Appropriate Rejection of Incorrect AI Advice in Clinical Decision Support: A Review and Safety Framework.

Abstract:

Background: Explainable artificial intelligence is frequently presented as a route to transparent and safer clinical decision support. In high-risk decisions, increased trust is not equivalent to safer use. When artificial intelligence advice is incorrect, incomplete, or clinically under-specified, a plausible explanation may convert error into an apparently acceptable recommendation.

Materials and Methods: This narrative review synthesised DOI-indexed literature on clinical explainable artificial intelligence, clinical decision support, trust calibration, advice-taking, unsafe recommendations, false confirmation, overreliance, cognitive forcing, human-AI collaboration, reporting standards, and health AI assurance. The analysis was organised around appropriate rejection: the clinician's capacity to reject incorrect AI advice while preserving appropriate reliance when AI advice is clinically valid.

Results: The evidence indicates that explanations have heterogeneous effects on clinician trust and performance. Explanation type, task difficulty, expertise, AI correctness, perceived coherence, attention, cognitive effort, workflow timing, and interface friction shape whether XAI improves judgement or amplifies inappropriate reliance. Studies on false confirmation, unsafe AI recommendations, chest radiograph diagnosis, neurology decision support, cognitive forcing, and AI assurance indicate that XAI should be judged by the quality of acceptance and rejection, not by trust increase alone.

Conclusion:A safety-oriented XAI framework should measure four decision states: correct AI accepted, correct AI rejected, incorrect AI accepted, and incorrect AI rejected. Clinical XAI should incorporate pre-commitment, uncertainty display, counterevidence prompts, differential diagnosis, structured disagreement, and rejection logging. The safest explanation is not the one that makes AI advice easier to accept; it is the one that makes necessary disagreement clinically visible.

Introduction:

Artificial intelligence has entered clinical decision support as a producer of recommendations, risk estimates, diagnostic rankings, triage suggestions, treatment prompts, documentation summaries, and workflow alerts. Its value is not restricted to computational accuracy. Once introduced into a clinical workflow, an AI system becomes part of the physician's reasoning environment. It changes what is noticed, which hypothesis gains weight, what is deferred, and which decision feels sufficiently justified.

Explainable artificial intelligence (XAI) was introduced to address opacity. The usual claim is that an explanation helps the clinician understand the recommendation and therefore trust it more appropriately. That claim is incomplete in medicine. A clinical explanation is not neutral text placed next to a prediction. It is an intervention on attention, confidence, responsibility, and action. The question is not whether the system explains. The question is what clinical behaviour the explanation produces.

The dominant evaluative language around XAI still emphasises transparency, trust, satisfaction, perceived usefulness, and adoption. These variables matter, but they do not exhaust safety. A high level of trust in a correct system is desirable. A high level of trust in an incorrect system is dangerous. The same subjective score may therefore describe two opposite states. Trust becomes clinically meaningful only when it is calibrated to the validity, uncertainty, and scope of the recommendation [1,2].

Clinical decision support systems have always carried a double edge. They may improve memory, guideline adherence, and consistency, yet they may also introduce alert fatigue, inappropriate reliance, and decision displacement [3]. AI intensifies this tension because outputs often arrive with statistical authority

and, increasingly, with fluent natural-language rationales. A plausible explanation may make an incorrect recommendation harder to reject, especially when it overlaps with the physician's initial hypothesis.

The safest clinical AI system is not the one that produces the highest average trust score. It is the one that helps clinicians accept correct advice and reject incorrect advice. This review names the latter capacity appropriate rejection. Appropriate rejection is the physician's ability to refuse incorrect, incomplete, unsafe, or under-specified AI advice with a clinically defensible reason, while preserving willingness to use the system when advice is sound.

This paper reviews DOI-indexed literature on clinical XAI, trust calibration, unsafe AI advice, false confirmation, human-AI collaboration, overreliance, cognitive forcing, reporting standards, and health AI assurance. It then proposes an Appropriate Rejection Framework for evaluating clinical XAI not by persuasion, but by conditional behaviour under correct and incorrect AI advice. The central thesis is direct: in clinical decision support, the explanation should not be designed to make the recommendation easier to follow; it should be designed to make agreement earned and disagreement visible.

Materials and Methods:

This manuscript is a narrative review with a conceptual safety framework. The review was designed to answer a specific question: how should clinician-facing XAI be evaluated when AI advice is wrong, incomplete, or unsafe for the current patient? The analysis therefore prioritised studies that measured or theorised trust, reliance, advice-taking, unsafe recommendations, diagnostic performance, attention, cognitive effort, interface friction, reporting, or human-AI team behaviour.

Relevant literature was identified from DOI-indexed publications available through PubMed, Nature Portfolio, JMIR, Radiology, ACM Digital Library, ScienceDirect, Wiley, BMJ, JAMA, and publisher-hosted article pages. Search concepts included explainable artificial intelligence, clinical decision support, trust calibration, overreliance, advice-taking, unsafe AI recommendations, false confirmation, medical decision-making, cognitive forcing, human-AI collaboration, clinical XAI, reporting standards, and health AI assurance.

