Scientific Research Publishing

Health, 2026, 18(7), 672-699
https://www.scirp.org/journal/health
ISSN Online: 1949-5005
ISSN Print: 1949-4998

Auditing Artificial Intelligence Reasoning in Healthcare: The DEEP SEE™-MIRROR Dual-Layer Architecture for Structured AI Analysis and Meta-Cognitive Safety Oversight

Adel Omar Bataweel

Alhammadi Hospitals Group, Riyadh, Saudi Arabia
Email: adelbataweel@hotmail.com, adel.taweel@alhammadi.med.sa

How to cite this paper: Bataweel, A.O. (2026) Auditing Artificial Intelligence Reasoning in Healthcare: The DEEP SEE™-MIRROR Dual-Layer Architecture for Structured AI Analysis and Meta-Cognitive Safety Oversight. *Health*, **18**, 672-699.
https://doi.org/10.4236/health.2026.187041

Received: March 12, 2026
Accepted: July 10, 2026
Published: July 13, 2026

Abstract

Artificial intelligence is increasingly integrated into clinical decision support, predictive monitoring, and diagnostic interpretation across modern healthcare systems. While these technologies offer substantial analytical capability, their expanding use raises important concerns regarding the transparency, reliability, and accountability of how AI systems derive clinical conclusions. Existing research on trustworthy artificial intelligence has largely focused on model performance, data quality, explainability, and governance. However, comparatively less attention has been directed toward systematic evaluation of the reasoning pathways through which AI systems generate those conclusions. This paper proposes a dual-layer architecture for structuring and auditing artificial intelligence reasoning in healthcare. In this framework, AI reasoning is defined as the sequence of inferential steps through which an AI system transforms clinical data into analytical conclusions, while the reasoning pathway refers to the traceable chain of intermediate inferences, evidence use, and decision logic supporting those conclusions. The first layer adapts the DEEP SEE™ framework into a structured reasoning protocol that guides AI systems through seven stages—Describe, Expose, Examine, Probe, Scan, Explore, and Elevate—to support systematic signal detection, evaluation of contributing factors, identification of hidden assumptions, and generation of interpretable clinical insights. The second layer introduces the MIRROR framework as a meta-cognitive audit protocol that evaluates reasoning integrity, defined as the extent to which a reasoning pathway is evidence-based, logically coherent, considers alternative explanations, and appropriately reflects uncertainty. Together, the DEEP SEE™-MIRROR architecture provides a structured approach in which AI systems perform structured clinical analysis followed by

systematic reasoning audit. The framework is particularly applicable to AI systems that provide auditable intermediate outputs or reasoning traces, including large language model-based and hybrid clinical decision-support systems, while partial application may be feasible in conventional predictive models with limited transparency. An illustrative proof-of-concept demonstrates how the architecture can transform AI-generated outputs into structured and auditable reasoning pathways, while proposed evaluation dimensions outline how reasoning integrity may be assessed in practice. By integrating principles from patient safety science, cognitive psychology, and trustworthy AI, the DEEP SEE™-MIRROR architecture offers a practical approach for improving transparency, reliability, and bias awareness in AI-assisted clinical decision systems. More broadly, it highlights the importance of evaluating not only what artificial intelligence systems predict, but how they reason.

Keywords

Artificial Intelligence, Patient Safety, AI Reasoning, Explainable AI, Clinical Decision Support, DEEP SEE, MIRROR, AI Governance

1. Introduction

Artificial intelligence (AI) is rapidly reshaping modern healthcare. Advances in machine learning, deep neural networks, and large language models have enabled AI systems to analyse large volumes of clinical data and support tasks traditionally performed by clinicians, including diagnostic interpretation, predictive risk modelling, medical imaging analysis, and clinical decision support [1]-[3]. In several narrowly defined domains, AI systems now demonstrate performance comparable to, and in some cases exceeding, that of human experts.

As a result, AI technologies are increasingly embedded within healthcare environments. Predictive algorithms are used to identify early signs of deterioration such as sepsis, respiratory failure, and cardiovascular instability, while AI-assisted systems analyse laboratory results, imaging, and electronic health records to support diagnostic and therapeutic decisions. This expansion has created new opportunities to improve speed, consistency, and early detection in clinical care.

At the same time, growing reliance on AI raises important concerns regarding transparency, reliability, and accountability. Many AI systems operate through complex statistical architectures whose internal reasoning processes remain difficult to interpret. Clinicians may therefore receive recommendations or alerts without fully understanding how those conclusions were reached. In safety-critical environments such as healthcare, this lack of interpretability creates challenges for trust, oversight, and responsible deployment.

Recent research has highlighted additional limitations, particularly in generative AI and large language models. These systems may produce outputs that appear coherent and authoritative yet are not adequately grounded in verifiable ev-

idence. This phenomenon, commonly referred to as hallucination, involves the generation of plausible but unsupported information [4] [5]. In clinical contexts, such outputs may undermine trust and introduce safety risks when AI-generated recommendations influence decision-making.

In response, researchers and policymakers have increasingly emphasized the importance of trustworthy artificial intelligence. Core principles associated with trustworthy AI include transparency, accountability, fairness, reliability, and human oversight [6] [7]. Research areas such as explainable artificial intelligence (XAI) and AI alignment aim to improve interpretability and ensure that AI systems operate in ways consistent with human values and safety expectations [8]-[10].

However, much of the existing literature remains focused on outputs rather than reasoning processes. Explainable AI methods often identify influential variables or provide post-hoc explanations of model behaviour, but they do not necessarily establish whether the reasoning pathway itself is coherent, evidence-based, or free from hidden assumptions. In other words, many approaches explain what a model did without systematically examining how it reasoned. While current approaches to trustworthy AI emphasize prediction accuracy, explainability, and governance, comparatively little attention has been given to systematically auditing the reasoning processes through which AI systems generate clinical conclusions.

Explainable artificial intelligence methods have made important contributions by providing insight into model outputs and identifying influential features. However, these approaches primarily focus on explaining what a model has produced rather than evaluating whether the reasoning pathway itself is valid, complete, and appropriately supported. In this sense, explainability does not necessarily guarantee reasoning integrity. The present study extends beyond explainability by introducing a framework that not only structures AI reasoning but also systematically audits the reasoning process itself, thereby complementing existing explainability, monitoring, and governance approaches.

