



From signal to reasoning: Computational bottlenecks in the evolution of artificial intelligence in radiology

Şükrü Mehmet Ertürk

Sıtkı Safa Taflan

Istanbul University Faculty of Medicine, Department
of Radiology, Istanbul, Türkiye

ABSTRACT

Artificial intelligence (AI) in radiology is often described as a sequence of architectures. This conceptual narrative review instead organizes its evolution around two questions: Which computational constraint was relaxed, and where did the resulting capability enter the radiologic chain from signal formation to recommendation? We identify five analytical epochs: handcrafted computer-aided detection, deep and volumetric learning, a branching stage of global-context modeling and learned reconstruction, multimodal foundation models, and an emerging stage of inference-stage reasoning. Epochs I–IV are characterized primarily by changes in representation or image formation; Epoch V shifts the frontier toward adaptive case-specific inference, for which radiology-specific evidence remains largely preclinical. The framework also generates a clinical prediction: Verification should be matched to the level of participation and its characteristic failure mode. Existing studies provide partial empirical support, but later-stage verification requirements remain framework-derived proposals rather than established standards.

KEYWORDS

Artificial intelligence, radiology, deep learning, foundation models, clinical reasoning

1. Introduction

Radiology has repeatedly been a prominent clinical setting for artificial intelligence (AI) translation, from computer-aided detection (CAD) through deep learning, learned reconstruction, transformers, and multimodal foundation models. Reasoning-oriented systems now raise a more difficult question: Can AI participate not only in recognition but also in constructing, comparing, and verifying diagnostic hypotheses?

Previous histories have organized medical imaging AI around chronological milestones, technical families, and clinical applications.^{1,2} The present framework instead asks which constraint warrants separate historical status and what verification problem follows from the resulting chain position.

The central thesis is that radiologic AI advances when relaxing a computational constraint makes a new class of machine participation tractable. Here, “bottleneck” denotes that constraint; “analytical epoch” denotes a broad historical-computational period; “frontier” denotes the leading boundary of capabilities made tractable during that period; and “branch” denotes a distinct bottleneck-defined path within an epoch. Paradigm is descriptive only, not a formal taxonomic level. Bottleneck identity classifies; chain position predicts the dominant verification problem. Independent bottlenecks may be grouped as branches within one epoch only when neither is a necessary predecessor of the other and forcing a sequence adds no explanatory value; temporal overlap alone is insufficient. Research frontier status does not imply clinical maturity.

The framework has explicit limits. A bottleneck is weakened if its relaxation does not open a materially new capability, if that capability was already routine, or if gains arise mainly from

Received 08 September 2026; revision requested 16
September 2026; last revision requested 21 September
2026; accepted 22 September 2026.

Corresponding author: Şükrü Mehmet Ertürk

E-mail: smerturk@gmail.com



Epub: 30.09.2026

DOI: 10.4274/dir.2026.264457

scale, data, or engineering. Older architectures can persist within the framework when new strategies complement or hybridize with them.

Sutton’s “Bitter Lesson” and scaling-law work emphasize general methods, data, and computation; here, the unit of analysis is the constraint whose relaxation changes what a machine can contribute clinically.^{3,4}

The criterion also has retrospective discriminatory value. Classical handcrafted radiomics retains predefined quantitative descriptors and therefore remains within Epoch I rather than earning separate epoch status.⁵ The fuller classification is developed in Section 4.

Recent radiology literature has synthesized foundation models, agentic systems, and human–AI collaboration.^{6–8} Those reviews characterize technological families and implementation relationships; the present review asks why the frontier of machine participation moves and how that changes verification.

Rahmim al’s⁹ 2026 perspective uses bottleneck language for deployment and clinical-alignment constraints, including multimodal integration, data sharing, trust, deployability, and actionable guidance. Here, computational bottleneck instead denotes what representation or inference becomes technically tractable. Clinical alignment asks whether that capability becomes useful, trusted, deployable, and actionable. Either can remain binding after the other is relaxed; the accounts are complementary, not interchangeable.

Epochs I–IV are characterized primarily by changes in what is represented or how the image itself is formed: Engineered features give way to learned visual patterns, global context, learned reconstruction priors, and multimodal concepts. Epoch V is classified separately because the limiting problem shifts toward how available representations are interrogated during a case; Section 8 specifies the admission threshold.

Inference is not representation free: Case-specific hidden states, retrieved context, scratchpads, and tool outputs may still be generated. The distinction concerns which constraint makes a new class of participation tractable, not whether internal representations ever change (Figure 1).

2. Review approach

The literature search was last updated on September 4, 2026. Searches were iterative and question driven rather than systematic. PubMed, IEEE Xplore, arXiv, primary regulatory sources, and backward citation chains were used for landmark technical papers, representative radiologic translations, governance documents, and recent reasoning/ agentic work. Primary technical papers, original radiology studies, and official documents were prioritized; peer-reviewed sources were preferred, with preprints retained where mature evidence was limited. No formal screening counts, meta-analysis, or risk-of-bias assessment were undertaken. Epoch boundaries were adjudicated using the classification rule outlined above; overlap, coexistence, and hybridization were permitted.

