shared both code and complete experimental data, in stark contrast to the 40–50% code-sharing rates seen in general computer vision and NLP research [20]. Without broad collaboration, harmonized data sharing, and structured generation of multimodal imaging resources, AI advances in medicine will remain incremental.

A practical way to mitigate limited shareable cohorts is to expand data diversity via clinically grounded augmentation and constrained synthetic generation. Basic perturbations improve robustness but do not create the rare anatomies, edge-case disease patterns, and institution-specific acquisition signatures that drive real-world failure. Generative models can target these missing modes, but only if synthesis is conditioned on clinically meaningful factors and screened for anatomical plausibility and label consistency; otherwise, augmentation quietly injects hallucinated pathology and unrealistic features. A

PROOF

reliable strategy is controlled generation plus automated filtering and clinician spot-audits, prioritizing under-represented subgroups and hard cases. Another often-overlooked path is revisiting the ethics and governance of medical data sharing. It has been argued [21, 22] that patients do not "own" their medical data; instead, institutions should act as custodians responsible for stewardship and socially beneficial use. This perspective supports the creation of more permissive, standardized, and well-governed data-access frameworks, essential prerequisites for reproducible, multicenter AI research and for breaking the data bottleneck limiting clinical translation.

5) From Algorithms to Platforms. Algorithms, no matter how advanced, cannot transform practice into isolation. Translation requires platforms: integrated systems that plug into daily workflows, connect with hospital PACS, and evolve iteratively with clinician input. Such platforms must address the full integration challenge: bridging AI models with electronic health records, handling diverse imaging protocols across scanners and institutions, supporting human-in-the-loop feedback loops, and enabling deployment at scale without sacrificing safety or interpretability. Emerging open frameworks illustrate what clinically integrated platforms can look like in practice. MONAI Label, for instance, enables active learning workflows by connecting AI-assisted annotation tools directly with clinical annotation pipelines, reducing labeling burden while continuously refining model performance. Kaapana provides an open-source, container-based platform for deploying AI pipelines in federated, privacy-preserving clinical environments, enabling multi-institutional collaboration without centralizing sensitive data. These frameworks represent a meaningful step toward making AI deployment a practical, maintainable endeavor rather than a one-time research artifact. Realizing this philosophy at scale will require not only technical infrastructure but governance frameworks and reimbursement models that support sustainable clinical AI deployment as an ongoing institutional commitment.

6) Closing the Evidence–Practice Gap. Even when strong evidence exists for the value of a particular solution, implementation is not guaranteed. Historically, effective interventions have taken 15–17 years to reach routine practice [23, 24], and there is little reason to assume AI will be exempt from this inertia without deliberate strategies to accelerate translation. Implementation science, which is a systematic study of methods to promote the uptake of evidence-based interventions into routine practice, offers a structured path forward [25,26]. Its tools include identifying context-specific barriers to adoption, engaging stakeholders early, and designing interventions that fit existing workflows rather than demanding that workflows adapt to them.

For medical imaging AI specifically, this means deploying integrated knowledge translation (iKT) approaches: co-creating AI solutions with physicians from the initial definition of clinical problems through to the design and execution of evaluation studies. Clinicians who participate in defining the problem are far more likely to recognize the solution as meaningful, to flag implementation barriers early, and to advocate for adoption within their institutions. Without this co-design process, even technically validated models' risk being solutions to problems clinicians did not ask for, evaluated on metrics clinicians do not use, deployed in workflows clinicians were not consulted about.

The evidence–practice gap in medical imaging AI is also widened by a monitoring problem that implementation science has begun to address in other domains but that remains largely unsolved here: most health systems currently lack the infrastructure to detect when deployed AI tools begin to degrade, whether due to shifts in patient populations, changes in imaging protocols, or scanner upgrades. Closing this gap requires not only successful initial deployment but sustained institutional commitment to post-deployment surveillance, re-validation, and iterative improvement.

