Interface Design, Attention, and Overreliance in Clinical XAI.

Abstract: Background: Clinical explainable artificial intelligence (XAI) is commonly evaluated as explanation content: text, probability, saliency map, feature ranking, counterfactual, or analogous case. In practice, an explanation is encountered through an interface that controls when advice appears, where attention is directed, how much effort verification requires, and whether acceptance or rejection is documented.
Materials and Methods: This narrative review synthesised DOI-indexed literature on clinician-centred XAI design, trust in AI-based clinical decision support systems, eye tracking, explanation type, cognitive forcing, false confirmation, advice-taking, overreliance, and reporting guidance for clinical AI.
Results: The evidence indicates that explanation effects depend on timing, visibility, task risk, attention, user expertise, AI correctness, and interaction cost. Explanations can improve clinical reasoning, but may also anchor decisions, increase inappropriate reliance, or reassure users when data are incomplete.
Conclusion: Clinical XAI should be designed as an attention architecture rather than a passive interpretability layer. The proposed Attention-Gated XAI framework uses clinical pre-commitment, adaptive timing, risk-stratified friction, counterevidence prompts, missing-data display, acceptance and rejection logging, and post-decision audit to reduce overreliance while preserving useful AI support.

Keywords:
Clinical XAI; Interface Design; Human-AI Interaction; Attention; Cognitive Friction; Trust Calibration; Clinical Decision Support; Overreliance

Introduction:

The clinical value of an artificial intelligence recommendation is not determined at the moment a model generates its output. It is determined when a clinician sees that output, compares it with a working hypothesis, decides whether to inspect the explanation, accepts or rejects the advice, and records the final clinical action. The interface is therefore not a neutral container. It is the pathway by which a model enters medical reasoning.

This point is frequently underestimated in explainable artificial intelligence (XAI). Clinical XAI is often discussed through explanation categories: heatmaps for imaging, feature importance for tabular models, textual rationales for large language models, counterfactuals for individual predictions, probability scores for risk models, and example-based explanations for pattern comparison. These categories describe what is displayed. They do not describe when, where, with what salience, under which workflow pressure, and with what behavioral demand the explanation is displayed.

The same explanation can support or distort clinical judgment depending on interface conditions. A recommendation displayed before a physician forms an initial judgment may anchor the case. A recommendation displayed after a physician has selected a diagnosis may become a justification. An explanation hidden behind a small icon may be absent in practice even though it is technically available. A long textual explanation placed inside a time-pressured workflow may be skipped. A probability score without context may be read as certainty. A counterevidence prompt may reopen the differential diagnosis. These effects are not properties of the model alone; they emerge from the encounter between the model, the interface, and the clinician.

Recent work supports this shift from interpretability as information to interpretability as interaction. Bienefeld and colleagues showed that clinicians and developers often bring different expectations to XAI, creating a gap between technical explanations and clinical needs [1]. Subramanian and colleagues identified the use of predictions, information accompanying predictions, personalization, and case customization as central concerns in medical XAI design [2]. Prince and colleagues likewise framed user-centered design as necessary for translating explainable systems into clinical decision support for central nervous system tumors [3]. Trust in AI-based clinical decision support also depends on transparency, reliability, usability, customization, human-centered design, and user control [4].

The problem becomes measurable when attention is studied rather than assumed. Nagendran and colleagues used eye tracking to examine how intensive care physicians responded to safe and unsafe AI recommendations with explanations [5]. Unsafe recommendations attracted attention, but explanations themselves did not necessarily receive greater attention when unsafe advice was presented. Prinster and colleagues showed that explanation type affected chest radiograph diagnostic performance and physician trust even when physicians were unaware of the effect [6]. Gombolay and colleagues found domain-specific

differences in how explanation approaches are perceived in neurology decision support [7]. The implication is direct: the mere presence of an explanation does not establish that the explanation was clinically processed.

This review develops the argument that clinical XAI should be evaluated as attention architecture. The aim is not to place an explanation beside a recommendation and assume safe interpretation. The aim is to design the conditions under which clinicians attend, compare, verify, disagree, accept, reject, and document. The proposed Attention-Gated XAI framework treats explanations as decision events rather than passive text. A physician does not merely receive AI advice. The physician is positioned inside an interaction that shapes whether agreement is easy, whether verification is feasible, and whether disagreement remains visible.

Materials and Methods:

This manuscript is a narrative review with a design framework. It synthesises DOI-indexed literature related to clinical XAI, AI-based clinical decision support systems, user-centered design, trust calibration, attention, overreliance, false confirmation, cognitive forcing, AI reporting guidance, and health AI assurance. The aim was not to estimate pooled effects. The aim was to identify design mechanisms that determine whether explanations shape safe clinical behavior.

Sources were identified through PubMed, Nature Portfolio, JMIR, Radiology, ACM Digital Library, ScienceDirect, Wiley Online Library, BMJ, JAMA, Frontiers, and publisher pages indexed through DOI records. Search concepts included clinical XAI, explainable AI clinical decision support, user-centered design, AI-CDSS, clinician trust, eye tracking, explanation type, attention, cognitive forcing, overreliance, false confirmation, advice-taking, human-AI team performance, pro-hoc explanations, TRIPOD+AI, TRIPOD-LLM, and health AI assurance.

Inclusion criteria were: DOI-indexed article; direct relevance to clinical decision support, clinical AI, XAI, clinician interaction, user-centered design, attention, trust, reliance, reporting standards, or AI governance; preference for publications from 2023 onward when available; and inclusion of older DOI-indexed studies when foundational to the design mechanism. Exclusion criteria were: preprints without DOI, purely algorithmic papers without clinician-facing or governance implications, texts about general AI ethics without clinical interface relevance, and papers treating explainability only as technical interpretability without human evaluation or workflow implications.

The synthesis followed five analytic questions. First, what does the literature indicate about clinician needs and stakeholder expectations for XAI? Second, how do timing, attention, and explanation type modify diagnostic or therapeutic behavior? Third, which interaction mechanisms increase or reduce overreliance? Fourth, which reporting and assurance standards are needed when explanations are evaluated in clinical AI studies? Fifth, which interface components are necessary if an explanation is treated as a clinical decision event rather than an information supplement?

Because the literature is heterogeneous, the review used mechanism-oriented synthesis. Studies were mapped to recurrent design functions: clinical pre-commitment, adaptive timing, risk-stratified friction, counterevidence prompting, missing-data display, acceptance and rejection logging, and post-decision audit. These functions became the components of the proposed Attention-Gated XAI framework.