3. Why radiology is an informative clinical laboratory

Radiology provided unusually favorable infrastructure for AI development. The transition from film to picture archiving and communication systems converted examinations into durable, transferable, machine-accessible objects, whereas the Digital Imaging and Communications in Medicine standard imposed common structures for images, metadata, acquisition parameters, and study organization.^{10,11} Each examination also routinely generated an expert report, creating large image–text corpora later used for weak supervision, alignment, report generation, and foundation-model training.¹² Lesions can be detected, segmented, measured, and followed, and diagnostic performance can be tested in reader studies. These properties make many radiologic tasks unusually operationalizable. In one systematic review of

950 Food and Drug Administration (FDA)-authorized AI and machine-learning devices through June 2024, 723 (76%) were radiology devices.¹³

Infrastructure, however, explains only why translation was comparatively easy. A deeper reason radiology is analytically informative is that its work can be decomposed into transformations from physical measurement to reconstructed image, finding, diagnostic concept, and recommendation. Major modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) also require computational reconstruction from acquired measurements, creating a pre-interpretive intervention point at which AI can modify the evidentiary object before diagnostic interpretation begins. The chain is an analytical scaffold rather than a literal one-way workflow: Clinical context, prior imaging, and provisional interpretation can feed backward across protocol selection, reconstruction, image search, and interpretation. AI systems have entered this scaffold at different levels, from marks on finished images through learned patterns, pre-finalization reconstruction, multimodal reporting, and reasoning or coordinated action.

This correspondence makes radiology more than an early adopter. From the framework we derive a prospective, normative prediction: The dominant verification problem should track chain position more closely than architecture. Systems that alter image formation should be judged primarily by evidence preservation and lesion fidelity; systems that produce semantic assertions by factuality, omission, and contextual consistency; and systems that recommend actions by calibration, provenance, and downstream clinical effects. These are proposed verification priorities, not established universal standards.

The literature provides partial convergent support. At the finding level, handcrafted CAD and deep-learning detection are architecturally distinct yet judged through reader performance and workflow effects.^{14–18} In the retrospective paired-reader PI-CAI study, an AI system detected clinically significant prostate cancer more accurately than 62 radiologists reading with PI-RADS 2.1, although non-inferiority to standard-of-care multidisciplinary routine practice was not confirmed.¹⁹ At image formation, adversarial CT reconstruction and diffusion MRI converge on fidelity, lesion preservation, and perturbation stability.^{20–22} At the semantic level, generalist multimodal and radiology-specific

Main points
• Radiologic artificial intelligence (AI) is better understood by the computational bottleneck each generation relaxed than by a simple chronology of model names. • Radiology is especially informative because clinical work can be analyzed from signal to image, finding, diagnostic concept, and recommendation; different AI systems enter this chain at different levels and times. • Epoch V changes the inference regime more than image representation, and current radiology-specific evidence for this transition remains early and predominantly preclinical. • As machine participation becomes less directly inspectable, oversight should shift from a final-glance model toward verification matched to chain position and failure mode.

Computational Evolution Across the Radiologic Representational Chain

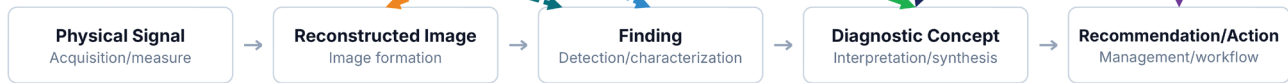
Bottleneck identity classifies the historical frontier; chain position predicts the verification problem.

A. Historical Classification — Bottleneck-Defined Frontiers

Rule: Bottleneck identity defines epoch/branch; chain position does not.



B. Clinical Position — An Analytical Scaffold, Not a One-Way Workflow



Feedback Loop: Clinical context, prior imaging, and provisional interpretation can feed backward across the scaffold

C. Predicted Dominant Verification Demand



Note: Horizontal position indicates typical level of entry. Architectures overlap, persist, and hybridize; Epoch V is distinguished by case-specific inference control.

Figure 1. Computational evolution across the radiologic representational chain. The horizontal backbone is an analytical scaffold from physical signal to reconstructed image, finding, diagnostic concept, and recommendation/action; it is not a literal unidirectional workflow. Bottleneck identity determines epoch or branch membership; chain position predicts the characteristic verification problem. Arrows indicate characteristic entry regions rather than exclusive scope: III-A can operate at image/finding as well as concept levels, and multimodal or reasoning systems may span multiple later levels. Epoch III contains two independent bottleneck-defined branches; Epoch V is an emerging inference regime on the research frontier. Positions are schematic and non-synchronous; older architectures persist and hybridize. CAD, computer-aided detection.

report generators converge on factuality and omission errors.²³⁻²⁶ These studies compare different architectures at one chain level. The reciprocal prediction—that one architectural lineage should require different verification across levels—remains prospective because no common design has yet tested it across the chain.

The fit is not deterministic: Reports are imperfect labels, archives reflect who was imaged, and measurable tasks can dominate research without being the most clinically important. The chain should therefore be treated as an explanatory scaffold whose predictions require empirical testing, not as a universal workflow law (Table 1).

4. Epoch I: handcrafted features and classical computer-aided detection

The first bottleneck was representational rigidity. Early systems classified only what engineers had formalized in advance: edges, textures, shapes, intensity distributions, or geometric descriptors passed to shallow classifiers. In radiology, this produced CAD for mammography and pulmonary nodules.²⁷ The representational unit was the predefined feature, and the role was narrow:

Mark suspicious locations for radiologist review.

CAD nevertheless established an enduring human-machine arrangement: The algorithm functioned as a second set of eyes rather than an interpreter of the examination as a whole. Its limitations exposed the cost of explicit feature engineering. In screening mammography, large real-world studies showed inconsistent or absent improvements and the burden of false-positive marks.¹⁴⁻¹⁶ Radiomics illustrates the same boundary: Classical pipelines expanded the inventory of predefined descriptors without changing the representational class or machine role. Medical images contained more variation than engineers could anticipate, making learned representation, rather than more elaborate feature engineering, the next necessary step.

5. Epoch II: deep learning and volumetric representation

Deep convolutional neural networks (CNNs) addressed that limitation through representation learning. AlexNet showed that layered networks could learn increasingly abstract visual features from raw pixels;²⁸ residual connections enabled deeper net-

works;²⁹ and U-Net and its three-dimensional (3D) variants provided encoder-decoder designs for biomedical segmentation.^{30,31} Representation expanded from engineered features to learned patterns and anatomically meaningful objects.