Inclusion criteria were: DOI-indexed article; direct relevance to clinical AI, clinical decision support, XAI, clinician trust, reliance, human-AI decision-making, reporting, or governance; and contribution to at least one analytic category: trust calibration, incorrect advice, explanation type, attention, acceptance behaviour, rejection behaviour, cognitive forcing, reporting, or design of interaction. Articles without DOI, institutional web pages, and purely technical model papers without user-facing decision consequences were not cited as evidence.

The review used a mechanism-oriented synthesis rather than a statistical meta-analysis. The included studies were heterogeneous in clinical domain, task, participant expertise, explanation format, AI correctness, and outcome measure. Quantitative pooling would obscure the decision mechanisms of interest. The synthesis therefore grouped findings by behavioural function: trust formation, advice-taking, false confirmation, attention to explanations, overreliance, friction-based mitigation, reporting, and assurance.

The framework was constructed by mapping the reviewed outcomes into four reliance states: accepting correct AI advice, rejecting correct AI advice, accepting incorrect AI advice, and rejecting incorrect AI advice. This mapping exposes why overall trust, overall agreement, and overall usage are insufficient endpoints. Safety depends on whether clinician behaviour changes with the correctness and limits of the AI recommendation.

Table 1. Search Strategy and Synthesis Domains.

Domain	Search concepts	Role in synthesis
Clinical XAI and CDSS	clinical decision support; explainable AI; clinician trust; advice-taking	Defined the base relationship between explanation, clinician perception, and behaviour.
Unsafe advice and false confirmation	unsafe AI recommendations; false confirmation; incorrect advice; diagnostic error	Identified the condition in which explanation becomes a safety stress test.
Cognitive forcing and reliance	overreliance; cognitive forcing; human-AI team performance; reliance calibration	Mapped mitigation strategies for automatic acceptance of AI advice.
Reporting and assurance	TRIPOD+AI; TRIPOD-LLM; health AI assurance; clinical AI evaluation	Connected appropriate rejection to transparent reporting, evaluation, and post-deployment

Domain	Search concepts	Role in synthesis
		monitoring.

Conceptual Definitions:

Several terms require separation. Trust is a subjective expectation that the system will perform in a way that supports the user's goal. Reliance is behavioural dependence on the system's output. Acceptance is adoption of a recommendation into the clinical decision. Rejection is refusal, modification, or disregard of a recommendation. Appropriate reliance occurs when correct advice is accepted and incorrect advice is rejected. Underreliance occurs when correct advice is rejected. Overreliance occurs when incorrect advice is accepted [4,5].

Appropriate rejection is narrower than distrust. Distrust may reject a system globally. Appropriate rejection rejects a specific recommendation for a specific clinical reason. It may occur because required data are missing, the advice conflicts with patient-specific evidence, the recommendation is outside the model's validated scope, the explanation is physiologically weak, or a safer alternative remains insufficiently evaluated.

Clinical reasoning adds further complexity. A recommendation may be diagnostically correct but therapeutically unsafe; statistically likely but unacceptable given a contraindication; correct at population level but poor for the current patient; or plausible but incomplete because it omits a dangerous differential diagnosis. Appropriate rejection includes these conditions. It is not restricted to binary right-wrong classification.

Results: Thematic Synthesis

1. Trust is a weak endpoint unless it is calibrated

Empirical work in clinical XAI shows that explanations do not have a uniform effect on trust. Bussone and colleagues found that explanations in clinical decision support influence both trust and reliance, but also showed why higher trust cannot be interpreted as a universal benefit [4]. Panigutti and colleagues examined advice-taking in an AI-based clinical decision support system and made the behavioural dimension of reliance explicit [5].

Naiseh and colleagues compared explanation classes in a clinical decision support setting and reported differences in trust calibration across explanation types [6]. The finding is central to clinical safety. A local feature explanation, a counterfactual explanation, an example-based explanation, and a natural-language rationale do not merely express the same meaning in different formats. They alter how clinicians perceive causality, uncertainty, and authority.

Rosenbacke and colleagues' systematic review found that XAI may increase, decrease, or fail to alter clinician trust depending on clarity, coherence, complexity, and clinical relevance [7]. This undermines a simplistic design principle. The aim cannot be more explanation or higher trust. The aim must be the right level of trust in the right condition. A confusing explanation that lowers trust may prevent adoption of useful support. A coherent explanation that raises trust in wrong advice may create harm.

The false hope critique by Ghassemi and colleagues is therefore not an argument against explanation. It is a warning against treating explanation as a safety guarantee [1]. A post-hoc rationale may improve user comfort while failing to represent model logic, failing to capture causal reasoning, or failing to identify clinical conditions under which the output should be ignored. In clinical settings, comfort without calibration is not safety.

2. Incorrect advice is the real stress test for XAI

Incorrect advice is the real stress test for clinical XAI. A system evaluated only under correct recommendations is not fully evaluated as a safety technology. The clinical problem emerges when advice is

inaccurate, incomplete, overconfident, mis-specified for the patient, or correct in one dimension but unsafe in another.