In this context, it is important to clarify the key constructs used in this study. Artificial intelligence reasoning is defined as the sequence of inferential steps through which an AI system transforms clinical data into analytical conclusions. The reasoning pathway refers to the traceable chain of intermediate inferences, evidence use, and decision logic supporting a given conclusion. Reasoning integrity is defined as the extent to which this pathway is evidence-based, logically coherent, considers plausible alternative explanations, and appropriately reflects uncertainty. A meta-cognitive audit refers to the systematic evaluation of how a conclusion was reached, focusing on the quality, completeness, and reliability of the reasoning process rather than the outcome alone.

The applicability of reasoning audit varies across different types of AI systems. Conventional predictive models typically generate probabilistic outputs with limited access to intermediate reasoning processes, making full reasoning audit chal-

lenging. In contrast, large language model-based systems and hybrid clinical decision-support systems often generate interpretable reasoning traces, explanations, or intermediate outputs that can be examined and evaluated. The DEEP SEE™-MIRROR architecture is therefore most directly applicable to AI systems that expose reasoning pathways or supporting artifacts, while partial application may be feasible in more opaque predictive systems where only outputs and selected explanatory features are available.

Insights from cognitive psychology suggest that this distinction is important. Human decision failures often arise not simply from incorrect conclusions but from flawed reasoning processes shaped by uncertainty, cognitive bias, and contextual pressures [11]. In clinical medicine, biases such as anchoring, confirmation bias, and premature closure are well-recognized contributors to diagnostic error [12]. Patient safety science has therefore developed structured investigative methods that reconstruct reasoning pathways and identify the deeper contributors to adverse outcomes [13] [14].

Interestingly, while structured investigative frameworks are routinely applied to the analysis of human reasoning failure, comparable approaches have rarely been extended to artificial reasoning systems. As AI becomes more active in clinical decision-making, this gap becomes increasingly significant. AI systems may exhibit behaviours analogous to cognitive bias, including unsupported inference, premature analytical closure, overconfidence, and pattern overgeneralization derived from training data.

This paper addresses this gap by proposing a dual-layer architecture for structured AI reasoning and meta-cognitive reasoning audit in healthcare. The first layer adapts the DEEP SEE™ framework [15], originally developed as a bias-aware investigative methodology for root cause analysis in healthcare, into a structured reasoning protocol that guides how AI systems analyse clinical data and generate interpretable reasoning outputs. The second layer introduces the MIRROR framework as a meta-cognitive audit protocol designed to evaluate the integrity of AI-generated reasoning pathways by examining evidential support, logical consistency, alternative explanations, and uncertainty calibration. Rather than replacing existing artificial intelligence methods, the DEEP SEE™-MIRROR architecture operates as a complementary framework that enhances transparency, accountability, and reliability by making AI reasoning both structured and auditable. In doing so, the study extends principles from patient safety science and cognitive psychology to the evaluation and governance of artificial reasoning systems in healthcare.

2. Background and Related Work

The rapid advancement of artificial intelligence has generated extensive discussion regarding the safety, reliability, and governance of AI systems. As AI technologies become increasingly integrated into high-stakes environments such as healthcare, researchers and policymakers have emphasized the need for systems

that are transparent, robust, and accountable. Within this context, the concept of trustworthy artificial intelligence has become central to contemporary AI discourse.

Trustworthy AI generally refers to systems that operate according to principles such as transparency, fairness, robustness, accountability, and human oversight [6] [7]. These principles have informed several policy and governance initiatives, including the European Commission's ethics guidelines, which highlight transparency, explainability, and human agency as key requirements for responsible AI deployment [7].

A closely related line of research concerns AI alignment, which addresses the challenge of ensuring that AI systems behave in ways consistent with human intentions, ethical constraints, and safety expectations [8] [10]. Alignment considerations are particularly important in healthcare, where incorrect or poorly supported outputs may influence clinical decisions and directly affect patient outcomes.

In response to these concerns, significant effort has been directed toward improving interpretability and explainability in AI systems. Explainable artificial intelligence (XAI) aims to make machine learning models more understandable by providing insight into how predictions are generated [9]. Techniques such as feature attribution, model visualisation, and post-hoc explanation methods have improved transparency and may help support user trust.

Despite these advances, important limitations remain. Many explainability approaches focus primarily on identifying which variables influenced a prediction rather than evaluating whether the reasoning process itself is logically coherent, evidence-based, and sufficiently robust. In other words, existing approaches often explain what a model produced without systematically examining how that conclusion was derived.

A further concern involves hallucination, particularly in generative AI and large language models. Hallucinations occur when AI systems generate information that appears plausible but is factually incorrect or insufficiently supported by evidence [5]. In healthcare settings, such outputs may lead to unsafe interpretations or recommendations. Current strategies for reducing hallucination often focus on training improvements, retrieval mechanisms, or output verification rather than on systematic evaluation of the reasoning pathway itself.

Recent work on large language model evaluation in healthcare has highlighted the importance of assessing not only factual accuracy but also reasoning quality, uncertainty expression, and clinical reliability [16] [17]. Studies examining clinical large language models have demonstrated that outputs may appear coherent and authoritative while lacking sufficient evidential grounding or exhibiting inconsistent reasoning patterns. This has led to increasing interest in evaluation approaches that consider reasoning transparency, traceability of evidence, and alignment with clinical logic. However, current evaluation methods remain largely focused on output-level assessment rather than systematic examination of underly-

ing reasoning pathways.

Insights from cognitive psychology provide useful perspective on this issue. Decades of research have demonstrated that human judgment frequently relies on heuristics and may be influenced by cognitive biases under conditions of uncertainty [11]. In clinical medicine, biases such as anchoring, confirmation bias, and premature closure are well-recognized contributors to diagnostic error [12]. These insights have informed patient safety science, which employs structured investigative approaches to reconstruct decision pathways and identify how cognitive, contextual, and system factors contribute to adverse outcomes [13] [14].

A key lesson from patient safety research is that harmful outcomes rarely arise from a single incorrect decision. Instead, they typically emerge from interactions among incomplete information, cognitive bias, system pressures, and environmental constraints. Effective safety analysis therefore requires structured methods capable of examining reasoning processes in depth.

Despite this recognition, comparable approaches have rarely been applied to artificial reasoning systems. Most AI evaluation methods continue to focus on performance indicators such as accuracy, calibration, robustness, and dataset quality. While these metrics remain essential, they do not directly evaluate how AI systems reason. As AI systems become more involved in clinical interpretation, risk stratification, and decision support, this limitation becomes increasingly significant.