Radiologic translation was rapid; CNNs classified examinations, detected urgent findings, segmented organs and tumors, quantified disease burden, and prioritized worklists. U-Net variants made voxel-level delineation central to imaging AI, whereas nnU-Net demonstrated how a common framework could self-configure across heterogeneous segmentation tasks.³² Systems for intracranial hemorrhage and other critical findings showed how image analysis could alter workflow before the examination was opened.^{33,34}

Generalization remained unresolved. Supervised models were label dependent and could degrade across scanners, protocols, institutions, prevalence conditions, and populations.³⁵ Self-supervised and label-efficient pretraining partly relaxed annotation dependence by learning reusable representations from unlabeled or weakly labeled data.³⁶ Self-supervised learning nevertheless fails

Analytical epoch/branch	Dominant bottleneck	Computational response	Newly tractable capability	Radiologic translation/typical chain level	Characteristic risk/evidentiary demand
Epoch I: Handcrafted CAD	Relevant features must be specified in advance	Engineered features + shallow classifiers	Automated marking of predefined patterns	Mammographic and pulmonary nodule CAD/finding	Brittleness and false-positive marks; reader-performance testing
Epoch II: Deep/volumetric learning	Hand-engineered representations do not scale to image variability	CNNs, residual networks, U-Net/3D U-Net; later self-supervised and label-efficient pretraining	Learned patterns, objects, voxel-level segmentation, triage	Segmentation, quantification, urgent-finding prioritization/image → finding	Label dependence and domain shift; external and workflow testing
Epoch III-A: Global context (parallel bottleneck-defined branch)	Local operations make distributed relationships cumbersome	Self-attention, vision transformers, hybrid CNN-transformer models	Direct content-dependent long-range contextual modeling	3D/global-context segmentation and context-aware interpretation/image-finding → concept	Computational cost and task-dependent benefit; comparative testing rather than assumed replacement
Epoch III-B: Learned reconstruction (parallel bottleneck-defined branch)	Incomplete/noisy measurements limit acquisition	Variational, adversarial, and diffusion reconstruction	Learned inverse imaging	Accelerated MRI, low-dose CT/signal → image	Error embedded in image evidence; signal-fidelity and lesion-preservation testing
Epoch IV: Multimodal foundation models	Tasks and modalities remain semantically isolated	Large-scale multimodal pretraining and alignment	Image-text concepts, retrieval, report generation	Collaborative report assistance, prior comparison, multimodal interpretation/finding → concept → recommendation	Fluent omission or unsupported assertion; factuality and omission analysis; prospective role-specific testing†
Epoch V: Inference-stage reasoning and test-time scaling	Rich learned representations are limited by shallow or poorly controlled case-level inference	Reasoning-oriented post-training; iterative hypothesis generation, retrieval, critique, verification, and tool-mediated action; additional test-time compute as an enabling lever	Hypothesis comparison, evidence reconciliation, adaptive verification, and controlled coordination	Preclinical/emerging differential diagnosis, multimodal evidence reconciliation, and recommendation verification/concept → recommendation/action	Overthinking, irrelevant retrieval, tool error, unfaithful reasoning; calibration, provenance, auditability, and prospective human-in-the-loop/sociotechnical testing†

Bottleneck identity is the primary classificatory principle; chain position is a separate predictive dimension for verification. Epoch III is an umbrella analytical period containing two independent bottleneck-defined branches; the grouping does not imply shared mechanism, causal sequence, or synchronous invention. Chain-level entries indicate characteristic, not exclusive, participation. Epoch V is an inference-regime transition on the research frontier. †Epoch IV–V evidentiary demands are proposed verification priorities derived from failure modes and clinical roles, not estimates of current testing prevalence. Post-hoc explainability methods are cross-cutting rather than a distinct epoch. 3D, three-dimensional; CAD, computer-aided detection; CNN, convolutional neural network; CT, computed tomography; MRI, magnetic resonance imaging.

the primary epoch criterion: Annotation dependence constrains the efficiency and data requirements for acquiring an existing visual representational class; it does not make a new representational relation or inferential organization tractable. Unchanged chain position corroborates rather than determines the classification. Self-supervised learning therefore remains a bridge within Epoch II and a scaling condition for later foundation models. Convolutional architectures can still make long-range relationships cumbersome to model; broader context and image formation became the next distinct constraints.

6. Epoch III: Global context and learned reconstruction—parallel branches

The third period is best understood as a branching transition. One branch addressed contextual integration. Transformers used

self-attention to model relationships dynamically across a sequence,³⁷ and vision transformers allowed image-region tokens to interact beyond fixed local receptive fields.³⁸ Unlike self-supervised learning, attention changes the representational relation itself: Distant elements can interact directly rather than only through repeated local operations. The UNETR model provides a radiology-specific instantiation: It reformulated 3D medical-image segmentation as sequence-to-sequence prediction because localized convolutional receptive fields limited long-range spatial dependency modeling, using a transformer encoder to capture global multiscale information.³⁹ Medical imaging remains heavily hybridized with convolution because global modeling and fine localization impose different demands; III-A is defined by tractable global context, not CNN displacement.

A parallel branch addressed inverse imaging. Variational networks, adversarial methods, and later score-based diffusion models learned priors for producing clinically usable images from incomplete or degraded measurements.^{20,21,40,41} In accelerated MRI and low-dose CT, AI entered the chain before interpretation by participating in image formation. Failure could therefore become embedded in the evidence rather than remain a visible mark a reader could dismiss. Learned priors can suppress noise but may remove subtle pathology or introduce unsupported plausible structure, and perturbation instability remains an important concern.²²

Epoch III is an umbrella analytical period, not a claim of shared mechanism or simultaneous invention. Both branches occupy a post-Epoch-II research frontier in which learned models extend beyond discrimina-

tive analysis of finalized images; III-A makes direct non-local context tractable within interpretation, whereas III-B moves learned modeling upstream into the inverse problem that constructs the evidentiary image, replacing or augmenting fixed reconstruction priors with learned priors. Therefore, III-B is not another Epoch II application merely because it uses deep learning: Its bottleneck is image formation, not representation of an already reconstructed image. The branches remain distinct; their common epoch label is an analytical grouping, not a chronology or causal sequence.