Gaube and colleagues studied physicians given chest radiographs and diagnostic advice, some of which was inaccurate, while varying whether the advice appeared to come from AI or human sources [9]. Diagnostic accuracy worsened when participants received inaccurate advice, regardless of the stated source. This demonstrates that advice itself exerts a pull on clinical judgement. The label AI is not the only issue; the presence of advice alters the decision environment.

Nagendran and colleagues studied ICU physicians exposed to safe and unsafe AI recommendations, accompanied by four types of XAI, using eye tracking as an objective behavioural measure [10]. Unsafe recommendations drew more attention than safe recommendations, yet explanations did not rescue decision-makers in unsafe scenarios. Self-reported usefulness of explanations did not correlate with visual attention to those explanations. The study shows that subjective evaluation misses behaviour at the point of risk.

Prinster and colleagues reported that explanation type affected physicians' diagnostic performance and trust in chest radiograph diagnosis, even when physicians were not aware of these effects [11]. This finding is clinically significant. Explanations shape decision-making beneath explicit self-report. A clinician may not recognise how a particular explanation format has changed confidence, agreement, or willingness to override AI advice.

Gombolay and colleagues examined XAI techniques in neurology decision support and found that explanation formats differ in perceived explainability, trust, and interaction outcomes [12]. This supports a domain-sensitive approach. A visual or probability-based explanation suitable for one task may be poorly matched to another. Unsafe advice should therefore be tested across explanation formats, specialties, and user expertise levels.

Table 2. Four Reliance States in Clinical AI Advice.

AI advice status	Clinician action	Decision state	Safety interpretation
Correct advice	Accepted	Appropriate reliance	Desired state: the system improves or supports clinical judgement.
Correct advice	Rejected	Underreliance	Potential loss of benefit, delayed diagnosis, unnecessary resource use, or missed risk.
Incorrect advice	Accepted	Overreliance or false confirmation	High-risk state: AI becomes part of the error chain.
Incorrect advice	Rejected	Appropriate rejection	Desired safety state: clinical judgement interrupts unsafe automation.

3. False confirmation is a clinical AI failure mode

False confirmation is the distinctive hazard of AI as a second opinion. It occurs when AI erroneously validates a human decision that is already wrong or incomplete. The system does not need to replace the physician. It only needs to confirm the physician's initial path strongly enough that reconsideration becomes less likely [8].

This failure mode is plausible in clinical cognition. Physicians often form an early working diagnosis, then integrate additional data. This process is necessary, but it is vulnerable to anchoring and confirmation bias. If AI enters after the initial hypothesis and agrees with it, the output may be experienced as independent validation. The physician may become less likely to reopen the differential diagnosis or ask what finding contradicts the shared conclusion.

False confirmation is especially dangerous when the recommendation is not absurd. A poor AI answer that names irrelevant diseases is easy to dismiss. A poor answer that correctly recognises some symptoms, uses appropriate vocabulary, and proposes a plausible diagnosis is harder to resist. The explanation contains true fragments organised around a false conclusion. Clinical reasoning often accepts pattern coherence before it verifies every premise.

Rosenbacke and colleagues argue that false confirmation may be pervasive and safety-relevant in physician-AI collaboration [8]. The practical conclusion is direct: AI second opinions should be integrated after human pre-commitment and should include prompts that make disagreement possible. A second opinion has clinical value when it challenges reasoning, not when it merely confirms the first available path.

4. Explanations, agreement, and human-AI team performance

Human-AI collaboration studies extend the caution. Bansal and colleagues showed that explanations do not automatically improve complementary team performance. Their work found that explanations can increase the chance that humans accept AI recommendations regardless of correctness [13]. This is a central lesson for clinical decision support. Agreement is not the endpoint. Selective agreement is.

Alufaisan and colleagues questioned whether XAI reliably improves human decision-making, finding that AI prediction itself may dominate outcomes more than explanation content [20]. Chen and colleagues showed that human intuition shapes reliance on AI explanations, helping explain why users sometimes override correct advice and sometimes follow incorrect advice [21]. These findings illuminate a mechanism that clinical systems share: users do not process explanations as pure evidence. They interpret them through prior beliefs, task intuition, and perceived model limits.

Vasconcelos and colleagues provide a cost-benefit account of overreliance: users choose whether engaging with the explanation is worth the cognitive effort [14]. In clinical work, this mechanism is amplified. Physicians manage interruptions, fatigue, time pressure, electronic documentation, uncertain data, and patient flow. If the effort of checking the AI exceeds the perceived benefit, the user may accept advice as a shortcut. If the explanation is too dense, too late, or too detached from the clinical question, it becomes decorative.

Buçinca and colleagues demonstrated that cognitive forcing functions reduce overreliance compared with simple explanation approaches, although they may lower subjective ratings [15]. This trade-off is clinically acceptable when risk is high. A system that feels less convenient but prevents acceptance of dangerous advice may be safer than a frictionless system that clinicians like but follow too readily. The design problem is not whether friction should exist; it is where and how much friction is justified.

5. Reporting standards and assurance make rejection measurable

Appropriate rejection also requires better reporting. TRIPOD+AI updated reporting guidance for prediction models that use regression or machine learning methods and emphasised transparent reporting of design, data, modelling, performance, and evaluation [26]. Although TRIPOD+AI is not a clinical XAI framework, it reinforces a central point for this review: safety claims require visible evidence about context, population, inputs, model behaviour, and evaluation conditions.