7) AI As Extension of the Physician. The goal is not to replace physicians but to extend their capabilities; e.g., augmenting what they can perceive, synthesize, and act upon, rather than substituting for the judgment that remains irreducibly theirs. For subtasks such as segmentation, dosimetry and lesion detection, physician-in-the-loop AI can improve reproducibility and efficiency. For clinical decision-making, physicians remain central. The true vision is AI-in-the-loop physicians: clinicians empowered by AI to make more informed, autonomous and effective decisions, a physician equipped with AI-enabled

PROOF

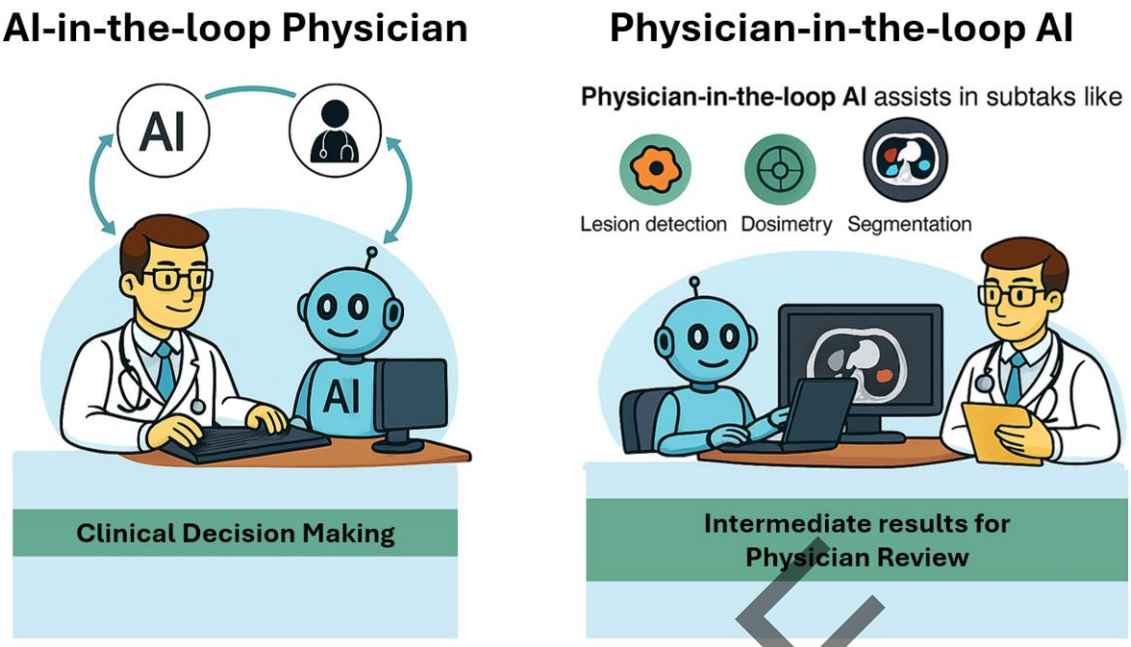


Figure 2. Empowering physicians to make more autonomous, informed and effective decisions; interlinking of physician-in-the-loop AI and AI-in-the-loop physician

armor rather than replaced by it (Figure 2). The distinction between these two modes, elaborated in the Building Trust section above, is operationally meaningful: the former improves the model iteratively via human correction; the latter improves the clinician's decision-making in real time via AI-generated insights, without requiring the physician to act as a data labeler.

Realizing this vision requires AI that adapts to how the individual physician works. Physician behavior is measurable, through cursor movements, dwell times, report edits, confidence ratings, and gaze patterns, and these signals can be used to build physician-specific models of reasoning. Preliminary work has demonstrated that an agent trained on a physician's history of accepted, revised, and overridden segmentation suggestions can anticipate the type and location of the next correction, significantly cutting interactions while improving accuracy (see last section). This personalization must be robust: when insufficient data exists for a given physician on a given task, the system should default to a shared non-personalized model, and the balance between the two should shift transparently as data accumulates. The goal is not a single fixed interaction style imposed on all clinicians, but an AI collaborator that becomes recognizably adapted to each physician, knowing when to propose, when to hold back, and how much explanation to provide.

8) Actionable and Agentic AI: Aligning Imaging Predictions with Clinical Decisions. This dimension concerns the failure of prediction-centric AI to generate actionable clinical guidance, and the agentic architectures needed to close this gap. A recurring limitation of medical imaging AI is not its ability to predict, but its failure to translate those predictions into concrete clinical action [27]. Clinicians do not primarily ask whether an outcome can be predicted; they inquire as to what should change in patient management, when a change should occur, and perhaps how this prediction happened (explainability). Risk estimates that are not clearly linked to a specific clinical choice, such as adjusting follow-up strategies, modifying treatment assessments, or altering the timing of intervention, remain abstract and difficult to use in practice.