Table 1: Literature map used for the thematic synthesis.

Cluster	Core references	Contribution to the review
Clinician-centered design	Bienefeld et al. [1]; Subramanian et al. [2]; Prince et al. [3]	Defines XAI as a socio-technical design problem involving clinicians, developers, context, workflow, and user needs.
Trust and usability in AI-CDSS	Tun et al. [4]; Naiseh et al. [9]; Rosenbacke et al. [10]	Shows that trust depends on transparency, reliability, usability, explanation class, complexity, coherence, and user control.
Attention and explanation type	Nagendran et al. [5]; Prinster et al. [6]; Gombolay et al. [7]	Demonstrates that explanation effects depend on attention, clinical domain, explanation format, and AI correctness.
Advice-taking and overreliance	Panigutti et al. [8]; Bucinca et al. [12]; Vasconcelos et al. [13]; Bansal et al. [14]	Connects explanation design to behavior, cognitive effort, team performance, and acceptance of AI advice.
False confirmation and clinical limits	Rosenbacke et al. [11]; Ghassemi et al. [15]; Cabitza et al. [16]	Frames the risk that explanations may justify incorrect conclusions or exceed proven safety effects.
Reporting and assurance	Collins et al. [17]; Gallifant et al. [18]; Shah et al. [19]; Goodman et al. [20]	Provides standards for transparent reporting, LLM evaluation, institutional assurance, and clinical documentation quality.

Results: Thematic Synthesis

1. Explanation content is inseparable from interface placement

A clinical explanation is usually described by content: variables that influenced the model, image regions that were highlighted, similar prior cases, counterfactual changes, or a textual rationale. Content is necessary, but it is only one dimension of use. A clinician-facing system must also decide whether the explanation appears before, during, or after independent judgment; whether it is visible by default or hidden behind a click; whether it interrupts workflow; whether it requires a response; whether it displays uncertainty; whether it forces comparison with alternatives; and whether acceptance or rejection is recorded.

Interface placement can convert the same explanatory object into different clinical behavior. A saliency map displayed before image search may guide attention to relevant anatomy, but it may also narrow search. A text explanation displayed after a diagnosis may improve confidence, but it may also rationalize a premature conclusion. A probability score displayed without calibration context may look precise while saying little about patient-specific applicability. The explanation therefore cannot be evaluated apart from its sequence, salience, and interaction cost.

This is where co-design literature becomes clinically important. Bienefeld and colleagues showed that explainability requirements differ between clinicians and developers [1]. Developers may focus on technical transparency and debugging, while clinicians may need patient-specific relevance, causal coherence, workflow fit, and actionability. Subramanian and colleagues identify prediction use, contextual information, personalization, and case customization as central design themes [2]. Prince and colleagues extend the point to neuro-oncology CDSS, where explainable tools must be aligned with user experience, translation, workflow, and outcome evaluation [3].

For clinical XAI, the question is not whether a model is interpretable in the abstract. The question is whether the interface makes interpretation cognitively available at the moment when the physician can still act on it. Explanation content without interface specification tells reviewers what was theoretically available, not what shaped the decision.

Table 2: Interface placement converts explanation content into clinical behavior.

Design variable	Clinical risk when ignored	Safer design implication
Timing of AI output	Premature recommendations anchor the clinician; late explanations justify decisions already made.	Delay AI advice until after a preliminary clinical hypothesis in diagnostic tasks; display earlier only when triage speed is the priority.
Default visibility	Hidden explanations become unused explanations.	Make core explanation visible by default when risk is intermediate or high; keep secondary detail expandable.
Response requirement	Passive displays invite passive agreement.	Require accept, reject, or revise when recommendation affects diagnosis, treatment, or disposition.
Uncertainty display	Point estimates are misread as certainty.	Show uncertainty, missing variables, calibration limits, and conditions under which the recommendation loses validity.
Alternative display	Single-output explanations suppress differential reasoning.	Display clinically plausible alternatives and the discriminating data between them.
Documentation	Agreement and disagreement remain invisible to audit.	Log acceptance, override, modification, rationale, and later outcome review.

2. Attention is a measurable clinical variable

Many XAI evaluations rely on self-reported usefulness, perceived trust, or satisfaction. These instruments are valuable, but they are insufficient. A clinician may describe explanations as useful while giving them minimal attention during the decision. A clinician may also be influenced by an explanation without awareness of that influence. The gap between perceived use and behavioral use is a design problem, not only a measurement problem.

Nagendran and colleagues separated explanation presence from attention to explanation by using eye tracking in intensive care physicians exposed to safe and unsafe AI recommendations [5]. Unsafe recommendations drew more attention than safe recommendations, but unsafe scenarios did not necessarily produce greater attention to the explanatory material. Self-reported usefulness also failed to correlate reliably with attention to explanations. The study demonstrates why an explanation should not be treated as used simply because it existed on the screen.

Prinster and colleagues show a related effect in chest radiography: explanation type modified diagnostic performance and trust, even when physicians were not aware that the explanation had influenced them [6]. Explanations may therefore operate beneath explicit self-report. Gombolay and colleagues add that explanation perception varies by clinical domain and user type [7]. These findings support the use of

objective and behavioral measurements: first fixation, dwell time, transitions between patient data and explanation, click path, time from advice to action, and the quality of post-decision rationale.

Attention is not a surrogate for correctness, but it is a precondition for processing. A heatmap that is not seen, a text explanation that is skipped, or an uncertainty panel that is hidden cannot calibrate clinical decision-making. Clinical XAI studies should therefore report what users could see, what they actually attended to, and whether their final decision changed in a clinically defensible direction.

Table 3: Attention metrics for clinician-facing XAI.

Metric	Interpretation	Use in evaluation
First fixation on AI advice	Whether the recommendation enters visual attention early or late.	Detects hidden, poorly positioned, or overly salient recommendations.
Dwell time on explanation	Approximate visual processing time for explanatory material.	Distinguishes displayed explanations from attended explanations.
Transitions between patient data and explanation	Whether the clinician compares AI output against clinical evidence.	Measures verification rather than passive reading.
Click or expansion behavior	Whether optional explanation layers are accessed.	Identifies whether secondary details are usable or effectively absent.
Time from AI display to decision	Whether AI accelerates, delays, or prematurely closes the decision.	Useful for workflow safety and cognitive load analysis.
Rationale concordance	Whether documented reasoning reflects the explanation or independent clinical evidence.	Connects attention to final decision quality.