7. Epoch IV: Multimodal foundation models

Foundation models addressed task and modality isolation through large-scale pre-training and shared representation spaces. Visual encoders and language models could align images, text, and other clinical information within a common semantic framework.²³ Recent radiology reviews describe the resulting adaptability, multimodal integration, opportunities, and risks in detail.⁶ Within the present framework, their historical significance is narrower and more specific: They relax semantic and task isolation sufficiently for a model to operate across multiple levels of the radiologic chain.

Multimodal systems now support image interpretation, retrieval, visual question answering, and report generation. Generalist biomedical models demonstrated cross-modality transfer,²³ while radiology-specific work advanced clinician–AI report composition, large-scale two-dimensional/3D representation learning, and 3D CT report generation.^{24–26} Their most plausible current role remains collaborative report and context assistance rather than autonomous interpretation.

This transition moves AI into radiology's principal professional assertion. A CAD mark is an intermediate prompt; a report can redirect care. Standard next-token training is not directly optimized to minimize clinically consequential omissions, unsupported findings, incorrect negation, or misstated interval change. A report can therefore be fluent while clinically wrong, and multiple phrasings may be equivalent, but one omitted finding can matter greatly. Foundation models qualify as a distinct epoch not because they are larger, but because relaxing semantic and task isolation makes cross-modal interpretation and report-level participation

tractable. Their success also exposes the next bottleneck: Broad representation does not guarantee controlled case-specific inference.

8. Epoch V: Inference-stage reasoning and test-time scaling

Epoch V contains mechanisms that operate at different stages. Reasoning-oriented post-training can shape a policy for generating, selecting, or revising responses; reinforcement learning is one route to encouraging reflection and verification behavior.⁴² Radiology-specific work now provides a closer analog. Fan et al. trained ChestX-Reasoner with supervised fine-tuning and reinforcement learning using process supervision and introduced RadRBench-CXR, comprising 59,000 visual-question-answering samples and 301,000 clinically validated reasoning steps, together with RadRScore, which evaluates reasoning factuality, completeness, and effectiveness. The model improved benchmark reasoning and diagnostic performance, but the evidence remains benchmark-based rather than prospective clinical reasoning.⁴³

Test-time scaling adds per-case computation through longer budgets, multiple candidates, self-consistency, revision, retrieval, or tools. In medical question answering, m1, a text-only test-time scaling framework for large language models (LLMs), found gains only up to an optimal budget; beyond it, overthinking could degrade performance, while insufficient knowledge remained a separate bottleneck.⁴⁴ Byun et al.⁴⁵ found gains using multiple candidates and self-consistency across text and medical imaging tasks. Oh et al. found model- and task-dependent benefits, limited gains from simple token expansion in current vision-language models, and vulnerability to misleading expert-physician cues.⁴⁶ These studies establish case-specific computation as a conditional lever but do not all clear our threshold: m1, Byun,⁴⁵ and much of Oh mainly test budget allocation, sampling, or revision and are enabling evidence rather than standalone demonstrations.

Epoch V is therefore different from its predecessors in what primarily limits progress. The stronger claim rests on reorganized case-level computation—iterative hypothesis generation, evidence acquisition, verification, or tool-mediated action—not simply more tokens, samples, thresholds, or prompts. ChestX-Reasoner moves closer to this threshold because stepwise verification is part of the learned reasoning policy. Rou-

tine test-time augmentation, ensembling, threshold adjustment, or prompt variation may improve outputs but do not qualify on their own. Reasoning-oriented post-training lets policies exploit this organization. This is a research frontier claim, not a claim of equivalent clinical maturity.

Radiology-specific studies occupy different positions relative to this threshold. ChestX-Reasoner is principally training-side evidence: Supervised fine-tuning and reinforcement learning with process supervision train a policy to produce explicit intermediate reasoning, whereas RadRBench-CXR and RadRScore provide reasoning-focused evaluation.⁴³ Likewise, CMR-R is training-side. A multistage framework used quantitative cardiac MR parameters and semantic report descriptions from 16,104 multicenter cases to generate diagnostic chains across eight cardiac disease categories; because its inputs are structured and semantic rather than raw images, it tests semantic diagnostic reasoning rather than end-to-end image reasoning.⁴⁷ By contrast, MedRAX is a cleaner inference-side example: Without additional training, it dynamically selects specialized chest X-ray tools and a multimodal language model to answer case-specific queries.⁴⁸ Thus, training-side studies show that structured reasoning policies can be learned, whereas MedRAX illustrates case-specific orchestration. All remain benchmark based or preclinical; none establishes prospective autonomous radiologic reasoning in clinical workflow.

This remains an emerging research direction rather than established autonomous clinical practice. General LLM evidence shows that chain-of-thought explanations can be plausible yet unfaithful, irrelevant retrieved passages can distract generation, and intrinsic self-correction can fail or even degrade reasoning.^{49–51} These are not radiology-specific validation studies; reasoning over an incorrect perceptual premise is therefore best treated as an anticipated compound failure mode. Latency, cost, provenance, calibration, and auditability remain part of the clinical problem.

9. Cross-epoch synthesis: From representation to clinical participation

Three linked but non-identical trajectories emerge. Representation changes from predefined features to learned patterns and objects, broader context, reconstruction priors, and multimodal concepts. Inference

becomes more flexible through hypothesis comparison, retrieval, critique, and verification. The scope of clinical participation extends from detection and measurement through reconstruction and reporting toward reasoning and controlled coordination. These trajectories are related but not synchronous: Representational capability can advance without equivalent inferential reliability or clinical readiness.