Actionable AI must therefore be built around decisions rather than predictions. Models should be designed to answer clinically meaningful questions, aligned with relevant time horizons, acceptable levels of uncertainty, and the consequences of acting versus not acting [28]. This shift requires moving beyond generic performance metrics toward evaluation frameworks that reflect how model outputs influence different clinical strategies (task-based evaluation)

PROOF

[8,29,30]. Without this alignment, even well-validated models may produce results that are technically correct yet offer little practical value at the bedside. Embedding actionability from the outset requires early co-definition of use cases with clinicians, explicit consideration of decision trade-offs, and interfaces that allow clinicians to interpret, question, and adapt outputs within the context of individualized patient care.

2. Towards Physician-Centered AI Solutions with Agentic Reasoning

To bridge the gap between pixel-level accuracy and patient-level care, the field must explore agentic AI solutions. Traditional deep learning treats diagnosis as a static, one-shot classification task, brittle when faced with ambiguity and blind to clinical history. Agentic systems, by contrast, solve the missing context problem by actively orchestrating the diagnostic pipeline. Instead of passively accepting an input image, an agent can reason dynamically: querying the Electronic Health Record (EHR) for lab trends to resolve a radiologic uncertainty, requesting alternative views, or cross-referencing prior reports. Community efforts such as Med Agent Bench [31], illustrate a shift toward self-improving, human-in-the-loop workflows that go beyond passive prediction to enable context-aware planning, iterative reflection, and adaptive action.

We propose that agentic architectures in medical imaging should be conceived around three functional capacities that traditional models lack. First, context orchestration: the ability to autonomously retrieve and synthesize relevant clinical information from EHRs, prior imaging, laboratory systems, and genomic databases before generating a recommendation. Second, uncertainty-aware reasoning: the capacity to recognize the limits of its own competence, flag cases that fall outside its reliable operating range, and escalate appropriately to human review rather than producing confident but unreliable outputs. Third, action-oriented output: producing not just predictions or probabilities, but structured recommendations linked to specific decision points, such as "recommend biopsy given PSA trajectory and lesion characteristics" rather than "probability of malignancy: 74%". This shift moves AI from "System 1" pattern matching to "System 2" deliberative reasoning. Yet significant challenges remain. Current systems lack robust mechanisms for recognizing the boundaries of their competence, and autonomous actions introduce new accountability concerns that existing regulatory frameworks are not yet equipped to address. Regulatory sandboxes and staged deployment models, analogous to Phase II/III trial logic but adapted for adaptive AI systems, will be needed to generate the post-market evidence required for safe scaling. Agentic AI should be understood not as autonomous replacement of clinical reasoning, but as a structured capability augmentation: a system that meets the physician inside their workflow, brings the relevant information to bear, and flags what it does not know, leaving final judgment unambiguously with the clinician. This is the architecture of AI that medicine needs.

These capacities are not merely theoretical. Recent agentic segmentation systems illustrate how each can be realized concretely in an imaging pipeline, addressing 3 significant existing gaps, as also depicted in Figure 3:

(i) *Repetitive annotation burden:* A physician who must repeatedly re-annotate the same case until an output is acceptable faces a real time and effort burden; Clinical-SAM [32] instead trains a frozen segmentation agent to learn an individual physician's acceptance standard from nothing more than binary accept/reject feedback, closing this loop without additional annotation [33].

(ii) *Repeated AI mistakes:* Without carried-forward context, the same errors can recur across cases; LORE [34] maintains a memory bank of prior physician–image interactions so that a correction made once is applied automatically on subsequent cases, rather than discarded [35].

(iii) *Uncalibrated uncertainty:* Where conventional outputs give no indication of their own reliability, PILOT [36] reports calibrated confidence alongside each proposed output and suggests the appropriate next action, operationalizing uncertainty-aware reasoning directly at the point of decision.