In parallel, the field of AI assurance and model monitoring has emerged to address the need for ongoing evaluation of AI systems after deployment [18] [19]. These approaches typically focus on detecting performance drift, calibration errors, bias across populations, and compliance with predefined safety thresholds. While such methods are essential for ensuring system-level reliability, they primarily operate at the level of outputs and aggregate performance metrics, without directly evaluating the reasoning processes underlying individual predictions or recommendations.

Taken together, existing approaches to explainability, model monitoring, and AI assurance provide valuable insights into model behaviour and performance. However, they do not offer a structured mechanism for examining the reasoning pathway itself as a unit of analysis. In particular, they do not systematically assess whether individual inferences are supported by evidence, whether reasoning chains are logically coherent, whether alternative explanations have been sufficiently considered, or whether uncertainty is appropriately represented. This limitation highlights the need for complementary frameworks that focus explicitly on reasoning integrity rather than solely on output explanation or system performance.

The present study addresses this gap by integrating two complementary ideas. First, it adapts the DEEP SEE™ framework—originally developed as a seven-step method for deeper, bias-aware root cause analysis in healthcare—into a structured reasoning protocol for artificial intelligence systems. Second, it introduces the

MIRROR framework as a meta-cognitive auditing protocol designed to evaluate the integrity of AI-generated reasoning pathways. Together, these frameworks extend beyond existing explainability and monitoring approaches by focusing explicitly on the structure and evaluability of reasoning processes, forming the foundation of the DEEP SEE-MIRROR architecture proposed in this study.

3. Conceptual Gap and Research Contribution

Despite substantial progress in AI safety, explainability, and governance, an important conceptual gap remains: artificial intelligence systems are typically evaluated based on their outputs rather than on the integrity of the reasoning processes that produce those outputs. While existing methods can indicate which variables influenced a prediction or whether a model performs well at a population level, they do not systematically assess whether the underlying reasoning pathway is evidence-based, logically coherent, and appropriately considers alternative explanations and uncertainty.

Current explainability approaches provide insight into model behaviour by identifying influential features or generating post-hoc explanations of outputs. However, these methods primarily address interpretability rather than reasoning validity. They may clarify why a model produced a particular output, but they do not determine whether the reasoning pathway itself is complete, logically consistent, or sufficiently supported by clinical evidence. Similarly, model monitoring and AI assurance frameworks focus on performance stability, bias detection, and calibration at a system level, yet they do not directly evaluate the reasoning processes underlying individual predictions or recommendations [18] [20].

This limitation is particularly evident in generative artificial intelligence systems, where outputs may appear coherent and clinically plausible despite being based on incomplete or weakly supported reasoning. Hallucination illustrates how AI systems can generate conclusions that are linguistically convincing yet insufficiently grounded in verifiable evidence [5] [16]. Existing mitigation strategies typically focus on improving training data, retrieval mechanisms, or output verification, but they do not systematically evaluate the reasoning pathway that led to the generated conclusion.

By contrast, patient safety science has long recognized that understanding failure requires examination of reasoning processes rather than outcomes alone. Investigations routinely reconstruct decision pathways, contextual influences, and cognitive patterns to identify how errors arise and propagate within complex systems. However, while such structured approaches are widely applied to human decision-making, they have rarely been adapted to examine artificial reasoning systems, despite increasing reliance on AI in clinical environments.

This gap reflects a broader limitation in current AI evaluation paradigms, which tend to prioritise predictive performance and output interpretability over systematic examination of reasoning processes. As AI systems become more deeply embedded in clinical workflows, the ability to evaluate how conclusions are gener-

ated—not only what conclusions are produced—becomes increasingly important for ensuring safety, trust, and accountability.

This study addresses that gap by proposing a dual-layer architecture designed to both structure and audit artificial intelligence reasoning in healthcare. The first contribution is the adaptation of the DEEP SEE™ framework into a structured reasoning protocol that guides how AI systems interpret clinical data, identify emerging safety signals, evaluate contributing factors, and generate interpretable reasoning outputs. The second contribution is the introduction of the MIRROR framework as a meta-cognitive auditing protocol that evaluates the integrity of AI-generated reasoning pathways by examining evidential support, logical consistency, alternative explanations, and uncertainty calibration. Together, these frameworks form the DEEP SEE-MIRROR architecture, which does not replace existing AI methods but operates as a complementary layer that enhances transparency, accountability, and reliability by making reasoning processes both structured and auditable. In doing so, the study extends principles from patient safety science and cognitive psychology to the governance and evaluation of artificial reasoning systems in healthcare.

4. The DEEP SEE-MIRROR Architecture

The DEEP SEE-MIRROR architecture consists of two sequential analytical layers designed to structure and audit artificial intelligence reasoning in healthcare systems. Rather than functioning as a standalone artificial intelligence model, the architecture operates alongside existing AI systems, including large language model-based and hybrid clinical decision-support systems, by structuring their reasoning processes and auditing the resulting reasoning pathways. The first layer applies the DEEP SEE™ framework as a structured reasoning protocol through which AI systems analyse clinical data and generate interpretable reasoning outputs. The second layer applies the MIRROR framework as a meta-cognitive auditing protocol that evaluates the evidential support, logical coherence, alternative explanations, and uncertainty calibration of the reasoning pathway produced by the AI system.

Within this architecture, DEEP SEE structures reasoning prospectively or during the analytical process by guiding how the AI system progresses from signal detection to structured clinical interpretation. MIRROR then operates as an auditing layer applied to the generated reasoning pathway, evaluating whether the resulting analysis is adequately supported by evidence, logically coherent, and appropriately calibrated in terms of uncertainty. The two layers therefore serve complementary but distinct functions: DEEP SEE structures how reasoning is generated, while MIRROR evaluates the integrity of that reasoning before it is presented or used in clinical decision-making.

In practical terms, the architecture can be implemented either within the reasoning process itself or as a post-processing layer applied to AI-generated outputs. In systems that allow access to intermediate reasoning steps, DEEP SEE can guide

the reasoning as it is generated, while MIRROR evaluates it before final presentation. In more opaque systems, both layers may be applied after output generation, restructuring and auditing the reasoning based on available information.

Figure 1 illustrates the conceptual workflow of the DEEP SEE™-MIRROR architecture, showing how clinical data is processed by AI systems, structured through the DEEP SEE reasoning framework, and subsequently audited through the MIRROR meta-cognitive protocol before informing clinical decision support.