The important change is not simply increasing accuracy but changing where machine contribution sits in the radiologic representational chain. Characteristic errors therefore change in both consequence and inspectability. A false-positive CAD mark is usually visible and can often be readily dismissed by the reader. A segmentation error may distort a measurement but remain geometrically reviewable. A reconstruction error can be embedded in the evidence itself. A generated report may introduce semantic errors that require comparison with the images and record. A reasoning system may organize several correct observations into a persuasive but incorrect conclusion. Increasing abstraction does not make every error less visible, but it weakens the assumption that a final human glance will reliably catch failure.

Post-hoc attribution is one cross-cutting response to the inspectability problem, not a separate analytical epoch. Saliency maps and gradient-weighted class activation mapping can expose image regions associated with an output, but they do not independently establish that the attribution faithfully tracks the computation or that the conclusion is correct.⁵² Radiology studies have shown limited sensitivity and robustness of popular saliency methods, whereas experimental mammography work found that explanation inputs did not reliably overcome radiologists' tendency to follow incorrect AI suggestions.^{53,54} These post-hoc attribution tools can support interrogation, localization, and error discovery, but they cannot substitute for outcome-appropriate testing or independent verification.

The verification requirements that follow are framework-derived proposals, not established universal standards. When direct inspection becomes less reliable, systematic verification should increasingly replace reliance on superficial oversight. Detection tools warrant reader and workflow studies; reconstruction systems warrant signal-fidelity and lesion-preservation testing; report generators warrant factuality and omission analysis; and reasoning or agentic systems

warrant calibration, robustness, provenance, tool-use auditing, and assessment of effects on human judgment. Of 723 FDA-authorized radiology devices in the cited review, submission documentation was available for 717; 208 (29%) incorporated clinical testing, 33 (5%) reported prospective testing, and 56 (8%) included human-in-the-loop testing.¹³ Given the authorization era and task mix, these figures mainly characterize earlier radiology-AI functions and should not be read as prevalence estimates for foundation, reasoning, or agentic systems.

Clinical and implementation evidence illustrates why the tested endpoint matters. ScreenTrustCAD showed non-inferiority of one radiologist plus AI compared with standard double reading for screening mammography.¹⁷ The randomized MASAI trial subsequently found a non-inferior interval-cancer rate, higher sensitivity, and unchanged specificity with AI-supported screening, and PRAIM provided large-scale real-world implementation evidence across 12 German sites; because radiologists voluntarily chose AI use in PRAIM, residual selection effects remain possible.^{55,56} LungIMPACT provides the complementary negative result: Across 93,326 chest radiographs, AI prioritization accelerated reporting but did not significantly shorten time to CT, diagnosis, referral, treatment, or stage.¹⁸ Together, these studies show that evidentiary maturity depends on testing the workflow or patient-level endpoint actually claimed, not on whether the result is positive.

Architecture determines computational possibility, not clinical value. A technical advance may relax a computational bottleneck yet fail to solve the clinically binding one if it cannot survive external testing, workflow friction, procurement, reimbursement, regulation, liability, or professional trust. Research prominence is not evidence of clinical progress. Predetermined change control plans, the European Union AI Act, general health-AI guidance, and World Health Organization guidance specific to large multimodal models increasingly address lifecycle governance.⁵⁷⁻⁶⁰

Human-AI collaboration work emphasizes complementarity, automation bias, oversight, and workflow effects, supporting the view that AI should currently function primarily as a clinical teammate rather than an autonomous replacement for physicians.^{8,61} Our framework adds a location-specific question: At which chain level does the machine contribute, and what evidence verifies it? Earlier systems mainly marked, measured,

or prioritized; foundation and reasoning systems can help formulate meaning. Collaboration therefore requires provenance, recognition of mismatch between formal patterns and patient context, communication of uncertainty, and explicit responsibility for consequential judgment. The issue is not benchmark performance alone, but whether the sociotechnical system performs reliably across institutions while preserving evidence and enabling meaningful verification (Figure 2, a conceptual and qualitative map).

10. Beyond reasoning: The emerging frontier

The framework is also useful prospectively: What unresolved constraint does an emerging system actually relax? Agentic radiology already combines language models with tools, databases, planning, memory, and multistep orchestration.^{7,62-64} Tool invocation alone, however, does not earn a new epoch. Agentic radiology would warrant promotion only if controlled delegation makes a genuinely new class of consequential coordination tractable and exposes a distinct failure regime involving permissions, provenance, recovery, tool boundaries, and authorization. Otherwise, it remains an extension of Epoch V plus workflow engineering.

Other emerging directions are useful mainly as stress tests of the criterion. Predictive world models target forward prediction and planning; radiologic forecasting of tumor evolution or treatment response remains extrapolative.^{65,66} Neuro-symbolic systems target verification through rules or knowledge structures;⁶⁷ continual learning targets model stasis but introduces drift, forgetting, versioning, and governance problems;⁶⁸ and interventional AI pushes participation from interpretation toward action, where real-time imaging, planning, robotics, and procedural safety become inseparable.⁶⁹ None should be granted epoch status by name alone.

These trajectories share one test: Which bottleneck do they remove? A larger model is not a revolution if it reproduces the same capability at greater cost. A meaningful advance should expand representational access, change the organization or controllability of inference, remove a genuine workflow constraint, or make a previously inaccessible task safe and tractable. A computational constraint may be relaxed while data access, trust, interoperability, regulation, or workflow integration remains clinically binding.⁹ The framework would be weakened if future

Error inspectability, clinical consequence, and evidentiary burden

QUALITATIVE CONCEPTUAL MAP — neither axis is calibrated; ranges overlap and do not constitute a measured rank order.