TRIPOD-LLM extends this reporting logic to studies using large language models [27]. This matters because LLM-mediated clinical explanations may appear as fluent narratives. If studies do not report prompt structure, data context, output evaluation, clinician role, and failure modes, readers cannot distinguish a useful explanation from persuasive language. For appropriate rejection, reporting should state whether advice correctness varied and whether rejection behaviour was measured.

Health AI assurance literature adds the institutional layer. Shah and colleagues argued for a national network of health AI assurance laboratories to support evaluation and monitoring of health AI systems [28]. This logic is relevant beyond model performance. Hospitals and vendors should monitor not only accuracy and adoption, but also patterns of acceptance, override, disagreement, and later-discovered error. Goodman and colleagues' argument that AI-generated clinical summaries require more than accuracy reinforces the same principle: clinical usefulness and safety depend on how outputs enter workflow and decision-making, not only on isolated correctness [29].

The Appropriate Rejection Framework:

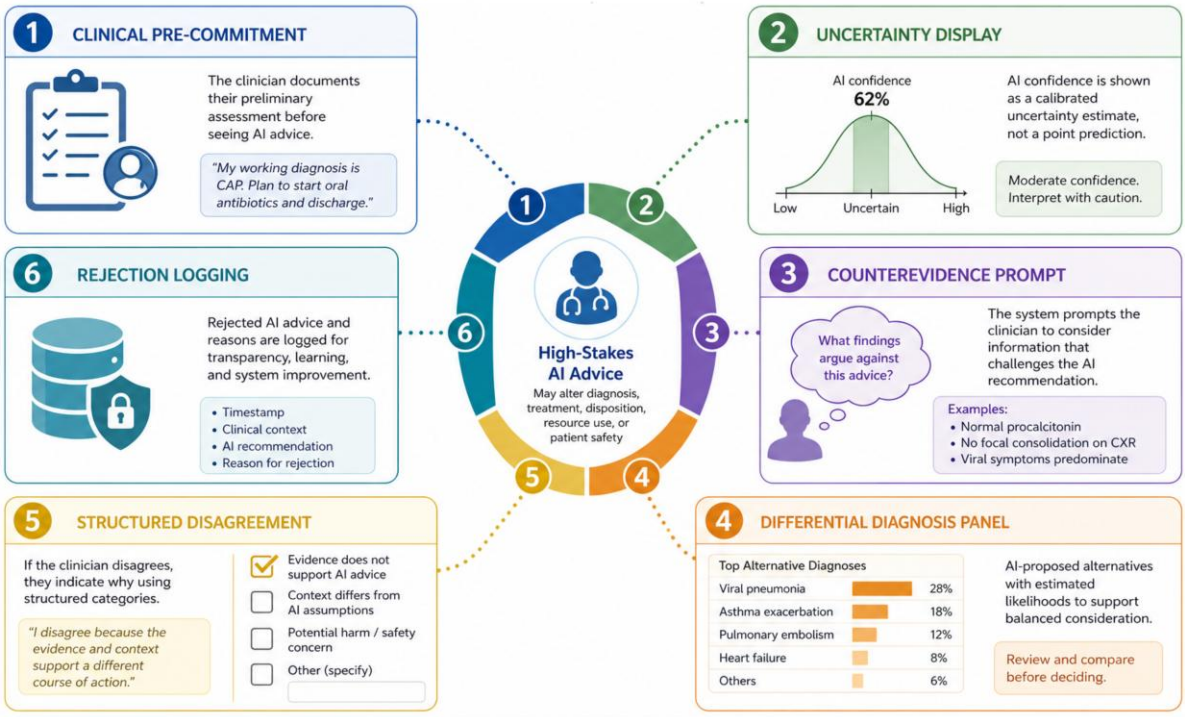
The Appropriate Rejection Framework translates these findings into design and evaluation criteria. Its purpose is to prevent an explanation from functioning as mere persuasion. It structures the clinical interaction so that AI advice is accepted, modified, or rejected through visible reasoning.

The framework has six components: clinical pre-commitment, uncertainty display, counterevidence prompt, differential diagnosis panel, structured disagreement, and rejection logging. These components are not intended for every low-risk suggestion. They are intended for decisions where AI advice could change diagnosis, treatment, disposition, resource use, or patient safety.

Table 3. Components of the Appropriate Rejection Framework.

Component	Operational definition	Safety function
Clinical pre-commitment	The clinician records a preliminary diagnosis, action, or confidence level before AI advice appears.	Reduces anchoring and preserves a human judgement for comparison.
Uncertainty display	The system states confidence, missing data, model boundary conditions, and patient-specific limits.	Prevents the recommendation from appearing stronger than the evidence permits.
Counterevidence prompt	The interface asks which finding would weaken or reverse the AI recommendation.	Activates disconfirmatory reasoning and reduces false confirmation.
Differential diagnosis panel	The system displays close alternatives and discriminating clinical data.	Keeps the problem open when AI advice is underdetermined.
Structured disagreement	The clinician selects a concise reason for accepting, modifying, or rejecting high-risk advice.	Turns disagreement into a traceable clinical act.
Rejection log	The system stores the rejection reason, evidence cited, and final human decision.	Creates audit data for governance, monitoring, and clinical learning.