3. Cognitive friction is a safety dose

Clinical interfaces are often designed to reduce friction. This is appropriate for routine documentation and administrative tasks. It becomes unsafe when friction reduction eliminates clinical thinking in high-risk AI-assisted decisions. Some friction is protective. The design question is dose. Insufficient friction makes agreement automatic. Excessive friction creates alert fatigue, workarounds, and abandonment.

Bucinca and colleagues showed that cognitive forcing functions can reduce overreliance in AI-assisted decision-making [12]. Their work also shows a trade-off: interventions that reduce overreliance may be less liked by users. This trade-off is acceptable in high-risk care when the cost of unexamined agreement is patient harm. Vasconcelos and colleagues add that users decide whether to engage with an explanation by weighing cognitive cost against expected benefit [13]. In medicine, that cost-benefit calculation is shaped by fatigue, time pressure, case difficulty, institutional expectations, and perceived liability.

Friction should therefore be adaptive. A low-risk coding suggestion may require a concise optional rationale. A diagnostic recommendation in an ambiguous case should show alternatives and ask for counterevidence. A high-risk prescription should require patient-specific safety checks and a reason for acceptance or override. A recommendation that conflicts with allergy, renal function, pregnancy status, hemodynamics, image quality, or missing essential data should trigger stronger friction than a low-impact reminder.

The aim is not to slow clinicians for its own sake. The aim is to make verification easier than blind agreement when the stakes are high. If the interface makes accepting advice easier than checking it, overreliance becomes a predictable design outcome.

Table 4: Risk-stratified friction model for clinical XAI.

Risk level	Example	Interface requirement
Low	Administrative coding suggestion; low-impact documentation support.	Optional explanation; no interruption; visible confidence and source link if opened.
Moderate	Diagnostic suggestion in non-urgent outpatient setting.	Evidence summary, alternative diagnosis, and one counterevidence prompt.
High	Anticoagulation, sepsis therapy, thrombolysis, ICU vasopressor guidance.	Clinical pre-commitment, uncertainty display, accept/reject rationale, and safety checklist.
Critical conflict	AI recommendation conflicts with allergy, contraindication, missing essential data, or patient-specific risk.	Block automatic acceptance; require human override rationale, escalation, or second review.
Post-event review	Adverse event, unexpected deterioration, reversal of diagnosis, or model disagreement with outcome.	Audit AI advice, human rationale, available data, and patient trajectory.

4. Trust should be designed as calibration, not acceptance

Trust is often treated as an adoption barrier. From that perspective, explanations are useful because they make clinicians more willing to use AI. In safety-critical settings, this framing is too narrow. The

problem is not whether clinicians trust AI. The problem is whether they trust it in proportion to validity, uncertainty, and fit to the case.

Rosenbacke and colleagues found that XAI may increase clinician trust, but the effect is not automatic; explanations can also have no effect or decrease trust when they are complex, incoherent, or poorly aligned with the clinical problem [10]. Tun and colleagues identified transparency, training, usability, reliability, credibility, ethics, customization, human-centered design, and user control as recurring factors shaping trust in AI-based clinical decision support [4]. Naiseh and colleagues showed that explanation classes affect trust calibration in a clinical decision support context [9]. Together, these studies point away from trust maximization and toward trust calibration.

A safe interface should sometimes reduce trust. If a recommendation is generated from missing data, out-of-distribution input, conflicting evidence, poor image quality, weak calibration, or a context outside the model's validation, the interface should not reassure. It should warn. If the system identifies pneumonia while vital signs and laboratory trajectory suggest sepsis, the interface should prevent single-diagnosis closure. If a radiology model highlights a region with low confidence and poor image quality, the system should not invite unqualified acceptance.

Acceptance rates are therefore poor primary safety metrics. High acceptance may signal usefulness or automation bias. Low acceptance may signal poor model performance or appropriate skepticism. The relevant outcome is conditional behavior: correct acceptance of useful advice and correct rejection of unsafe, incomplete, or inapplicable advice.

5. False confirmation is an interface-sensitive hazard

False confirmation occurs when AI validates a human judgment that is wrong or incomplete. The system does not need to replace the clinician. It only needs to confirm a premature working hypothesis strongly enough that reconsideration becomes less likely. The risk is heightened when the AI presents a fluent explanation that uses familiar clinical terms and contains partially true observations.

Rosenbacke and colleagues argue that false confirmation is a central danger when AI is positioned as a substitute for a second medical opinion [11]. The interface determines whether this risk is amplified or mitigated. If AI appears before the clinician has generated alternatives, false confirmation may prevent alternatives from forming. If it appears after a diagnosis has been selected, it may validate the existing path. If the interface fails to ask for counterevidence, the clinician may not examine what would disconfirm the shared conclusion.

Cabitza and colleagues' work on pro-hoc explanations offers a useful design direction: present alternative explanations for possible outcomes rather than only a post-hoc justification for the recommended output [16]. This approach is clinically relevant because differential reasoning depends on contrast. A display that explains only why AI favours pneumonia narrows attention. A display that compares pneumonia, heart failure, atelectasis, aspiration, and pulmonary embolism with discriminating findings preserves reasoning.

Bansal and colleagues showed that explanations do not necessarily improve complementary human-AI team performance and may increase acceptance of AI recommendations irrespective of correctness [14]. In clinical systems, this is unacceptable. The physician and the model should not fail in the same direction without a mechanism for detecting shared error.

Table 5: Interface patterns that increase or reduce false confirmation.

Pattern	Likely effect	Safer alternative
AI advice shown before clinician hypothesis	Anchors diagnostic search and narrows differential reasoning.	Require initial hypothesis and confidence rating before revealing AI advice when the task permits.
Single recommended diagnosis with persuasive explanation	Converts uncertainty into apparent confirmation.	Display alternatives and discriminating clinical evidence.
High-salience confidence score without context	Promotes over-interpretation of numeric precision.	Show calibration context, missing data, and validity conditions.
No cost for accepting AI advice	Makes agreement easier than verification.	Require rationale for high-risk agreement and provide counterevidence prompt.
No record of disagreement	Overrides and uncertainty disappear from safety review.	Log accept, reject, revise, rationale, and follow-up outcome.

6. Reporting and assurance are part of interface safety

Clinical XAI interface studies require stronger reporting than many current evaluations provide. A study should state what the clinician saw first, whether the clinician committed to a judgment before AI advice, whether explanations were visible by default, what data were missing, whether the AI advice was correct, partially correct, or incorrect, and how acceptance or rejection was captured. Without these details, readers cannot determine whether a reported effect belongs to the model, the explanation content, the display timing, or the interaction requirement.