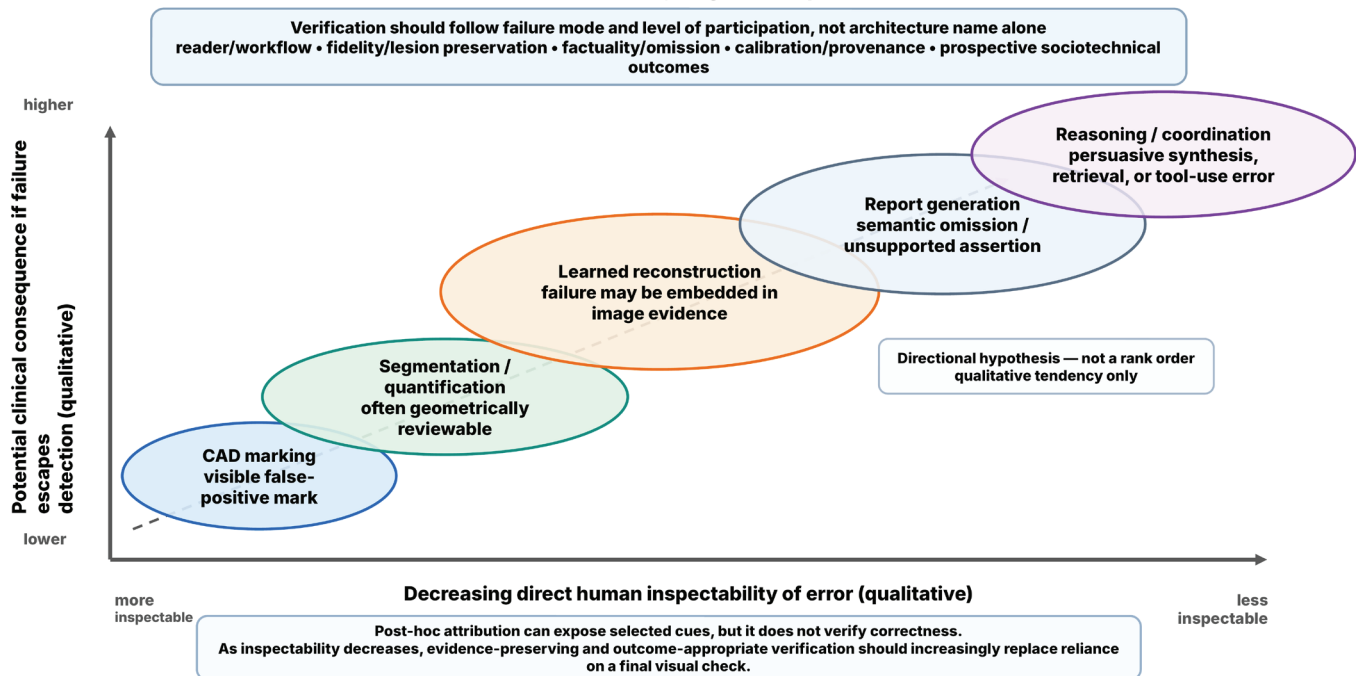


Figure 2. Conceptual, qualitative map of error inspectability, clinical consequence, and evidentiary burden. Representative contributions are shown as overlapping qualitative ranges against decreasing direct human inspectability of error and potential clinical consequence if failure escapes detection. Neither axis is calibrated or strictly monotonic, and clinical consequence remains task- and context dependent. The directional cue is a hypothesis, not a rank order. Reduced reliance on direct inspection should be accompanied by failure-mode-matched verification, progressing from reader/workflow and signal fidelity testing toward factuality/omission analysis, calibration, provenance/tool-use auditing, and prospective sociotechnical evaluation. Later-stage evidentiary escalation is partly normative rather than derived from device-landscape prevalence data.

systems repeatedly acquired genuinely new clinical roles without relaxing an identifiable representational, inferential, or coordination constraint. Promotion therefore depends on demonstrated constraint relaxation, not marketing vocabulary.

11. Limitations of the framework

This framework is explanatory rather than exhaustive. The review is selective and non-systematic, and its epochs are retrospective constructs rather than discrete events. Bottleneck identity is the primary taxonomic rule; chain position is a separate predictive dimension. Epoch III deliberately groups two distinct branches within one analytical period; treating them as occupants of a shared post-Epoch-II research frontier is an explanatory choice, not a claim of shared mechanism, synchronous invention, or causal sequence. The exact number of analytical epochs is therefore not a strong theoretical claim. Data, hardware, workflow, regulation, reimbursement, and professional practice may remain binding. Epoch V is especially provisional because its evidence mainly

concerns the research frontier. Later-stage evidentiary demands are normative inferences from failure modes and clinical roles, and Figure 2 presents a conceptual qualitative hypothesis rather than a quantitative risk model.

12. Conclusion

The history of AI in radiology is therefore not merely a sequence of architectures. Bottleneck identity explains why certain technical changes deserve separate historical status; chain position generates a distinct clinical prediction about how to verify those changes.

Radiology makes this relationship unusually visible because its practice can be analyzed from signal and image formation through finding, concept, and recommendation while retaining recursive clinical feedback. Successive AI paradigms have entered that scaffold at different levels and times, and older architectures continue to coexist. Epoch V remains the least mature empirically: Radiology-specific reasoning evidence is still predominantly preclinical. The broader

implication is methodological. As machine participation becomes more embedded, semantically richer, and less directly inspectable, safety cannot rest on the assumption that a supervising radiologist will simply see the error. The governing question should be whether evidence, provenance, and verification are appropriate to the level at which the system participates in care.

Acknowledgments

Use of AI-assisted technologies: ChatGPT (OpenAI; GPT-5.6 Sol; accessed through September 4, 2026) and Claude Opus 5 (Anthropic; accessed August 2026) were used for conceptual organization, drafting and editing assistance, figure development, and reference-checking support. The authors reviewed all AI-assisted content and independently verified the manuscript's claims and citations.