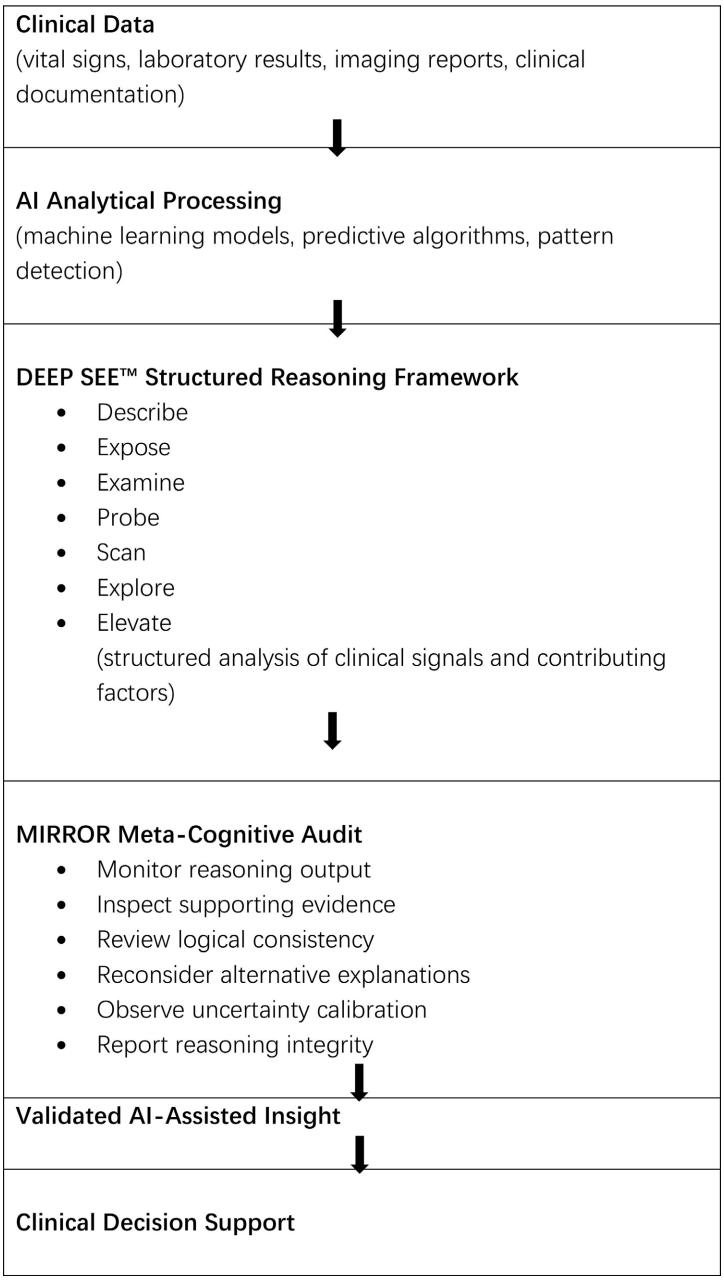

Figure 1. Conceptual workflow of the DEEP SEE™-MIRROR architecture, illustrating the progression from clinical data through AI analytical processing, structured reasoning using DEEP SEE™, and meta-cognitive audit using MIRROR, leading to validated AI-assisted clinical insight.

The architecture is intended for environments in which AI-generated conclusions may influence patient monitoring, diagnostic interpretation, escalation decisions, or clinical investigations. By introducing both structured reasoning and explicit reasoning audit, the DEEP SEE™-MIRROR model adds a complementary safety-oriented layer that enhances transparency, interpretability, and accountability in AI-assisted clinical decision systems. Rather than evaluating AI systems solely on predictive performance or output explainability, the architecture enables systematic examination of how conclusions are derived, thereby supporting more reliable and clinically meaningful use of artificial intelligence in healthcare.

Scope of Applicability across AI System Types

The applicability of the DEEP SEE™-MIRROR architecture depends on the level of transparency and accessibility of reasoning processes within the underlying AI system. In large language model-based and hybrid clinical decision-support systems, where intermediate reasoning steps, explanatory outputs, or evidence traces are available, the framework can be applied more comprehensively to structure and audit reasoning pathways. In contrast, conventional predictive models that produce probabilistic outputs with limited access to intermediate reasoning may only support partial application of the framework, focusing primarily on post-hoc structuring and audit of available outputs and explanatory features.

This distinction is important in practice, as it highlights that the effectiveness of reasoning audit is closely linked to the visibility of the reasoning pathway. Systems that expose richer reasoning information enable deeper and more reliable audit, whereas more opaque systems may limit the depth of evaluation that can be achieved.

5. DEEP SEE as a Structured Reasoning Protocol for Artificial Intelligence

Artificial intelligence systems used in healthcare frequently operate as statistical prediction tools that identify correlations within large datasets and generate probabilistic or language-based outputs. Although such systems can provide valuable insights, they often do so without explicitly structuring the reasoning pathway that connects clinical evidence to analytical conclusions. In safety-critical environments such as healthcare, this limitation reduces transparency and makes it difficult for clinicians to evaluate how AI-generated recommendations are derived.

The DEEP SEE™ framework, originally developed as a structured methodology for deeper, bias-aware root cause analysis in healthcare, can be adapted to guide artificial intelligence reasoning by transforming retrospective investigative steps into prospective or concurrent reasoning stages applied during data interpretation.

When applied to artificial intelligence systems, DEEP SEE functions as a structured reasoning protocol that organizes how clinical information is interpreted, how emerging safety signals are identified, and how conclusions are generated.

Rather than relying solely on statistical associations or unstructured language generation, an AI system guided by DEEP SEE™ analyses clinical data through a sequential reasoning process that progressively expands from observable signals to deeper contextual, process-related, and interpretive factors. This structured approach enhances transparency and supports more interpretable and clinically meaningful outputs.

The mapping of the seven stages of the DEEP SEE™ framework from human root cause analysis to artificial intelligence reasoning is presented in Table 1.

Table 1. Mapping of the DEEP SEE framework from human root cause analysis to artificial intelligence reasoning.