Image 1. Appropriate Rejection Framework for Safe Clinical XAI.



Clinical pre-commitment before AI advice

Pre-commitment requires the clinician to record a working hypothesis, planned action, or confidence estimate before viewing AI advice. The design principle is temporal. If AI appears first, it anchors. If AI appears after a documented human judgement, it can serve as comparison. Cabitza and colleagues' work on collaboration protocols supports this view: sequence and protocol alter how humans and AI interact in medical diagnosis [16].

Pre-commitment also improves auditability. If the physician changes a decision after AI advice, the system can record whether the change was justified by new evidence, uncertainty reduction, or inappropriate deference. Without pre-commitment, change in reasoning is difficult to reconstruct. The final order remains visible, but the cognitive path disappears.

Uncertainty as part of the explanation

Uncertainty display is the second component. Clinical AI should not present a recommendation as if it were detached from missing data, model scope, or patient-specific constraints. A confidence score alone is not enough. A useful uncertainty display should state what evidence supports the recommendation; what

evidence is missing; what patient factors may invalidate it; what alternatives remain plausible; and what test or clinical observation would change the conclusion.

This is where clinical explanation differs from generic interpretability. A clinician does not need a long description of every model feature. The clinician needs a reasoned account of why this recommendation is safe enough for this patient now. An explanation that cannot name its own boundary conditions is incomplete.

Counterevidence and differential diagnosis

The counterevidence prompt is the most direct anti-confirmation element. It asks the clinician to identify what would weaken the AI recommendation. In suspected pulmonary embolism, this may include low pretest probability, alternative diagnosis, or contraindication to imaging. In sepsis, it may include absence of organ dysfunction, noninfectious mimic, or lack of source. The prompt forces the decision-maker to look away from the supported conclusion and toward its conditions of failure.

The differential diagnosis panel complements this step. It should not present a long generic list. It should present the nearest dangerous alternatives and the discriminating data. For a chest radiograph system, the panel might distinguish pneumonia, pulmonary edema, atelectasis, and malignancy. For neurological decision support, it might distinguish stroke mimic, seizure, migraine, demyelination, and metabolic disorder. The point is not encyclopaedia. The point is structured contrast.

Structured disagreement and rejection logging

Structured disagreement transforms rejection into a clinical act. In many decision systems, ignoring or overriding AI advice leaves little trace. The system registers the final order but not why the AI was rejected. That design treats rejection as absence. In safety terms, rejection is data.

A structured disagreement interface should use concise options: contradictory clinical evidence, missing required data, patient-specific contraindication, alternative diagnosis stronger, recommendation outside validated scope, low clinical plausibility, guideline exception, or patient preference. A free-text field should be available for nuance, but the primary taxonomy should be simple enough for use during clinical workflow.

Rejection logging supports governance. Repeated rejections for the same reason may reveal model drift, population mismatch, poor explanation design, or flawed deployment. Repeated acceptance of later-proven incorrect advice may reveal over-persuasive interface design. In both cases, the human response is not noise; it is part of monitoring.

Measurement Model:

Appropriate rejection requires new metrics. Standard accuracy of the AI model remains necessary, but it is not sufficient. The system must be evaluated as a coupled human-AI decision process. The outcome is not only whether the model predicted correctly; it is whether the clinician responded correctly to the model's correctness and uncertainty.

The primary behavioural metrics should include acceptance rate for correct advice, rejection rate for correct advice, acceptance rate for incorrect advice, rejection rate for incorrect advice, time to override, request for additional evidence, change in differential diagnosis, and final clinical action. Cognitive metrics should include visual attention to advice and explanation, sequence of interaction, number of explanation elements opened, and time spent on counterevidence. Documentation metrics should include stated reason for acceptance, modification, or rejection.

These metrics should be stratified by expertise, task difficulty, time pressure, and risk. The same interface may reduce overreliance in an attending physician and increase cognitive burden in a resident. The same explanation may be helpful in a slow diagnostic task and intrusive in emergency triage. Appropriate rejection is therefore not a single score. It is a safety profile.

Table 4. Suggested Metrics for Evaluating Appropriate Rejection.

Metric domain	Examples	Interpretive value
Behavioural	Acceptance of correct advice; rejection of incorrect advice;	Shows whether clinician behaviour tracks AI

Metric domain	Examples	Interpretive value
	time to override; request for additional evidence; final clinical action.	correctness and patient-specific safety.
Cognitive-attentional	Fixation on recommendation; fixation on explanation; sequence of clicks; time spent on counterevidence; differential diagnosis review.	Shows whether the explanation was processed rather than merely displayed.
Documentary	Reason for acceptance; reason for modification; reason for rejection; evidence cited; uncertainty recorded.	Shows whether agreement or disagreement is clinically auditable.
Governance	Clusters of repeated rejection; repeated acceptance of later-identified incorrect advice; mismatch by specialty or expertise.	Identifies model drift, poor deployment, and interface-induced overreliance.