TRIPOD+AI provides updated reporting guidance for clinical prediction models using regression or machine learning and emphasises transparent, complete, and accurate reporting [17]. TRIPOD-LLM extends reporting principles to studies using large language models [18]. These reporting frameworks are not interface-design manuals, but they reinforce the same safety principle: clinical AI research must disclose enough information for users, reviewers, and institutions to understand what was evaluated and under which conditions. For XAI interface studies, that means reporting the interaction pathway, not only the algorithm.

Assurance also matters after deployment. Shah and colleagues proposed a nationwide network of health AI assurance laboratories to support local evaluation, monitoring, and best practices [19]. This is directly relevant to clinical XAI because interface behavior may vary by institution, specialty, workflow, user role, and patient population. A system that works in a simulation may behave differently in a high-volume emergency department or an ICU with different documentation culture.

Goodman and colleagues argued that AI-generated clinical summaries require more than accuracy [20]. The same principle applies to XAI interfaces. An explanation may be factually correct yet clinically incomplete, poorly timed, too difficult to verify, or too easy to accept. Safety depends on how information changes action. Reporting and assurance should therefore include attention, reliance, override behavior, rationale quality, and downstream decision quality.

Table 6: Minimum reporting items for clinical XAI interface studies.

Reporting item	Reason for inclusion
Timing of AI display	Determines whether advice could anchor, revise, or confirm clinician judgment.
Clinician pre-AI hypothesis	Allows measurement of whether AI changed the decision or merely reinforced it.
Default visibility of explanation	Separates available explanations from used explanations.
Attention measures	Documents whether explanatory material was visually or behaviorally processed.
AI correctness and confidence	Needed to distinguish appropriate reliance from overreliance or underreliance.
Missing data and applicability limits	Prevents retrospective inflation of system performance using data unavailable at the bedside.
Acceptance, rejection, modification	Captures real clinician behavior rather than subjective trust alone.
Rationale and outcome audit	Links interface behavior to downstream clinical quality and learning.

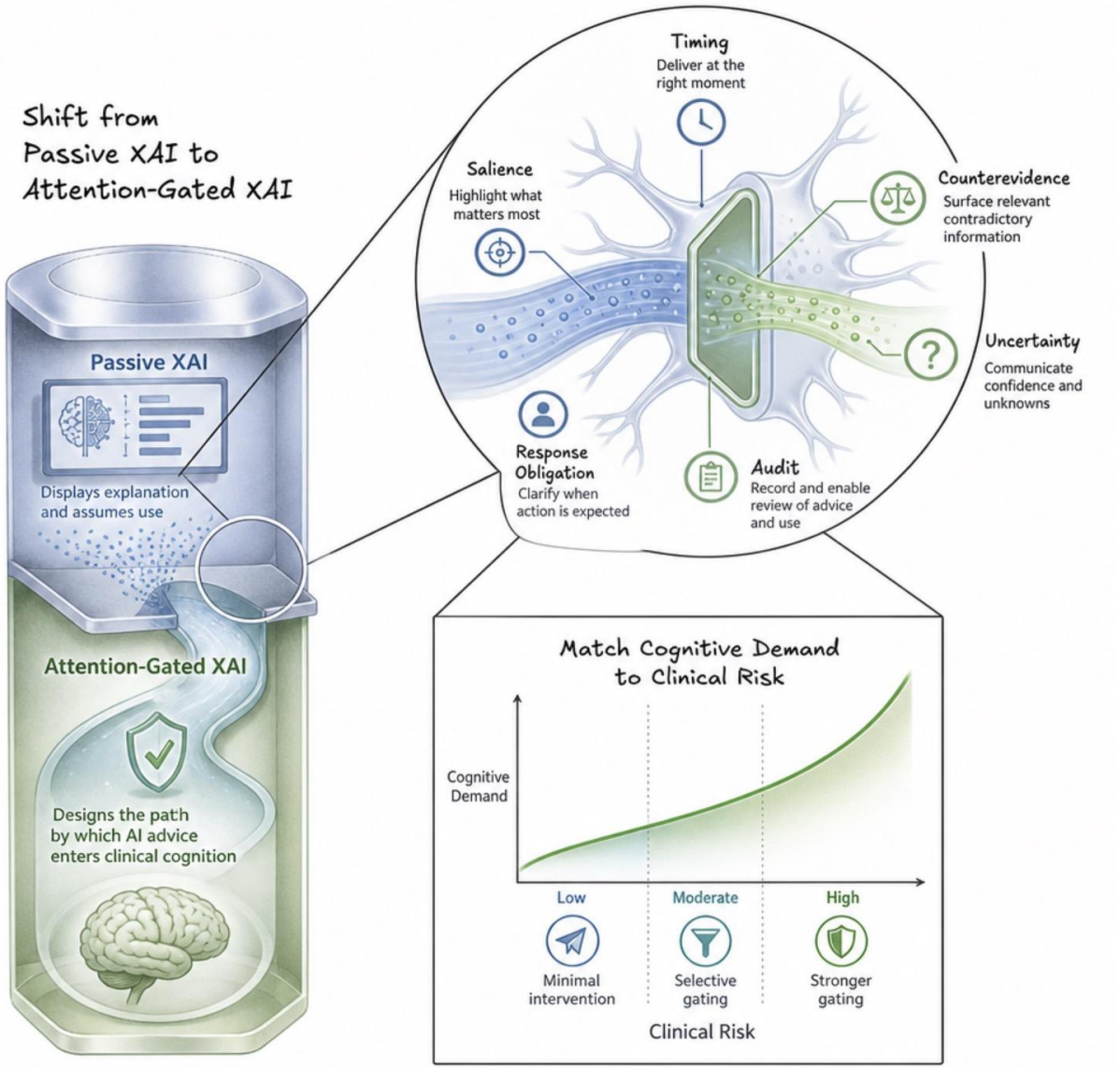
Attention-Gated XAI Framework:

The synthesis supports a shift from passive XAI to Attention-Gated XAI. Passive XAI displays an explanation and assumes that clinicians will use it. Attention-Gated XAI designs the route by which AI advice enters clinical cognition. It regulates timing, salience, response obligation, counterevidence, missing data, and audit. The purpose is not to slow every decision. The purpose is to match cognitive demand to clinical risk.

The framework has seven components: clinical pre-commitment, adaptive timing, risk-stratified friction, counterevidence prompting, missing-data display, acceptance and rejection logging, and post-decision audit. These components convert an explanation from an optional interpretability artifact into a structured decision event.