DEEP SEE Step	Function in Human Root Cause Analysis	Adapted Function in AI Reasoning	Example in AI Context
Describe	Identify the event or adverse outcome	Detect and summarise abnormal clinical signals	Elevated CRP, tachycardia, hypotension detected
Expose	Recognize underlying risk patterns	Identify emerging clinical risk trajectories	Pattern suggests possible early sepsis
Examine	Review actions taken and available evidence	Assess whether expected clinical actions or data are present	No cultures or antibiotics recorded
Probe	Identify process failures or missed steps	Detect gaps in expected clinical response pathways	Sepsis pathway not activated
Scan	Explore system-level and contextual factors	Evaluate workflow, documentation, and timing influences	Documentation attributes tachycardia to pain
Explore	Identify cognitive biases and assumptions	Detect potential reasoning bias patterns in interpretation	Anchoring on postoperative pain diagnosis
Elevate	Generate learning points and corrective actions	Produce structured clinical recommendations or alerts	Recommend sepsis workup and escalation

Table 1 provides a structured mapping of each DEEP SEE™ step from its original role in retrospective human investigation to its adapted function within artificial intelligence reasoning. This mapping clarifies how the framework transitions from analysing past events to actively guiding real-time or near-real-time reasoning processes in AI systems.

When implemented within artificial intelligence systems, these stages operate as sequential reasoning modules that structure how patient data are analysed and how conclusions are generated.

5.1. Operational Steps for AI Reasoning Using DEEP SEE™

Each step produces intermediate reasoning outputs that can be examined, traced, and subsequently evaluated within the MIRROR audit layer.

Step 1—Describe

The AI system identifies and summarises the observable clinical signal or abnormal event detected within the available data. Inputs may include abnormal vi-

tal signs, laboratory results, imaging findings, clinical documentation, or monitoring trends. The purpose of this step is structured signal detection.

Example output:

"Elevated inflammatory markers and haemodynamic instability detected over the past several hours, e.g. elevated CRP, tachycardia, and declining blood pressure detected in the last six hours."

Step 2—Expose

The AI system evaluates whether the detected findings represent a meaningful trajectory or emerging risk pattern that may not yet have been clinically recognised. This step focuses on identifying early patterns that suggest potential clinical deterioration.

Example: rising lactate, persistent tachycardia, and low blood pressure may expose a potential early sepsis trajectory.

Step 3—Examine

The system evaluates whether appropriate clinical actions or diagnostic steps have been initiated in response to the detected pattern. This includes assessing whether expected investigations or interventions have occurred in a timely manner.

Examples of checks:

- Were relevant diagnostic tests requested? e.g. Were blood cultures ordered?
- Were appropriate treatments initiated? e.g. Were antibiotics initiated?
- Was clinical escalation considered? e.g. Was diagnostic imaging requested?

Step 4—Probe

The AI system investigates potential process gaps or breakdowns in expected care pathways, such as delayed recognition, missing interventions, or incomplete evaluation. This step focuses on identifying where expected responses may not have occurred.

Example: sepsis indicators are present, but no cultures or antibiotics have been ordered.

Step 5—Scan

The system expands the analysis to consider contextual and system-level influences, including documentation patterns, workflow timing, communication signals, and monitoring processes that may affect clinical interpretation and response.

Step 6—Explore

This stage evaluates patterns that may reflect cognitive bias in clinical interpretation. Although AI systems do not possess human cognition, they can identify patterns in documentation or decision sequences that suggest potential biases such as anchoring, confirmation bias, or premature closure.

Example: tachycardia documented as postoperative pain without further reassessment.

Step 7—Elevate

The final stage converts analytical findings into actionable clinical insights,

such as escalation recommendations, clinical prompts, or structured explanations.

Example output:

"Findings suggest possible clinical deterioration: Possible early sepsis detected. Recommend further evaluation, appropriate investigations: e.g. blood cultures, repeat lactate measurement, and urgent clinical review."

5.2. Example of DEEP SEE™-Guided AI Analysis

Consider a hospital ward monitoring system that detects rising C-reactive protein, elevated lactate, persistent tachycardia, and mild hypotension.

Applying the DEEP SEE™ protocol:

Describe: Abnormal inflammatory markers and haemodynamic instability detected.

Expose: Pattern suggests possible early sepsis.

Examine: No blood cultures ordered and no antibiotics initiated.

Probe: Expected sepsis response pathway not activated.

Scan: Documentation attributes tachycardia to postoperative pain.

Explore: Possible anchoring bias in clinical interpretation.

Elevate: Recommend sepsis workup, repeat lactate, and urgent clinical review.

Through this structured pathway, the AI system generates not only an alert but also a transparent explanation of the reasoning process that produced the conclusion.

5.3. Role of DEEP SEE™ within the Architecture

Within the DEEP SEE™-MIRROR architecture, DEEP SEE™ functions as the primary reasoning layer that structures how AI systems interpret clinical information and generate analytical outputs. Its role is to ensure that reasoning progresses through a transparent, sequential, and clinically meaningful pathway.

However, structured reasoning alone does not guarantee validity. Even when reasoning is organised, conclusions may still be affected by incomplete evidence, unsupported assumptions, or overconfident interpretation. This highlights the need for a complementary mechanism that evaluates not only how reasoning is structured, but whether it is reliable.

For this reason, structured reasoning must be complemented by a mechanism capable of auditing the integrity of the reasoning process itself. The next section introduces the MIRROR framework as this meta-cognitive audit layer.

6. The MIRROR Framework: A Meta-Cognitive Audit Protocol for AI Reasoning

While DEEP SEE™ structures how AI systems analyse clinical data, it does not by itself guarantee that the resulting reasoning pathway is valid, complete, or free from bias. AI systems may still produce conclusions that contain unsupported assumptions, incomplete evidence, or overconfident interpretations. To address this limitation, this paper introduces MIRROR, a meta-cognitive audit protocol de-

signed to evaluate the integrity of AI reasoning processes.

In this context, meta-cognitive auditing refers to the systematic examination of how a conclusion was derived, focusing on the quality, completeness, and reliability of the reasoning pathway rather than the outcome alone. MIRROR therefore functions as a secondary analytical layer that evaluates whether AI-generated reasoning is evidence-based, logically coherent, and appropriately calibrated in relation to available data.

The framework is grounded in the concept of meta-cognition, understood here as the explicit evaluation of reasoning processes rather than conclusions alone. When applied to artificial intelligence systems, meta-cognitive auditing involves reconstructing the chain of inference, examining the evidential basis of each step, identifying potential gaps or unsupported assumptions, and assessing whether uncertainty is appropriately represented. This approach shifts the focus from what the AI system concludes to how that conclusion is formed and whether the reasoning pathway can be considered reliable.