Footnotes

Conflict of interest disclosure

Şükrü Mehmet Ertürk, MD, serves as Editor-in-Chief of Diagnostic and Intervention-

al Radiology. He had no involvement in the peer review of this article and had no access to information regarding its peer review. The author declared no conflicts of interest.

References

1. Avanzo M, Stancanello J, Pirrone G, Drigo A, Retico A. The evolution of artificial intelligence in medical imaging: from computer science to machine and deep learning. *Cancers (Basel)*. 2024;16(21):3702. [\[Crossref\]](#)
2. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics (Basel)*. 2023;13(17):2760. [\[Crossref\]](#)
3. Sutton RS. The bitter lesson. Incomplete Ideas. 2019 Mar 13 [Internet]. Accessed 2026 Sep 4. [\[Crossref\]](#)
4. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. *arXiv [Preprint]*. 2020. [\[Crossref\]](#)
5. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563-577. [\[Crossref\]](#)
6. Akinci D'Antonoli T, Bluethgen C, Cuocolo R, Klontzas ME, Ponsiglione A, Kocak B. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. *Diagn Interv Radiol*. 2026;32(3):259-272. [\[Crossref\]](#)
7. Koçak B, Meşe İ. AI agents in radiology: toward autonomous and adaptive intelligence. *Diagn Interv Radiol*. 2025. [\[Crossref\]](#)
8. Kocak B, Cuocolo R. Human-AI interaction and collaboration in radiology: from conceptual frameworks to responsible implementation. *Diagn Interv Radiol*. 2026;32(5):535-547. [\[Crossref\]](#)
9. Rahmim A, Bayasi N, Li X, Saboury B, Yousefirizi F. Rethinking artificial intelligence in medical imaging: assumptions, reality, and reframing [Preprint]. *arXiv*. 2026. [\[Crossref\]](#)
10. Bidgood WD Jr, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. *J Am Med Inform Assoc*. 1997;4(3):199-212. [\[Crossref\]](#)
11. DICOM Standards Committee. Digital Imaging and Communications in Medicine (DICOM) Standard, PS3.1 2026c: Introduction and Overview [Internet]. Rosslyn (VA): National Electrical Manufacturers Association; 2026 [cited 2026 Sep 4]. [\[Crossref\]](#)
12. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. 2019;6(1):317. [\[Crossref\]](#)
13. Sivakumar R, Lue B, Kundu S. FDA Approval of artificial intelligence and machine learning devices in radiology: a systematic review. *JAMA Netw Open*. 2025;8(11):e2542338. [\[Crossref\]](#)
14. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*. 2007;356(14):1399-1409. [\[Crossref\]](#)
15. Gilbert FJ, Astley SM, Gillan MG, et al. Single reading with computer-aided detection for screening mammography. *N Engl J Med*. 2008;359(16):1675-1684. [\[Crossref\]](#)
16. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175(11):1828-1837. [\[Crossref\]](#)
17. Dembrower K, Crippa A, Colón E, Eklund M, Strand F; ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *Lancet Digit Health*. 2023;5(10):e703-e711. [\[Crossref\]](#) Correction in: *Lancet Digit Health*. 2023;5(10):e646. [\[Crossref\]](#)
18. Woznitza N, Smith L, Rawlinson J, et al. AI-based chest X-ray prioritization in the lung cancer diagnostic pathway: the LungIMPACT randomized controlled trial. *Nat Med*. 2026;32(5):1737-1744. [\[Crossref\]](#)
19. Saha A, Bosma JS, Twilt JJ, van Ginneken B, et al. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. *Lancet Oncol*. 2024;25(7):879-887. [\[Crossref\]](#)
20. Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative adversarial networks for noise reduction in low-dose CT. *IEEE Trans Med Imaging*. 2017;36(12):2536-2545. [\[Crossref\]](#)
21. Chung H, Ye JC. Score-based diffusion models for accelerated MRI. *Med Image Anal*. 2022;80:102479. [\[Crossref\]](#)
22. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc Natl Acad Sci U S A*. 2020;117(48):30088-30095. [\[Crossref\]](#)
23. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. *NEJM AI*. 2024;1(3):Aloa2300138. [\[Crossref\]](#)
24. Tanno R, Barrett DGT, Sellergren A, et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nat Med*. 2025;31(2):599-608. [\[Crossref\]](#)
25. Wu C, Zhang X, Zhang Y, Hui H, Wang Y, Xie W. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nat Commun*. 2025;16(1):7866. [\[Crossref\]](#)
26. Li CY, Chang KJ, Yang CF, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nat Commun*. 2025;16:2258. [\[Crossref\]](#)
27. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph*. 2007;31(4-5):198-211. [\[Crossref\]](#)
28. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012;25:1097-1105. [\[Crossref\]](#)
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis Pattern Recognit*. 2016:770-778. [\[Crossref\]](#)
30. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. 2015. p. 234-241. [\[Crossref\]](#)
31. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. 2016. p. 424-432. [\[Crossref\]](#)
32. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. [\[Crossref\]](#)
33. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet*. 2018;392(10162):2388-2396. [\[Crossref\]](#)
34. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. 2018;24(9):1337-1341. [\[Crossref\]](#)
35. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683. [\[Crossref\]](#)
36. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit Med*. 2023;6:74. [\[Crossref\]](#)
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998-6008. [\[Crossref\]](#)
38. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR); 2021. [\[Crossref\]](#)
39. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: transformers for 3D medical image