Clinical pre-commitment asks the physician to record a provisional hypothesis, confidence level, or intended action before AI advice is displayed when the task allows. Adaptive timing determines whether AI advice should appear before diagnostic search, after a preliminary interpretation, before order signature, or during reassessment. Risk-stratified friction increases response demand as clinical consequence increases. Counterevidence prompting asks the clinician what would weaken the recommendation. Missing-data display prevents incomplete inputs from being experienced as complete advice. Acceptance and rejection logging preserves clinician agency as audit data. Post-decision audit links AI advice, human rationale, and later clinical trajectory.

Imagem 1. Passive XAI to Attention-Gated XAI



This framework is compatible with, but distinct from, model-centered explainability. It does not ask only whether the model can be interpreted. It asks whether the interface supports safe human-AI complementarity under clinical pressure.

Table 7: Attention-Gated XAI framework.

Component	Purpose	Operational design
Clinical pre-commitment	Preserve independent reasoning before AI influence.	Capture provisional hypothesis, confidence, and intended next step before AI display.
Adaptive timing	Align AI display with cognitive stage and clinical risk.	Delay, advance, or repeat explanations according to task type and urgency.
Risk-stratified friction	Prevent automatic agreement in high-risk decisions.	Escalate from optional reading to required rationale according to severity.
Counterevidence prompt	Activate disconfirmation and differential reasoning.	Ask what finding weakens the AI recommendation or supports an alternative.
Missing-data display	Prevent false certainty when input is incomplete.	Show absent, stale, low-quality, or out-of-distribution data beside the recommendation.
Acceptance/rejection log	Make clinician agency visible and auditable.	Record accept, reject, modify, rationale, and whether AI changed the initial plan.
Post-decision audit	Learn from disagreement and outcome.	Compare AI output, human rationale, final decision, and later clinical trajectory.

Clinical Scenarios Illustrating Interface Failure Modes:

Radiology illustrates visual anchoring. An AI system may mark a focal opacity and provide a heatmap. If the heatmap appears before the radiologist performs a systematic search, it may narrow exploration. If it appears after the independent read, it may function as verification. If it appears without

image-quality information or calibration context, the highlighted region may be over-weighted. The work by Prinster and colleagues suggests that explanation type can change performance and trust in chest radiograph interpretation [6]. A radiology interface should therefore distinguish search assistance from interpretive confirmation.

Intensive care illustrates high-risk action under time pressure. AI recommendations may concern fluids, vasopressors, antibiotics, ventilation, anticoagulation, transfusion, or monitoring intensity. Nagendran and colleagues show why the presence of explanations cannot be assumed to rescue unsafe advice [5]. ICU interfaces should require patient-specific safety checks and counterevidence review before high-risk acceptance, especially when advice conflicts with physiology, organ function, or missing data.

Emergency medicine illustrates premature closure. A system that flags pulmonary embolism, sepsis, stroke, or acute coronary syndrome may accelerate recognition. The same system may anchor the team if displayed too early or with excessive confidence. Interfaces should distinguish detection prompts from diagnostic claims. 'Consider sepsis' and 'sepsis is the most likely diagnosis' are different clinical acts and should not share the same visual grammar.

Medication ordering illustrates hidden harm. A system may recommend a drug correctly but fail to communicate dose adjustment, interaction, contraindication, or monitoring requirements. A safe interface should display missing creatinine, allergy history, pregnancy status, current interacting drugs, and required monitoring before order signature. Agreement should be decomposed: accept indication, modify dose, add monitoring, reject agent, or request more data.

Outpatient chronic care illustrates cumulative automation. Repeated acceptance of risk scores, reminders, screening suggestions, or deprescribing prompts can quietly reshape care over time. Here, intense friction at every decision would fail. The interface should instead detect exceptions, uncertainty, patient preference, and longitudinal deviation. The same principle holds: interface demand should match risk.

Table 8: Clinical scenario mapping for Attention-Gated XAI.

Clinical setting	Typical AI risk	Interface requirement
Radiology	Visual anchoring, incomplete search, over-weighting saliency.	Separate independent read from AI review; display image quality, uncertainty, and alternative findings.
Intensive care	Unsafe treatment recommendation under time pressure.	Require patient-specific safety checks, counterevidence, and rationale before high-risk acceptance.
Emergency medicine	Premature diagnostic closure or triage misdirection.	Use consider-this prompts before diagnostic claims; delay definitive advice until key data are reviewed.
Outpatient chronic care	Cumulative acceptance of low-salience recommendations.	Use longitudinal audit, exception capture, and patient preference documentation.
Medication ordering	Dose, contraindication, interaction, or monitoring failure.	Display missing data, renal/hepatic adjustment, contraindications, and required monitoring before signature.
Multidisciplinary board	AI framed as authority rather than evidence contributor.	Show evidence provenance, uncertainty, dissenting alternatives, and team annotation.

Proposed Empirical Evaluation Protocol:

The Attention-Gated XAI framework should be tested experimentally before clinical deployment. A practical design would use simulated cases with practicing clinicians and randomize interface timing and friction while holding AI advice constant. A factorial design could compare AI advice before versus after clinician pre-commitment; explanation hidden versus visible by default; no friction versus counterevidence prompt versus mandatory rationale; correct versus incorrect AI advice; and low-risk versus high-risk decisions.

Primary outcomes should include appropriate reliance: accepting correct AI advice and rejecting incorrect advice. Secondary outcomes should include diagnostic accuracy, treatment safety, time to decision, unnecessary actions, attention to explanation, confidence calibration, subjective workload, perceived usefulness, and quality of documented rationale. A study that measures only trust misses the central safety mechanism. A study that measures only accuracy misses why the decision changed.

The study should capture expertise. Medical students, residents, generalists, specialists, radiologists, intensivists, pharmacists, and nurses may need different explanation formats and friction levels. A junior clinician may benefit from explicit differential diagnosis. A senior clinician may need concise uncertainty and contraindication data. A radiologist may need image-based evidence. A pharmacist may need dose and interaction logic. Personalization should be specified rather than assumed.

Evaluation should include adverse interaction outcomes: acceptance of incorrect AI advice, rejection of correct advice, failure to view explanation, excessive time burden, increased confidence without improved accuracy, displacement of patient data by AI display, and documentation that copies AI rationale without independent clinical justification. These outcomes are not peripheral. They are the failure modes that determine whether the system is safe outside a controlled study.