6.1. Minimum Artifacts Required for MIRROR-Based Audit

The effectiveness of meta-cognitive reasoning audit depends on the availability of information that reflects how the AI system generated its conclusion. These audit artifacts may include, but are not limited to:

- Source clinical data (e.g., vital signs, laboratory results, imaging findings)
- Input prompts or task specifications provided to the AI system
- Retrieved knowledge sources or reference materials used during reasoning
- Intermediate reasoning outputs or inference steps (where available)
- Generated clinical interpretations or recommendations
- Confidence estimates or expressions of uncertainty
- Clinician inputs, overrides, or contextual annotations

In practical implementation, these artifacts may be captured through system logs, prompt-response traces, retrieval pipelines, or structured output formats. The availability and quality of these artifacts directly influence the depth and reliability of the audit process.

In systems where these artifacts are accessible—such as large language model-based or hybrid systems—MIRROR can be applied more comprehensively to reconstruct and evaluate the reasoning pathway. In contrast, in systems with limited transparency, the audit may rely on partial information, focusing primarily on output evaluation and available explanatory features. This distinction highlights that the depth of reasoning audit is dependent on the visibility of the reasoning process within the AI system.

6.2. MIRROR Auditing Steps

MIRROR evaluates AI reasoning through six sequential auditing steps, each of which examines a specific aspect of the reasoning pathway using available audit artifacts.

Step 1—Monitor Reasoning Output

The audit begins by reconstructing the reasoning pathway used by the AI system, including intermediate inferences and analytical steps.

Key question:

How did the AI reach this conclusion?

This step relies on available reasoning traces or outputs to reconstruct how the conclusion was generated.

Step 2—Inspect Supporting Evidence

Each inference should be linked to identifiable evidence such as laboratory results, imaging findings, patient monitoring data, or clinical documentation. Unsupported claims are flagged as potential hallucinations or weak inferences.

This evaluation depends on linking each inference to identifiable evidence within the available audit artifacts.

Step 3—Review Logical Consistency

This step evaluates whether the reasoning chain linking evidence to conclusions is logically coherent and whether unsupported reasoning shortcuts or inference gaps are present.

Step 4—Reconsider Alternative Explanations

To reduce premature analytical closure, the audit explicitly requires consideration of plausible alternative interpretations.

Example alternatives: pulmonary embolism, cardiogenic shock, dehydration.

Step 5—Observe Uncertainty and Confidence Calibration

This step evaluates whether the AI system's expressed confidence appropriately reflects the completeness, strength, and reliability of the available evidence.

Confidence evaluation considers whether uncertainty is appropriately aligned with the completeness and reliability of the underlying evidence.

Step 6—Report Reasoning Integrity

The final step produces a structured reasoning integrity report summarising supported inferences, missing data, alternative explanations considered, confidence calibration, and any identified reasoning risks. The MIRROR auditing steps are summarised in Table 2.

Table 2 summarises the MIRROR audit steps, including their purpose, key audit questions, and illustrative clinical applications.

6.3. MIRROR as the AI Reasoning Audit Layer

Within the DEEP SEE™-MIRROR architecture, MIRROR functions as a metacognitive safety layer that evaluates the integrity of AI reasoning after structured analysis has been completed. Its role is to ensure that conclusions are not accepted solely based on plausibility or predictive confidence, but are critically examined in terms of evidential support, logical consistency, consideration of alternative explanations, and appropriate uncertainty representation.

Importantly, MIRROR does not alter the original AI output; rather, it evaluates and qualifies that output by providing an explicit assessment of reasoning integ-

rity. This distinction allows the framework to be integrated into existing systems without modifying underlying models, while still enhancing transparency and safety.

Table 2. The MIRROR framework for meta-cognitive auditing of AI reasoning.

MIRROR Step	Purpose	Key Question	Example Application
M—Monitor Reasoning Output	Capture and reconstruct the reasoning pathway used by the AI system rather than examining only the final prediction.	How did the AI reach this conclusion?	AI detects fever, hypotension, and elevated lactate and links them to possible sepsis risk.
I—Inspect Supporting Evidence	Verify whether each inference is supported by identifiable data.	What evidence supports this claim, and is anything missing?	Check laboratory markers, cultures, imaging, or infection source supporting the interpretation.
R—Review Logical Consistency	Evaluate whether the reasoning chain logically connects evidence to the conclusion.	Does the evidence logically justify the inference?	Assess whether infection is inferred without confirming an infection source.
R—Reconsider Alternative Explanations	Encourage counterfactual reasoning to prevent premature analytical closure.	What other explanations could explain the same data?	Consider cardiogenic shock, pulmonary embolism, or dehydration rather than assuming sepsis.
O—Observe Uncertainty and Confidence Calibration	Evaluate whether the AI system expresses an appropriate level of confidence relative to the evidence.	Is the system overconfident given the available data?	AI expresses the diagnosis as a hypothesis rather than a definitive conclusion.
R—Report Reasoning Integrity	Produce a structured summary of reasoning quality, evidence, and uncertainty.	Can clinicians evaluate the reliability of the reasoning?	Generate a reasoning integrity report highlighting supported findings and missing evidence.

Depending on the level of transparency of the underlying AI system, this audit may be applied either comprehensively—when full reasoning traces are available—or partially, when only outputs and limited explanatory information can be accessed.

The architecture therefore operates as a two-stage protocol:

Stage 1—Structured reasoning using DEEP SEE™

The AI system analyses clinical data through a systematic reasoning process.

Stage 2—Meta-cognitive auditing using MIRROR

The resulting reasoning pathway is evaluated for evidential support, logical consistency, alternative explanations, and appropriate uncertainty calibration.

By combining structured reasoning with systematic audit, the DEEP SEE™-MIRROR architecture introduces a complementary safety-oriented mechanism

for evaluating artificial intelligence reasoning in healthcare. This dual-layer approach enables not only the generation of interpretable reasoning but also the critical examination of how that reasoning is constructed, thereby supporting greater transparency, accountability, and reliability in AI-assisted clinical decision-making.

7. Illustrative Proof-of-Concept Application of the DEEP SEE™-MIRROR Architecture

To demonstrate the practical application of the DEEP SEE-MIRROR architecture, this section presents an illustrative clinical scenario involving early detection of patient deterioration. The example is intended to show how structured reasoning and meta-cognitive audit can be applied to artificial intelligence outputs in a realistic healthcare context. **This proof-of-concept is illustrative rather than empirical and is designed to demonstrate feasibility and operational flow rather than clinical performance.**

7.1. Clinical Scenario

A hospitalized adult patient develops rising inflammatory markers, including elevated C-reactive protein and lactate levels, accompanied by persistent tachycardia and mild hypotension over several hours. An AI-based monitoring system processes these data and generates a preliminary clinical interpretation suggesting possible early sepsis.