PROOF

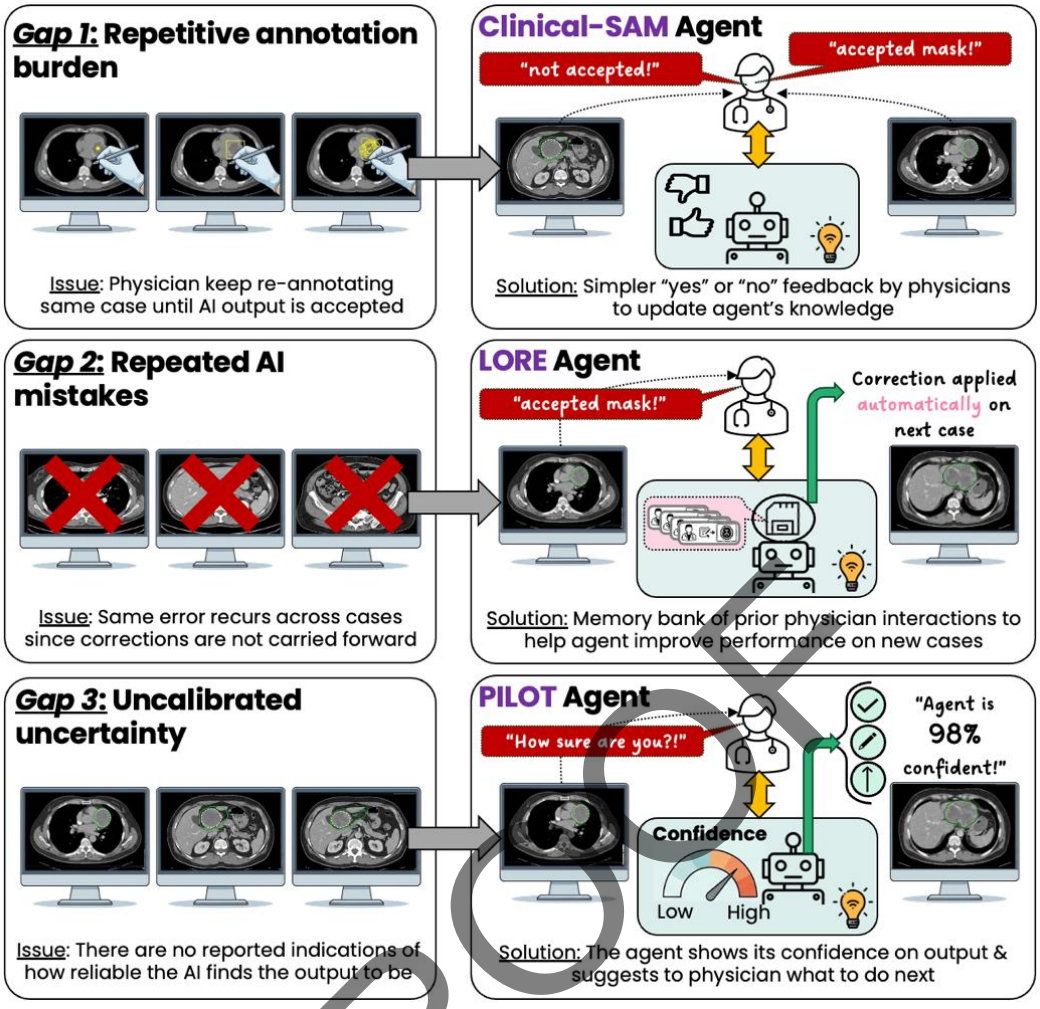


Figure 3. Three practical gaps limiting physician-facing agentic AI, and how minimal-feedback agentic systems close them. *(i) Repetitive annotation burden*: physicians re-annotate the same case until an AI output is accepted; Clinical-SAM instead learns a physician's individual acceptance standard from simple accept/reject feedback. *(ii) Repeated AI mistakes*: without memory, the same error recurs across cases; LORE maintains a memory bank of prior physician–image interactions so corrections carry forward automatically. *(iii) Uncalibrated uncertainty:* conventional outputs give no indication of their own reliability; PILOT reports calibrated confidence alongside each output and recommends the appropriate next step

PROOF

3. Conclusion

The eight reframing's advanced in this work converge on a single design imperative: medical imaging AI must be built for the physician, not merely handed to the physician. This means co-designing systems with clinicians from problem definition to deployment; measuring success by decision quality and patient outcomes, not benchmark accuracy; building platforms that integrate into workflows rather than disrupting them; and designing agentic systems that bring context, reasoning, and appropriate humility to every clinical encounter. None of this is technically impossible, frameworks like MONAI, Kaapana, and emerging agentic benchmarks provide concrete starting points. What is missing is not capability but orientation. The field has spent a decade asking, "can AI match physician performance on task X?" The more important question is "does AI make physicians better at caring for patients?" Answering this question requires sustained collaboration across machine learning, clinical medicine, implementation science, and health policy, and a shared commitment to measuring progress not by papers published but by patients helped.