Table 9: Experimental evaluation matrix for clinical XAI interfaces.

Factor	Levels to compare	Safety question answered
Advice timing	Before pre-commitment; after pre-commitment; at order signature.	Does timing anchor, revise, or verify clinician decision?
Explanation visibility	Hidden; optional; visible by default; mandatory in high risk.	Does availability translate into actual attention and use?
Friction level	None; counterevidence prompt; mandatory rationale; second review.	Which dose reduces overreliance without unacceptable burden?
AI correctness	Correct; incorrect; partially correct; correct with weak explanation.	Can clinicians reject unsafe advice and accept useful advice?
Risk level	Low; moderate; high; critical conflict.	Should the interface change its demands with harm potential?
User expertise	Student; resident; generalist; specialist; multidisciplinary team.	Which explanation format fits which clinical user?

Design and Governance Requirements:

Translation of Attention-Gated XAI into practice requires obligations for developers, hospitals, evaluators, and governance boards. Developers should not treat explanation widgets as optional post-processing modules. They should define the clinical action that each explanation supports: notice, compare, verify, prescribe, withhold, escalate, or document. This action must be known before the interface is built; otherwise, explanations may be technically correct but clinically inert.

Hospitals should treat AI interface configuration as part of clinical governance. Institutions deploying AI-based decision support should decide which recommendations are advisory, which are interruptive, which require a response, which require second review, and which should be suppressed until required data are available. Governance should monitor acceptance, override patterns, missing-data warnings, time burden, near misses, and clinician reports of misleading explanations.

The health AI assurance literature supports this approach. Shah and colleagues argue for infrastructure capable of evaluating and monitoring health AI in local contexts [19]. This matters because interface effects are local. A sepsis model deployed in a teaching hospital ICU, a community emergency department, and a tertiary oncology service may have the same algorithmic output but different user behavior. Assurance should therefore include interface behavior, not only model discrimination or calibration.

A minimal governance checklist should answer five questions before high-risk deployment: when does the clinician see the AI; what must the clinician do before accepting it; what missing data are displayed; how does the clinician reject or modify the recommendation; and how are disagreements reviewed? If these questions cannot be answered, the system is not ready for high-risk clinical use regardless of model performance.

Table 10: Design and governance requirements for clinical XAI deployment.

Actor	Required responsibility	Operational artifact
Developers	Define the clinical action supported by each explanation.	Explanation-action map: notice, compare, verify, prescribe, escalate, document.
Developers	Expose missing data and model applicability limits.	Validity panel beside AI recommendation.
Hospitals	Classify AI recommendations by clinical risk and workflow location.	Local AI-CDSS risk policy.
Hospitals	Monitor acceptances, overrides, and disagreement patterns.	Monthly safety review of AI-human interaction logs.
Clinical leaders	Align friction level with specialty workflow.	Specialty-specific interface configuration.
Evaluators	Measure behavior in addition to subjective trust.	Study protocol including attention, reliance, rationale, and outcomes.
Governance board	Define escalation for high-risk or conflicting recommendations.	Override and second-review procedure.

Implementation Pathway:

A practical implementation pathway can be organized in four phases. The first phase is decision mapping. The team identifies the decision points where AI advice will appear and classifies each point by risk, time pressure, reversibility, and data dependency. This phase prevents a common implementation error: using the same display pattern for a low-risk reminder and a high-risk treatment recommendation.

The second phase is explanation-action alignment. Each explanation must be tied to a clinical action. If the action is diagnostic comparison, the explanation should include alternatives and discriminating evidence. If the action is prescription, the explanation should include contraindications, dosing logic, and monitoring requirements. If the action is triage, the explanation should prioritize urgency, uncertainty, and missing data. This phase avoids the generic explanation box that sounds plausible but guides no decision.

The third phase is simulation testing. Clinicians should interact with realistic cases under varied time pressure and varied AI correctness. The test set should include obvious cases, ambiguous cases, misleading AI advice, incomplete data, and cases where AI is correct but conflicts with the clinician's initial impression. The purpose is to observe what clinicians saw, what they ignored, what they accepted, what they rejected, and what they wrote.

The fourth phase is monitored deployment. If a prototype enters clinical workflow, the institution should monitor interaction logs, override patterns, time burden, alert fatigue, near misses, and reported confusion. Deployment is not a single approval event. It is a continuous feedback system in which clinicians, developers, and governance teams examine how AI changes decision behavior over time.

Table 11: Four-phase pathway for Attention-Gated XAI implementation.

Phase	Core activity	Expected output
1. Decision mapping	Identify decision points, risk level, time pressure, reversibility, and required data.	Clinical AI decision map.
2. Explanation-action alignment	Match explanation type to clinical action and user role.	Interface specification for each AI recommendation type.
3. Simulation testing	Test correct and incorrect AI advice under realistic cases and workflow constraints.	Reliance, attention, accuracy, workload, and rationale report.
4. Monitored deployment	Review acceptances, overrides, near misses, time burden, and clinician feedback.	Governance dashboard and iteration plan.

Why Current XAI Evaluations Miss Interface Effects:

A recurring limitation in XAI evaluation is that explanation is treated as an isolated object rather than a situated clinical event. A study may compare heatmaps, feature attribution, counterfactuals, examples, and textual rationales while underreporting the route by which clinicians encounter them. When performance improves, the cause may be explanation content, timing, salience, task framing, perceived AI authority, or a demand to respond. When performance worsens, the cause may be cognitive overload, poor placement, excessive confidence display, weak alternatives, or explanation mismatch with clinical work.

This attribution problem matters because the same model and the same explanation can produce different behavior under different interface states. A radiology explanation presented as optional review after an independent read is not the same intervention as the same explanation presented before search. A sepsis prediction shown as a banner before full vital-sign review is not the same as one shown after lactate, mental status, blood pressure, urine output, and infection source are reviewed. A medication warning displayed before order signature is not the same as one displayed after the clinician has already completed the order and is trying to close the chart.

The literature on advice-taking and overreliance indicates that users do not process AI explanations as neutral information. They process them through prior beliefs, clinical intuition, workload, perceived risk, and system framing [8,12-14]. Therefore, an interface that reduces verification cost changes reliance. An interface that hides explanation detail changes reliance. An interface that displays one conclusion without alternatives changes reliance. These are not usability details. They are causal features of the decision environment.

Future studies should therefore report the interaction pathway with the same care used to report model inputs and outputs. The report should identify whether AI advice appeared before or after clinician judgment, whether explanation was default-visible, whether counterevidence was requested, whether users could ignore the advice without record, whether uncertainty was displayed, and whether the final human rationale was captured. Without these elements, readers cannot determine what was actually tested.