- segmentation. In: Proceedings of the IEEE/ CVF Winter Conference on Applications of Computer Vision (WACV); 2022. p. 1748-1758. [\[Crossref\]](#)
40. Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. *Magn Reson Med*. 2018;79(6):3055-3071. [\[Crossref\]](#)
41. Knoll F, Murrell T, Sriram A, et al. Advancing machine learning for MR image reconstruction with an open competition: overview of the 2019 fastMRI challenge. *Magn Reson Med*. 2020;84(6):3054-3070. [\[Crossref\]](#)
42. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*. 2025;645(8081):633-638. [\[Crossref\]](#)
43. Fan Z, Liang C, Wu C, Zhang Y, Wang Y, Xie W. Advancing radiology foundation models with reasoning through step-by-step verification from daily reports. *Commun Med (Lond)*. 2026;6(1):449. [\[Crossref\]](#)
44. Huang X, Wu J, Liu H, Tang X, Zhou Y. M1: unleash the potential of test-time scaling for medical reasoning with large language models. In: Proceedings of the Fifth Machine Learning for Health Symposium. *Proc Mach Learn Res*. 2026;297:369-383. [\[Crossref\]](#)
45. Byun JY, Park YJ, Azizan N, Chellappa R. Test-time scaling in clinical decision making. In: Proceedings of the 9th International Conference on Medical Imaging with Deep Learning. *Proc Mach Learn Res*. 2026;315:2118-2142. [\[Crossref\]](#)
46. Oh G, Kim S, Park S, Kim BH. Model and task-aware test-time scaling strategies for large language and vision-language models in medicine: evaluation study. *J Med Internet Res*. 2026;28:e90693. [\[Crossref\]](#)
47. Li Y, Zheng J, Tian Y, et al. Explicit chain of diagnosis for CMR semantic interpretation via large reasoning model. *NPJ Digit Med*. 2026. [\[Crossref\]](#)
48. Fallahpour A, Ma J, Munim A, Lyu H, Wang B. MedRAX: medical reasoning agent for chest X-ray. In: Proceedings of the 42nd International Conference on Machine Learning. *Proc Mach Learn Res*. 2025;267:15661-15676. [\[Crossref\]](#)
49. Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Adv Neural Inf Process Syst*. 2023;36:74952-74965. [\[Crossref\]](#)
50. Amiraz C, Cuconasu F, Filice S, Karnin Z. The distracting effect: understanding irrelevant passages in RAG. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2025. p. 18228-18258. [\[Crossref\]](#)
51. Huang J, Chen X, Mishra S, et al. Large language models cannot self-correct reasoning yet. In: International Conference on Learning Representations (ICLR); 2024. [\[Crossref\]](#)
52. Durak G, Aktas HE, Akinci D'Antonoli T, Bagci U. Explainable artificial intelligence in medical imaging: how to interpret, evaluate, and use artificial intelligence explanations. *Diagn Interv Radiol*. 2026. [\[Crossref\]](#)
53. Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P. Revisiting the trustworthiness of saliency methods in radiology AI. *Radiol Artif Intell*. 2024;6(1):e220221. [\[Crossref\]](#)
54. Rezazade Mehri MH, Mol F, Peter M, Ranschaert E, et al. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Sci Rep*. 2023;13(1):9230. [\[Crossref\]](#)
55. Gommers J, Hernström V, Josefsson V, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *Lancet*. 2026;407(10527):505-514. [\[Crossref\]](#)
56. Eisemann N, Bunk S, Mukama T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat Med*. 2025;31(3):917-924. [\[Crossref\]](#)
57. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and Food and Drug Administration staff. Final guidance. Silver Spring (MD): U.S. Food and Drug Administration; 2025 Aug. Accessed 2026 Sep 4. [\[Crossref\]](#)
58. European Parliament, Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024;L 2024/1689. Accessed 2026 Sep 4. [\[Crossref\]](#)
59. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. [\[Crossref\]](#)
60. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2024. Accessed 2026 Sep 4. [\[Crossref\]](#)
61. Durak G, Medetalibeyoglu A, Cicek V, Keles E, Bagci U. Beyond autonomy: why medicine needs artificial intelligence teammates, not artificial intelligence doctors. *Diagn Interv Radiol*. 2026. [\[Crossref\]](#)
62. Khosravi B, Rouzrokh P, Akinci D'Antonoli T, et al. Agentic AI in radiology: evolution from large language models to future clinical integration. *Radiol Artif Intell*. 2026;8(2):e250651. [\[Crossref\]](#)
63. Tzanis E, Adams LC, Akinci D'Antonoli T, et al. Agentic systems in radiology: principles, opportunities, privacy risks, regulation, and sustainability concerns. *Diagn Interv Imaging*. 2026;107(1):7-16. [\[Crossref\]](#)
64. Chinniah P, Gibson J, Chapala S, Saran S, Botchu R. Review of agentic artificial intelligence (AI) in radiology: from current clinical integration to future innovations. *Clin Radiol*. 2026;97:107351. [\[Crossref\]](#)
65. LeCun Y. A path towards autonomous machine intelligence. Version 0.9.2. OpenReview; 2022 Jun 27. OpenReview ID: BZ5a1r-kVsf. Accessed 2026 Sep 4. [\[Crossref\]](#)
66. Assran M, Bardes A, Fan D, et al. V-JEPA 2: self-supervised video models enable understanding, prediction and planning [Preprint]. arXiv. 2025. [\[Crossref\]](#)
67. Gorenshstein A, Omar M, Barash Y, Nadkarni GN, Klang E. Neuro-symbolic LLM Integration in Clinical Medicine: A Systematic Review and Taxonomy. *Res Sq* [Preprint]. 2026;rs.3.rs-8902022. [\[Crossref\]](#)
68. Kumari P, Chauhan J, Bozorgpour A, Huang B, Azad R, Merhof D. Continual learning in medical image analysis: a comprehensive review of recent advancements and future prospects. *Med Image Anal*. 2025;106:103730. [\[Crossref\]](#)
69. Lesaunier A, Khlaoui J, Dancette C, Tselikas L, Bonnet B, Boeken T. Artificial intelligence in interventional radiology: Current concepts and future trends. *Diagn Interv Imaging*. 2025;106(1):5-10. [\[Crossref\]](#)