Acknowledgments

We acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Horizons Grant DH-2025-00119, and the BC Cancer Foundation.

References

1- Amirhosein Toosi, Andrea G Bottino, Babak Saboury, Eliot Siegel, and Arman Rahmim, "A brief history of AI: how to prevent another winter (a critical review)." *PET clinics,* Vol. 16 (No. 4), pp. 449–69, (2021).

2- Tugba Akinci D'Antonoli, Christian Bluethgen, Renato Cuocolo, Michail E Klontzas, Andrea Ponsiglione, and Burak Kocak, "Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects." *Diagnostic and Interventional Radiology,* Vol. 32 (No. 3), p. 259, (2026).

3- Henry Litt, Paras Mehta, Aarush Sahni, and Ronac Mamtani, "FDA-authorized oncology artificial intelligence and machine learning devices and their clinical evidence: A cross-sectional analysis." *Journal of Cancer Policy,* p. 100743, (2026).

4- Xiaoxuan Liu, Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M. & Kern, C. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis." *The lancet digital health,* Vol. 1 (No. 6), pp. e271–e97, (2019).

5- Tomasz M Beer, Tangen, C.M., Bland, L.B., Hussain, M., Goldman, B.H., DeLoughery, T.G. & Crawford, E.D, "The prognostic value of hemoglobin change after initiating androgen-deprivation therapy for newly diagnosed metastatic prostate cancer: a multivariate analysis of Southwest Oncology Group Study 8894." *Cancer,* Vol. 107 (No. 3), pp. 489–96, (2006).

6- Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data." *Scientific Reports,* Vol. 12 (No. 1), p. 17981, (2022).

7- Navid Hasani, Morris, M.A., Rhamim, A., Summers, R.M., Jones, E., Siegel, E. & Saboury, B, "Trustworthy artificial intelligence in medical imaging." *PET clinics,* Vol. 17 (No. 1), p. 1, (2022).

8- Abhinav K Jha, Myers, K.J., Obuchowski, N.A., Liu, Z., Rahman, M.A., Saboury, B., Rahmim, A. & Siegel, B.A., "Objective task-based evaluation of artificial intelligence-based medical imaging methods: framework, strategies, and role of the physician." *PET clinics,* Vol. 16 (No. 4), pp. 493–511, (2021).

9- Matthias De Lange, Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. & Tuytelaars, T., "A continual learning survey: Defying forgetting in classification tasks." *IEEE transactions on pattern analysis and machine intelligence,* Vol. 44 (No. 7), pp. 3366–85, (2021).

10- Pengzhen Ren, Xiao, Y., Chang, X., Huang, P.-.Y., Li, Z., Gupta, B.B., Chen, X. & Wang, X., "A survey of deep active learning." *ACM computing surveys (CSUR),* Vol. 54 (No. 9), pp. 1–40, (2021).

11- Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi, "Gc 2: Generalizable continual classification of medical images." *IEEE Transactions on Medical Imaging,* Vol. 43 (No. 11), pp. 3767–79, (2024).

12- Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi, "Continual-Zoo: Leveraging zoo models for continual classification of medical images." in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2024): *IEEE*, pp. 4128–38.

13- Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam, "The false hope of current approaches to explainable artificial intelligence in health care." *The lancet digital health,* Vol. 3 (No. 11), pp. e745–e50, (2021).

14- Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer, "Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care." in *Proceedings of the 2023 CHI conference on human factors in computing systems*, (2023), pp. 1–18.

15- Shaoting Zhang and Dimitris Metaxas, "On the challenges and perspectives of foundation models for medical image analysis." *Medical image analysis,* Vol. 91p. 102996, (2024).

16- Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi, "BoosterNet: Improving domain generalization of deep neural nets using culpability-ranked features." in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022): *IEEE*, pp. 528–38.

17- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie, "Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data." *Nature Communications,* Vol. 16 (No. 1), p. 7866, (2025).

18- Suraj Pai, Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H. & Aerts, H.J, "Vision foundation models for computed tomography." *arXiv preprint arXiv:2501.09001,* (2025).

19- Siyi Du, Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi, "Avit: Adapting vision transformers for small skin lesion segmentation datasets." in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2023): *Springer*, pp. 25–36.

PROOF

20- Kesavan Venkatesh, Samantha M Santomartino, Jeremias Sulam, and Paul H Yi, "Code and data sharing practices in the radiology artificial intelligence literature: a meta-research study." *Radiology: Artificial Intelligence,* Vol. 4 (No. 5), p. e220081, (2022).