7.2. DEEP SEE™-Guided AI Reasoning

The AI system applies the DEEP SEE™ structured reasoning protocol as follows:

- Describe: Elevated CRP, rising lactate, tachycardia, and mild hypotension detected.
- Expose: Pattern indicates a potential early sepsis trajectory.
- Examine: No blood cultures ordered; no antibiotic therapy initiated.
- Probe: Expected sepsis response pathway not activated.
- Scan: Documentation attributes tachycardia to postoperative pain without further reassessment.
- Explore: Possible anchoring bias in clinical interpretation.
- Elevate: Recommend sepsis workup, repeat lactate measurement, and urgent clinical review.

Through this structured process, the AI system generates not only an alert but also a transparent reasoning pathway that links clinical evidence to its conclusion.

7.3. MIRROR-Based Reasoning Audit

The MIRROR framework is then applied to evaluate the integrity of the generated reasoning:

- Monitor Reasoning Output: The reasoning pathway is reconstructed from the DEEP SEE™ stages.

- Inspect Supporting Evidence: Elevated CRP, lactate, tachycardia, and hypotension support the sepsis hypothesis; however, no microbiological confirmation is present.
- Review Logical Consistency: The inference from inflammatory markers and haemodynamic instability to possible sepsis is clinically plausible.
- Reconsider Alternative Explanations: Alternative causes such as dehydration, postoperative inflammation, or cardiogenic instability are considered.
- Observe Uncertainty and Confidence Calibration: The conclusion is appropriately framed as a probable diagnosis rather than a confirmed condition.
- Report Reasoning Integrity: The reasoning is partially supported, clinically plausible, and appropriately cautious, but dependent on further diagnostic confirmation.

7.4. Reasoning Integrity Outcome

The combined DEEP SEE™-MIRROR process produces a structured and auditable output:

- The reasoning pathway is transparent and traceable
- Key evidence supporting the conclusion is explicitly identified
- Alternative explanations are considered
- Uncertainty is acknowledged rather than overlooked
- Clinical recommendations are appropriately framed

In addition, the audit layer explicitly highlights the strengths and limitations of the reasoning process, allowing clinicians to understand not only what is suggested, but how reliable that suggestion is based on available evidence.

Compared with a standard AI alert that may simply indicate "high sepsis risk," the DEEP SEE™-MIRROR approach provides a more interpretable, clinically meaningful, and critically evaluated reasoning output.

7.5. Implications for Clinical Practice

This illustrative example demonstrates how the DEEP SEE™-MIRROR architecture can enhance AI-assisted clinical decision-making by transforming opaque predictions into structured and auditable reasoning processes. By combining systematic reasoning with explicit audit, the framework supports improved transparency, facilitates clinician trust, and enables more informed clinical judgment when interpreting AI-generated recommendations.

Importantly, this approach does not replace clinical decision-making but supports it by making the reasoning process visible, examinable, and open to critical evaluation.

8. Evaluation and Implementation Considerations for the DEEP SEE™-MIRROR Architecture

The practical adoption of the DEEP SEE™-MIRROR architecture requires consideration of how the framework can be integrated into existing artificial intelli-

gence systems, how reasoning integrity can be evaluated, and what operational implications may arise in clinical environments.

8.1. Integration within AI System Workflows

The DEEP SEE™-MIRROR architecture is designed to operate alongside existing AI systems rather than replace them. In practice, integration may occur at different stages of the AI workflow depending on system design and available interfaces.

In large language model-based and hybrid clinical decision-support systems, DEEP SEE™ can be implemented as a structured reasoning layer that guides how clinical data are interpreted and how outputs are generated. This may be achieved through prompt engineering, modular reasoning pipelines, or intermediate processing layers that enforce structured reasoning stages.

MIRROR can be implemented as a post-reasoning audit layer that evaluates the generated reasoning pathway using available artifacts such as intermediate outputs, retrieved evidence, and final recommendations. In more advanced implementations, elements of MIRROR may be partially automated, enabling real-time reasoning audit prior to presentation of outputs to clinicians.

In practice, this integration may take the form of an additional reasoning interface or audit panel within clinical systems, where structured reasoning outputs and audit summaries are presented alongside standard AI predictions.

8.2. Automation and Human Oversight

The extent to which the DEEP SEE™-MIRROR framework can be automated depends on system capabilities and the availability of reasoning artifacts. Certain components, such as structured reasoning steps and basic consistency checks, may be implemented algorithmically. However, higher-level evaluation—such as interpretation of alternative explanations or contextual clinical judgment—may still require human oversight, particularly in complex or ambiguous cases.

A hybrid approach is therefore likely to be most practical, in which automated reasoning structure and preliminary audit are supplemented by clinician review. This aligns with existing models of human-AI collaboration in healthcare, where AI systems support but do not replace clinical decision-making.

For example, an AI system may automatically generate a structured reasoning output and preliminary audit, while the clinician reviews flagged uncertainties or alternative explanations before acting on the recommendation.

8.3. Computational and Workflow Implications

The introduction of structured reasoning and audit layers may increase computational and workflow complexity. Additional processing steps may introduce latency, particularly in systems requiring real-time decision support. Furthermore, the effectiveness of reasoning audit is dependent on the availability and quality of clinical data and documentation.

From a workflow perspective, the presentation of structured reasoning and au-

dit outputs must be carefully designed to avoid cognitive overload for clinicians. Clear summarisation, prioritisation of key findings, and integration into existing clinical interfaces are essential to ensure usability and adoption.

In practice, this may involve presenting a concise summary of reasoning integrity (e.g., "moderately supported with missing evidence") rather than displaying the full reasoning chain unless further detail is requested.

8.4. Evaluation of Reasoning Integrity

Evaluating the effectiveness of the DEEP SEE™-MIRROR architecture requires metrics that extend beyond traditional performance indicators such as accuracy or calibration. The concept of reasoning integrity can be assessed through a combination of qualitative and quantitative indicators, including:

- Evidence traceability: the extent to which each inference is supported by identifiable clinical data

Example: The system explicitly links elevated CRP, lactate, and hypotension to the sepsis hypothesis rather than making an unsupported general statement.