This is also why reporting frameworks such as TRIPOD+AI and TRIPOD-LLM are relevant even when the manuscript is not building a new model [17,18]. They reinforce a broader principle: clinical AI studies must describe enough of the system, data, task, and evaluation context to allow interpretation and reproduction. For clinician-facing XAI, the interface is part of that context.

Table 12: Common XAI evaluation blind spots and corresponding interface variables.

Evaluation blind spot	Why it matters clinically	Variable to report or test
Explanation treated as content only	The same content changes effect according to timing and salience.	Timing, location, default visibility, interruption level.
Trust measured without correctness state	Trust may be high for correct and incorrect advice alike.	Trust stratified by AI correctness and case uncertainty.
Self-report used without behavior	Clinicians may report usefulness without attending to explanations.	Eye tracking, dwell time, clicks, decision changes, rationale text.
No clinician pre-commitment	The study cannot distinguish AI revision from AI anchoring.	Pre-AI hypothesis, confidence, and intended action.
No reject/override analysis	The most important safety behavior is hidden.	Acceptance, rejection, modification, reason, and outcome.
No workflow context	Time pressure and specialty workflow alter use of explanations.	Clinical setting, urgency, user expertise, and task complexity.

Minimum Acceptance Criteria for High-Risk Clinical XAI Interfaces:

Before a high-risk clinical XAI interface is deployed, the institution should define minimum acceptance criteria. Model performance alone is insufficient. An AI system that is well calibrated in retrospective validation may still be unsafe if the interface promotes premature closure, hides missing data, or makes acceptance easier than verification. Conversely, a system with modest but transparent decision support may be safer if clinicians understand its limits and can reject advice efficiently.

The first criterion is scope clarity. The interface should state what clinical action the recommendation supports and what action it does not support. A risk estimate is not a treatment order. A detection alert is not a diagnosis. A diagnostic suggestion is not a full therapeutic plan. Visual grammar should separate these categories so the clinician does not confuse possibility, probability, and instruction.

The second criterion is data completeness at the point of display. If an AI recommendation depends on laboratory values, image quality, medication history, allergy status, pregnancy status, renal function, or clinical timing, the interface should show whether those data are present, absent, stale, or outside expected range. An explanation that ignores missing data communicates more confidence than the evidence supports.

The third criterion is actionable disagreement. A physician must be able to reject, modify, or defer AI advice without leaving the decision unrecorded. The system should ask for a concise clinical reason in high-risk cases: missing data, contraindication, stronger alternative diagnosis, local protocol, patient preference, or low plausibility. This does not turn clinicians into data clerks. It converts disagreement into safety intelligence.

The fourth criterion is post-decision learning. The system should support audit of accepted and rejected advice against later information. This is especially important when the AI advice was wrong but accepted, correct but rejected, or modified by an expert. These cases identify whether the problem lies in model performance, interface design, local workflow, missing data, or clinical training.

Table 13: Minimum acceptance criteria before high-risk clinical XAI deployment.

Criterion	Operational requirement	Safety rationale
Scope clarity	State whether the output is alert, risk estimate, diagnosis, treatment suggestion, or monitoring prompt.	Prevents users from treating one type of output as a stronger clinical instruction.
Data completeness	Display absent, stale, low-quality, and out-of-scope inputs at the point of advice.	Prevents false certainty when the recommendation rests on incomplete context.
Uncertainty and alternatives	Show uncertainty and clinically plausible alternatives with discriminating evidence.	Preserves differential reasoning and reduces premature closure.
Actionable disagreement	Allow accept, reject, modify, defer, and request more data with concise clinical rationale.	Makes clinician agency visible and auditable.
Risk-adapted friction	Escalate interaction demands for high-risk or conflicting advice.	Reduces automatic agreement where harm potential is high.
Post-decision audit	Review AI advice, clinician rationale, final action, and later clinical trajectory.	Supports local assurance, learning, and iterative improvement.

Discussion:

The main implication of this review is that interface design belongs inside clinical AI safety. It should not be deferred until after model development, treated as cosmetic implementation, or reduced to convenience. When AI recommendations influence diagnosis, treatment, triage, imaging interpretation, or disposition, the display of the recommendation becomes part of the intervention.

This conclusion changes the evaluation agenda. A clinical XAI study should report when AI advice was displayed, whether the clinician had already committed to a hypothesis, whether explanations were visible by default, whether the clinician had to respond, whether attention was measured, whether uncertainty

was displayed, whether alternatives were visible, and whether acceptance or rejection was documented. Without these variables, the interpretation of XAI effects remains unstable.

The review also clarifies why clinician resistance should not automatically be interpreted as lack of trust or lack of training. Sometimes resistance reflects poor interface fit. If a system interrupts at the wrong time, provides too much text, hides the discriminating fact, fails to respect expertise, or offers confidence without actionable uncertainty, distrust may be rational. Conversely, a system that is smooth, fast, and easy to accept may be unsafe if it reduces verification.

The proposed framework does not imply that every AI recommendation should be surrounded by complex friction. That would be unworkable. The argument is for adaptive friction. Clinical environments already ration attention. The interface should help ration it well. Low-risk suggestions should remain light. High-risk, uncertain, or conflicting recommendations should ask for more thought at the exact point where automatic agreement would be dangerous.

The framework also shifts responsibility away from individual vigilance alone. Telling clinicians to avoid overreliance is insufficient if the interface makes overreliance cheap. Design should make verification feasible. It should place relevant evidence near the recommendation, disclose missing data, expose alternatives, and make rejection legitimate. A clinician who disagrees with AI should not disappear from the system as a non-user; that disagreement may be the most important safety signal.

Finally, the article suggests a more mature definition of human-centered clinical AI. Human-centered design in this domain does not mean making AI pleasant, visually clean, or acceptable. It means preserving the clinician's capacity to reason under pressure, disagree with plausible automation, and remain accountable while using computational support. The goal is structured complementarity, not human-only decision-making and not obedience to AI.

Limitations:

This review is conceptual and design-oriented. It synthesizes DOI-indexed literature but does not perform meta-analysis, pooled estimation, or formal risk-of-bias grading across all included studies. The argument is strongest as a framework for study design, reporting, and system development rather than as evidence of outcome improvement in live clinical deployment.