Clinical Scenarios:

Consider suspected sepsis in an older patient with pneumonia. An AI system recommends pneumonia alone and omits sepsis because lactate is unavailable. A persuasive explanation may list cough, fever, infiltrate, and antibiotic selection. A safe explanation would state that organ dysfunction cannot be excluded, that lactate is missing, and that arterial blood gas or serum lactate should be obtained before ruling out sepsis. If the physician rejects the AI's narrower diagnosis, that rejection is appropriate and should be recorded as such.

Consider suspected pulmonary embolism after hip surgery. An AI system recommends chest radiography, electrocardiography, and laboratory tests but omits CT pulmonary angiography in a high-probability case. A clinician who rejects the incomplete workup and orders definitive imaging is not resisting AI. The clinician is correcting under-specification. The explanation should make this path easier by showing guideline-sensitive boundary conditions rather than presenting the initial workup as sufficient.

Consider chest radiograph interpretation. If AI favours pneumonia with high confidence and the clinician initially agrees, false confirmation may occur. A counterevidence prompt would ask whether volume overload, atelectasis, malignancy, aspiration, or pulmonary embolic disease remains plausible. The physician is not forced to reject AI. The physician is forced to preserve differential reasoning before agreement.

Implementation Pathway for Clinical Systems:

A clinical system that adopts appropriate rejection should not present every AI output with the same interaction burden. The first implementation step is risk tiering. Low-risk outputs, such as administrative coding suggestions or non-urgent reminders, may receive concise optional explanations. Intermediate-risk outputs, such as diagnostic ranking or laboratory follow-up suggestions, should show uncertainty, missing data, and alternatives. High-risk outputs, such as discharge advice, anticoagulation, thrombolysis, antibiotic delay, ICU triage, or invasive testing, should require pre-commitment and structured acceptance or rejection.

Risk tiering prevents the framework from becoming a documentation tax. Excessive friction across all decisions would create alert fatigue and workarounds. Selective friction preserves usability while protecting decisions where error has material consequence. This design principle is consistent with cognitive-forcing evidence: forcing functions are most defensible when the cost of unexamined agreement is high [15].

The second implementation step is integration into the electronic health record or clinical decision support interface. The AI recommendation should not sit in an isolated panel detached from patient data. The clinician should see the recommendation beside the evidence used, evidence not used, missing variables, relevant contraindications, and alternatives. A separate explanation window that requires several clicks may satisfy formal transparency while failing practical cognition.

The third implementation step is human role specification. A student, resident, consultant, nurse practitioner, intensivist, radiologist, and emergency physician do not need the same explanation granularity. Stakeholder-driven design literature in medical XAI has emphasised that explainability requirements vary by user, task, and context [22,23]. Less experienced clinicians may need more explicit differential diagnosis. Experts may need concise uncertainty and boundary conditions. Team-based decisions may require shared visibility of both AI advice and human override reason.

The fourth implementation step is audit. Rejection data should not be reviewed as user disobedience. It should be analysed as a signal. A cluster of rejections due to missing renal function, for example, reveals a data-readiness problem. A cluster due to contraindication reveals a safety-rule gap. A cluster due to low plausibility may reveal model drift or poor local calibration. Conversely, a cluster of accepted recommendations later judged unsafe reveals over-persuasive design or insufficient uncertainty display.

Evaluation Protocol for Future Studies:

A rigorous study of appropriate rejection should use a factorial design. The first factor is AI correctness: correct advice, incorrect advice, and partially correct advice. The second factor is explanation design: no explanation, conventional supportive explanation, uncertainty-aware explanation, and counterevidence-prompted explanation. The third factor is timing: AI before human judgement versus AI after human pre-commitment. The fourth factor is task risk: low, intermediate, and high clinical consequence.

Participants should include multiple expertise levels. Medical students and residents are useful for controlled comparison, but they cannot substitute for practicing clinicians in high-stakes workflow evaluation. Senior clinicians often rely on tacit pattern recognition and may respond differently to explanation formats. Studies should therefore stratify by training stage, specialty, and prior exposure to clinical AI [9,12,23].

Primary endpoints should be behavioural: acceptance of correct advice, rejection of correct advice, acceptance of incorrect advice, and rejection of incorrect advice. Secondary endpoints should include diagnostic accuracy, treatment safety, time, number of additional tests ordered, confidence change, perceived usefulness, cognitive load, and documentation quality. Eye tracking and clickstream data should be used where feasible because self-report alone misses relevant attention behaviour [10].

The most important experimental condition is the plausible wrong recommendation. Obvious errors underestimate risk. A recommendation that is wrong but clinically absurd is easy to reject. A recommendation that is wrong yet coherent, fluent, and partially grounded is the true test. Such cases should be built from common clinical traps: pneumonia without sepsis, pulmonary embolism missed after surgery, stroke mimic treated as stroke, renal-dose omission, discharge despite red flags, and imaging interpretation with a plausible distractor.

Follow-up analysis should distinguish error prevention from decision delay. A system that reduces overreliance while doubling time for every low-risk case is operationally weak. A system that adds thirty seconds to a high-risk recommendation and prevents acceptance of unsafe advice is clinically defensible. Evaluation should therefore report time cost by risk tier, not as a single mean.