21- Barbara J Evans, "Much ado about data ownership." *Harv. JL & Tech.,* Vol. 25p. 69, (2011).

22- Jonathan Herington, McCradden, M.D., Creel, K., Boellaard, R., Jones, E.C., Jha, A.K., Rahmim, A., Scott, P.J., Sunderland, J.J. & Wahl, R.L., "Ethical considerations for artificial intelligence in medical imaging: data collection, development, and evaluation." *Journal of Nuclear Medicine,* Vol. 64 (No. 12), pp. 1848–54, (2023).

23- Zoë Slote Morris, Steven Wooding, and Jonathan Grant, "The answer is 17 years, what is the question: understanding time lags in translational research." *Journal of the royal society of medicine,* Vol. 104 (No. 12), pp. 510–20, (2011).

24- Shahnaz Khan, David Chambers, and Gila Neta, "Revisiting time to translation: implementation of evidence-based practices (EBPs) in cancer control." *Cancer Causes & Control,* Vol. 32 (No. 3), pp. 221–30, (2021).

25- Ahmad Fayaz-Bakhsh, Janice Tania, Syaheerah Lebai Lutfi, Abhinav K Jha, and Arman Rahmim, "What Is Implementation Science: And Why It Matters for Bridging the Artificial Intelligence Innovation-to-Application Gap in Medical Imaging." *PET clinics,* Vol. 21 (No. 1), pp. 1–16, (2026).

26- Julia Adler-Milstein, Aggarwal, N., Ahmed, M., Castner, J., Evans, B.J., Gonzalez, A.A., James, C.A., Lin, S., Mandl, K.D. & Matheny, M.E, "Meeting the moment: addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis." *NAM perspectives,* Vol. 2022p. 10.31478/202209c, (2022).

27- VI Madai and DC Higgins, "Artificial intelligence in healthcare: Lost in translation?. arXiv." *arXiv preprint arXiv:2107.13454,* Vol. 10(2021).

28- Marije Wijnberge., Geerts, B.F., Hol, L., Lemmers, N., Mulder, M.P., Berge, P., Schenk, J., Terwindt, L.E., Hollmann, M.W. & Vlaar, A.P. "Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial." *Jama,* Vol. 323 (No. 11), pp. 1052–60, (2020).

29- Andrew J Vickers and Elena B Elkin, "Decision curve analysis: a novel method for evaluating prediction models." *Medical Decision Making,* Vol. 26 (No. 6), pp. 565–74, (2006).

30- Baptiste Vasey, Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B. & Ibrahim, M., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI." *Nature medicine,* Vol. 28 (No. 5), pp. 924–33, (2022).

31- Yixing Jiang, Black, K.C., Geng, G., Park, D., Zou, J., Ng, A.Y. & Chen, J.H, "MedAgentBench: a virtual EHR environment to benchmark medical LLM agents." *Nejm Ai,* Vol. 2 (No. 9), p. AIdbp2500144, (2025).

32- Bayasi N, Fayyad J, Yousefirizi F, Rahmim A., "Clinical-SAM: Agentic Test-Time Adaptation of SAM via Physician-in-the-Loop for Clinical Tumor Segmentation." *IEEE Transactions on Medical Imaging*, In Review (2026).

33- Monica Luo, Yousefirizi, F., Rouzrokh, P., Jin, W., Alberts, I., Gowdy, C., Bouchareb, Y., Hamarneh, G., Klyuzhin, I. & Rahmim, A., "Physician-in-the-loop active learning in radiology artificial intelligence workflows: opportunities, challenges, and future directions." *American Journal of Roentgenology,* Vol. 225 (No. 5), p. e2533364, (2025).

34- Nourhan Bayasi, Fereshteh Yousefirizi, and Arman Rahmim, "Learning to Reason Over Physician Corrections: An Interactive Agentic Framework for 3D Tumor Segmentation." Accepted at MICCAI CLiMeM Workshop (2026).

35- Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone, "When combinations of humans and AI are useful: A systematic review and meta-analysis." *Nature Human Behaviour,* Vol. 8 (No. 12), pp. 2293–303, (2024).

36- Bayasi, N., F. Yousefirizi, A. Rahmim, "PILOT: An Uncertainty-based Agentic-AI Framework for Interactive 3D CT Tumor Segmentation." *SPIE Medical Imaging*, In Review (2027).

PROOF