The reviewed literature is heterogeneous. It spans radiology, intensive care simulation, neurology decision support, transplantation and neuro-oncology design work, general AI-assisted decision-making, and systematic reviews of clinician trust. Interface effects may differ across emergency medicine, intensive care, oncology, outpatient primary care, radiology, pathology, and medication ordering. The framework should therefore be adapted to specialty-specific workflows.

Clinical AI safety also depends on model performance, calibration, dataset representativeness, institutional governance, legal responsibility, user training, and local workflow. Interface design cannot compensate for an unsafe model, biased data, poor validation, or lack of accountability. The claim is narrower: even a technically strong model may become clinically unsafe if the interface promotes uncalibrated attention and reliance.

Future empirical work should test Attention-Gated XAI in controlled simulations with practicing clinicians and then in prospective clinical environments. Studies should randomize timing, pre-commitment, friction level, counterevidence prompts, and explanation type. Outcomes should include diagnostic accuracy, treatment safety, overreliance on incorrect AI advice, rejection of correct advice, attention to explanations, time burden, workload, and documented rationale quality.

Conclusion:

Clinical XAI should not be evaluated as explanation content alone. In practice, an explanation reaches the physician through an interface that controls timing, visibility, salience, effort, comparison, and documentation. That interface determines whether AI advice becomes a second opinion, an anchor, a shortcut, or a structured object of review.

The Attention-Gated XAI framework treats explanation as a decision event. It requires clinical pre-commitment, adaptive timing, risk-stratified friction, counterevidence prompting, missing-data display, acceptance and rejection logging, and post-decision audit. These components preserve clinician agency while reducing automatic agreement with AI advice.

The practical standard is direct: an explanation that is not attended to has not explained; an explanation that cannot be challenged has not protected the patient; and an interface that makes agreement easier than

verification has already shaped the decision before the clinician recognizes it. Safe clinical AI begins before the answer. It begins at the point where the interface decides how the answer is allowed to enter thought.

Acknowledgments:

No external editorial or institutional support was used in the preparation of this review manuscript.

Ethical Approval:

Ethical approval was not required because this manuscript is a review and conceptual framework based exclusively on previously published literature and did not involve human participants, patient records, animal subjects, or identifiable private data.

Conflict of Interest:

The author declares no conflict of interest.

Funding:

No specific funding was received for this work.

Data Availability:

No original dataset was generated. All sources discussed in the manuscript are available through their respective publishers and DOI records.

Use of Artificial Intelligence Tools:

Artificial intelligence tools were used for language drafting assistance and formatting support. The author reviewed, revised, verified the cited literature, and assumes full responsibility for the manuscript content.

References:

1. Bienefeld N, Boss JM, Lüthy R, Brodbeck D, Azzati J, Blaser M. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digital Medicine. 2023;6:94. DOI: 10.1038/s41746-023-00837-4.

2. Subramanian HV, Canfield C, Shank DB. Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review. Artificial Intelligence in Medicine. 2024;149:102780. DOI: 10.1016/j.artmed.2024.102780.

3. Prince EW, Mirsky DM, Hankinson TC, Görg C. Current state and promise of user-centered design to harness explainable AI in clinical decision-support systems for patients with CNS tumors. Frontiers in Radiology. 2025;4:1433457. DOI: 10.3389/fradi.2024.1433457.

4. Tun HM, Rahman HA, Naing L, Malik OA. Trust in Artificial Intelligence-Based Clinical Decision Support Systems Among Health Care Workers: Systematic Review. Journal of Medical Internet Research. 2025;27:e69678. DOI: 10.2196/69678.

5. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. NPJ Digital Medicine. 2024;7:202. DOI: 10.1038/s41746-024-01200-x.

6. Prinster D, Mahmood A, Saria S, Jeudy J, et al. Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI. Radiology. 2024;313(2):e233261. DOI: 10.1148/radiol.233261.

7. Gombolay GY, Silva A, Schrum M, Gopalan N, Hallman-Cooper J, Dutt M, Gombolay M. Effects of explainable artificial intelligence in neurology decision support. Annals of Clinical and Translational Neurology. 2024;11(5):1224-1235. DOI: 10.1002/acn3.52036.

8. Panigutti C, Beretta A, Giannotti F, Pedreschi D. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems. Proceedings of the CHI Conference on Human Factors in Computing Systems. 2022. DOI: 10.1145/3491102.3502104.

9. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. International Journal of Human-Computer Studies. 2023;169:102941. DOI: 10.1016/j.ijhcs.2022.102941.

10. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. JMIR AI. 2024;3:e53207. DOI: 10.2196/53207.

11. Rosenbacke R, Melhus Å, McKee M, Stuckler D. AI and XAI second opinion: the danger of false confirmation in human-AI collaboration. Journal of Medical Ethics. 2025;51(6):396-400. DOI: 10.1136/jme-2024-110074.

12. Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW1):1-21. DOI: 10.1145/3449287.

13. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction. 2023;7(CSCW1):1-38. DOI: 10.1145/3579605.

14. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld DS. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the CHI Conference on Human Factors in Computing Systems. 2021. DOI: 10.1145/3411764.3445717.

15. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;3(11):e745-e750. DOI: 10.1016/S2589-7500(21)00208-9.

16. Cabitza F, Natali C, Famiglini L, Campagner A, Caccavella V, Gallazzi E. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. Artificial Intelligence in Medicine. 2024;150:102819. DOI: 10.1016/j.artmed.2024.102819.

17. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. DOI: 10.1136/bmj-2023-078378.

18. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine. 2025;31(1):60-69. DOI: 10.1038/s41591-024-03425-5.

19. Shah NH, Halamka JD, Saria S, Pencina M, Tazbaz T, Tripathi M, et al. A Nationwide Network of Health AI Assurance Laboratories. JAMA. 2024;331(3):245-249. DOI: 10.1001/jama.2023.26930.

20. Goodman KE, Yi PH, Morgan DJ. AI-Generated Clinical Summaries Require More Than Accuracy. JAMA. 2024;331(8):637-638. DOI: 10.1001/jama.2024.0555.

21. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nature Medicine. 2020;26(8):1229-1234. DOI: 10.1038/s41591-020-0942-0.

22. Cabitza F, Campagner A, Ronzio L, Cameli M, Elena G, Concetta M, et al. Rams, hounds and white boxes: Investigating human-AI collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine. 2023;138:102506. DOI: 10.1016/j.artmed.2023.102506.

23. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics. 2021;113:103655. DOI: 10.1016/j.jbi.2020.103655.

24. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digital Medicine. 2020;3:17. DOI: 10.1038/s41746-020-0221-y.