Table 5. Research Agenda for Validating Appropriate Rejection.

Research question	Suggested design	Primary endpoint
Does pre-commitment reduce false confirmation?	Randomised simulation with AI shown before or after human diagnosis.	Incorrect AI advice rejected with documented reason.
Does counterevidence prompting improve diagnostic safety?	Crossover clinical vignette study with and without counterevidence prompt.	Change in acceptance of plausible incorrect advice.
How does expertise modify appropriate rejection?	Stratified study involving students, residents, specialists, and consultants.	Interaction between expertise and rejection of unsafe advice.
Does rejection logging improve governance?	Silent deployment audit comparing free override with structured reason logging.	Identification of recurring model or interface failure modes.

Large Language Model Explanations and the Fluency Problem:

Large language models intensify the appropriate rejection problem because they produce explanations in the same medium clinicians use to reason: language. A table, heatmap, or probability score visibly belongs to a machine interface. A fluent paragraph resembles a colleague's explanation. That resemblance may help communication, but it also increases the risk that verbal coherence is mistaken for clinical validity.

The issue is not that natural-language explanations should be excluded. In many clinical settings, textual explanations are the most useful format because medicine is documented, handed off, and justified in

language. The issue is that fluency should be decoupled from authority. A well-written explanation should still display uncertainty, missing data, contraindications, alternative diagnoses, and reversal criteria. Otherwise the explanation becomes rhetorically strong and clinically under-disciplined.

For LLM-mediated decision support, appropriate rejection requires explicit source separation. The system should distinguish retrieved patient facts, guideline-derived claims, model inference, and generated rationale. If these layers are fused into one fluent narrative, the clinician cannot tell whether a statement came from the chart, a validated rule, a probabilistic model, or language generation. The rejection act becomes harder because the target of rejection is blurred.

A safety-oriented LLM interface should therefore include a traceable reasoning scaffold: patient data used, patient data missing, guideline or source basis, model recommendation, uncertainty, alternatives, and clinician decision. The purpose is not to expose hidden chain-of-thought. The purpose is to expose clinically relevant premises and limits so that the physician can accept, modify, or reject the advice with a documented reason. TRIPOD-LLM reinforces the need for structured reporting when LLMs are used in prediction or clinical evaluation studies [27].

Discussion:

The main implication is methodological. Clinical XAI studies should include incorrect advice by design. Unsafe advice should not be an incidental error; it should be an experimental condition. Without that condition, investigators cannot distinguish helpful trust from dangerous reliance. The evaluation should report four separate outcomes: correct advice accepted, correct advice rejected, incorrect advice accepted, and incorrect advice rejected.

The second implication is design-related. Explanation content cannot be separated from interface sequence. Advice shown before human pre-commitment has a different cognitive effect from advice shown after a working diagnosis. Advice that includes uncertainty has a different effect from advice that presents only a conclusion. Advice that requires counterevidence has a different effect from advice that merely offers a paragraph. XAI is therefore a workflow intervention, not simply a model property.

The third implication is clinical governance. Rejection should be auditable. Hospitals and vendors often measure adoption, click-through, alert dismissal, and use frequency. These metrics are insufficient. A high dismissal rate may mean alert fatigue, but it may also mean clinicians are correctly rejecting weak advice. A high acceptance rate may mean usefulness, but it may also mean overreliance. Governance should inspect the reason for both behaviours.

The fourth implication is medico-legal. If an AI-influenced decision harms a patient, the relevant question will not be only whether the model was generally accurate. It will be whether the recommendation was appropriate for the patient, whether its limitations were visible, whether the clinician had a meaningful opportunity to disagree, and whether the disagreement or acceptance was documented. Structured acceptance and rejection logs create a record of clinical reasoning.

The fifth implication concerns training. Physicians should be trained not only to use AI but to interrogate it. Training should include examples of correct advice, incorrect advice, incomplete advice, overconfident advice, and advice that is plausible but unsafe. The aim is to cultivate conditional reliance. The physician should know when AI deserves attention, when it deserves acceptance, when it deserves modification, and when it deserves rejection.

The proposed framework also clarifies why a universal XAI standard is unlikely. A dermatology classifier, chest radiograph tool, ICU prescription assistant, neurology decision support system, and large language model triage agent have different failure modes. Appropriate rejection must be specified by domain. In every domain, however, the core requirement remains: the clinician must be able to see why the recommendation may fail.

Limitations:

This review is conceptual and mechanism-oriented. It does not estimate pooled effect sizes. The reviewed studies vary substantially in design, clinical domain, user population, explanation type, task stakes, and outcome measure. The framework should therefore be treated as a testable proposal rather than a validated instrument.

Much of the evidence comes from simulated settings. Simulation is necessary because intentionally exposing real patients to unsafe AI advice would be unethical. Yet simulation reduces consequence, responsibility, fatigue, institutional pressure, and medico-legal stakes. Future work should progress from controlled simulation to prospective silent trials, then staged clinical deployment with monitoring.

A further limitation is that the framework emphasises clinician-facing decision support. Patient-facing AI, administrative AI, and population-level prediction may require additional safeguards. The construct of appropriate rejection remains relevant, but the actor who rejects advice and the record of rejection may differ.

Finally, this review cites DOI-indexed literature to strengthen traceability. Some recent technical reports, policy statements, and implementation guides without DOI may be relevant to local deployment but were not used as cited evidence.

Conclusion:

Clinical XAI should not be judged by whether it increases trust. It should be judged by whether it calibrates action. The physician must be able to accept correct AI advice, reject incorrect AI advice, and document the reason for both. That conditional behaviour is the core of safe human-AI collaboration.

The reviewed literature shows that explanations affect clinician trust, attention, diagnostic performance, and advice-taking in non-uniform ways. Unsafe AI recommendations, false confirmation, explanation type, task difficulty, cognitive effort, and professional expertise shape whether AI support improves judgement or amplifies error. A safety-oriented framework must therefore treat rejection of incorrect advice as a positive outcome.

The Appropriate Rejection Framework proposes six design elements: pre-commitment, uncertainty display, counterevidence prompt, differential diagnosis panel, structured disagreement, and rejection logging. These elements shift clinical XAI from persuasion toward accountable reasoning. The goal is not a physician who follows AI more often. The goal is a physician whose agreement is earned and whose disagreement remains visible when the machine is wrong.

Acknowledgments:

None.

Ethical Approval:

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

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 or analysed. All cited material is publicly available through the referenced publications and DOI records.

Use of Artificial Intelligence Tools:

Artificial intelligence tools were used for language drafting assistance and formatting support. The author reviewed, corrected, verified, and assumes full responsibility for the scientific content, argument, references, and final manuscript.

References:

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

2. 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. J Biomed Inform. 2021;113:103655. doi:10.1016/j.jbi.2020.103655.

3. 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 Digit Med. 2020;3:17. doi:10.1038/s41746-020-0221-y.

4. Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. 2015 IEEE International Conference on Healthcare Informatics. 2015:160-169. doi:10.1109/ICHI.2015.26.

5. 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 2022 CHI Conference on Human Factors in Computing Systems. 2022. doi:10.1145/3491102.3502104.

6. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. Int J Hum Comput Stud. 2023;169:102941. doi:10.1016/j.ijhcs.2022.102941.

7. Rosenbacke R, Melhus A, 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.

8. Rosenbacke R, Melhus A, McKee M, Stuckler D. AI and XAI second opinion: the danger of false confirmation in human-AI collaboration. J Med Ethics. 2025;51(6):396-399. doi:10.1136/jme-2024-110074.

9. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Guttag JV, Colak E, Ghassemi M. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 2021;4:31. doi:10.1038/s41746-021-00385-9.

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

11. Prinster D, Bhatt A, Yang T, 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.

12. Gombolay GY, Silva A, Schrum M, Gopalan N, Hallman-Cooper J, Dutt M, Gombolay M. Effects of explainable artificial intelligence in neurology decision support. Ann Clin TranslNeurol. 2024;11(5):1224-1235. doi:10.1002/acn3.52036.

13. 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 2021 CHI Conference on Human Factors in Computing Systems. 2021. doi:10.1145/3411764.3445717.

14. Vasconcelos H, Jorke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations can reduce overreliance on AI systems during decision-making. Proc ACM Hum Comput Interact. 2023;7(CSCW1):1-38. doi:10.1145/3579605.

15. Bucinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum Comput Interact. 2021;5(CSCW1):1-21. doi:10.1145/3449287.

16. Cabitza F, Campagner A, Ronzio L, et al. Rams, hounds and white boxes: investigating human-AI collaboration protocols in medical diagnosis. Artif Intell Med. 2023;138:102506. doi:10.1016/j.artmed.2023.102506.

17. 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. Artif Intell Med. 2024;150:102819. doi:10.1016/j.artmed.2024.102819.

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

19. Jacovi A, Marasovic A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021:624-635. doi:10.1145/3442188.3445923.

20. Alufaisan Y, Marusich LR, Bakdash JZ, Zhou Y, Kantarcioglu M. Does explainable artificial intelligence improve human decision-making? Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35(8):6618-6626. doi:10.1609/aaai.v35i8.16819.

21. Chen V, Liao QV, Vaughan JW, Bansal G. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. Proc ACM Hum Comput Interact. 2023;7(CSCW2):1-32. doi:10.1145/3610219.

22. Bienefeld N, Boss JM, Luthy R, Brodbeck D, Azzati J, Blaser M. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digit Med. 2023;6:94. doi:10.1038/s41746-023-00837-4.

23. Subramanian HV, Canfield C, Shank DB. Designing explainable AI to improve human-AI team performance: a medical stakeholder-driven scoping review. Artif Intell Med. 2024;149:102780. doi:10.1016/j.artmed.2024.102780.

24. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. J Med Internet Res. 2025;27:e69678. doi:10.2196/69678.

25. Kostopoulos G, Karlos S, Kotsiantis S. Explainable artificial intelligence-based decision support systems: a survey. Electronics. 2024;13(14):2842. doi:10.3390/electronics13142842.

26. 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.

27. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5.

28. 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.

29. 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.