- Logical coherence: consistency and validity of the reasoning chain

Example: The reasoning does not infer infection without supporting clinical indicators or contradict available data.

- Alternative explanation coverage: whether plausible alternative interpretations are considered

Example: The system considers dehydration, postoperative inflammation, and cardiogenic causes alongside sepsis.

- Uncertainty calibration: alignment between confidence and strength of evidence

Example: The output states "possible sepsis" rather than presenting a definitive diagnosis when evidence is incomplete.

- Audit completeness: degree to which all reasoning steps are explicitly evaluated

Example: All MIRROR steps are addressed, including evidence review, alternative explanations, and uncertainty assessment.

- Clinical interpretability: clarity and usefulness of reasoning outputs for clinicians

Example: The output provides a clear, structured explanation that a clinician can quickly understand and act upon.

These dimensions provide a framework for assessing how effectively the architecture improves transparency and reliability in AI-assisted reasoning. Future empirical studies may operationalise these indicators into measurable metrics and evaluate their impact on clinical decision-making and patient safety outcomes. These dimensions are summarised in Table 3.

9. Broader Applications of the DEEP SEE™-MIRROR Architecture in Healthcare AI

While the preceding section demonstrates an illustrative proof-of-concept application, the DEEP SEE™-MIRROR architecture has broader relevance across mul-

tiple domains in healthcare where artificial intelligence supports clinical reasoning and patient safety. These applications extend beyond individual case scenarios and include clinical decision support, patient monitoring, safety investigation, governance, and education. In each context, the framework contributes by transforming AI outputs into structured and auditable reasoning processes.

Table 3. Proposed dimensions for evaluating reasoning integrity in AI systems.

Dimension	Description	Example Assessment
Evidence Traceability	Each inference linked to identifiable data	CRP and lactate explicitly referenced
Logical Coherence	Reasoning follows valid clinical logic	No contradiction between data and conclusion
Alternative Coverage	Other explanations considered	Sepsis vs dehydration vs pulmonary embolism
Uncertainty Calibration	Confidence matches evidence strength	"Possible" vs "confirmed" diagnosis
Audit Completeness	All reasoning steps evaluated	All MIRROR steps completed
Clinical Interpretability	Output understandable to clinicians	Clear, structured explanation provided

9.1. Clinical Decision Support Systems

Clinical decision support systems analyse patient history, laboratory data, imaging findings, and clinical documentation to assist diagnostic and therapeutic decision-making. Many existing systems generate predictions or alerts without clearly presenting the reasoning processes that led to those conclusions. This lack of transparency can limit clinician trust and make it difficult to assess the reliability of AI-generated recommendations.

The DEEP SEE™-MIRROR architecture addresses this limitation by structuring how clinical information is interpreted and by auditing the reasoning underlying AI-generated outputs. DEEP SEE™ organizes reasoning into a sequential, interpretable pathway, while MIRROR evaluates whether that pathway is evidence-based, logically coherent, and appropriately calibrated. Together, these layers enable clinical decision support systems to produce outputs that are not only explainable but also critically evaluated in terms of reasoning integrity.

9.2. AI-Based Patient Deterioration Monitoring

Artificial intelligence is increasingly used to monitor hospitalized patients for early signs of deterioration, including sepsis, respiratory failure, and cardiovascular instability. Although such systems may improve early detection, they may also generate excessive alerts or overconfident predictions when evidence is incomplete or ambiguous.

By applying DEEP SEE™, monitoring systems can analyse patient deterioration through a structured reasoning process rather than relying solely on statistical thresholds or opaque alerts. MIRROR then evaluates whether the resulting conclusions are adequately supported by evidence, whether alternative explanations have been considered, and whether uncertainty is appropriately represented. This approach may enhance the clinical credibility of AI-generated alerts and support safer and more timely escalation decisions.

9.3. Incident Investigation and Patient Safety Analysis

Healthcare organisations routinely investigate adverse events and near misses in order to identify system vulnerabilities and improve patient safety. As artificial intelligence becomes increasingly embedded within clinical workflows, future investigations may need to examine not only human decision-making but also AI-assisted reasoning processes.

The DEEP SEE™-MIRROR architecture provides a structured framework for examining both human and AI-assisted reasoning processes within incident investigations. DEEP SEE™ can support reconstruction of clinical events and contributing factors, while MIRROR can evaluate whether AI-generated reasoning contained unsupported assumptions, incomplete evidence, or overconfident conclusions. This dual-layer approach enables more comprehensive analysis of how artificial and human reasoning interact within complex healthcare systems.

9.4. Governance and Oversight of Medical AI Systems

The growing use of AI in healthcare has intensified discussion regarding governance, monitoring, and responsible deployment. Regulatory agencies and healthcare institutions increasingly emphasise the need for transparency, risk management, and ongoing oversight of AI-assisted decision systems.

The DEEP SEE™-MIRROR architecture may support governance efforts by introducing a structured protocol for evaluating reasoning processes in addition to traditional performance metrics. Rather than assessing AI systems solely through accuracy or calibration, organisations can examine how conclusions are derived, whether reasoning pathways are consistent and evidence-based, and whether recurring reasoning risks or bias patterns are present. This supports a more comprehensive approach to AI assurance and oversight in healthcare.

9.5. Education and Training for AI-Assisted Clinical Reasoning

The integration of AI into healthcare also introduces new educational challenges. Clinicians must be able not only to use AI tools but also to critically evaluate the reasoning underlying AI-generated recommendations.

The DEEP SEE™-MIRROR architecture may serve as an educational framework for developing AI reasoning literacy among clinicians. By applying structured reasoning and meta-cognitive audit to AI outputs, clinicians can better understand how AI systems generate conclusions, identify potential reasoning limi-

tations, and critically evaluate AI-assisted recommendations. This may support safer and more effective integration of AI into clinical decision-making.

10. Ethical and Governance Implications of AI Reasoning Audit

The integration of artificial intelligence into healthcare introduces significant ethical and governance challenges. While AI technologies may improve diagnostic accuracy, predictive monitoring, and clinical decision support, they also create risks related to transparency, reliability, and accountability. A central ethical concern is not only whether AI systems produce accurate outputs, but whether those outputs are derived through reasoning processes that are evidence-based, logically coherent, and appropriately calibrated. This highlights the importance of evaluating reasoning integrity as part of responsible AI deployment in clinical environments.

Current AI governance discussions often focus on issues such as privacy, fairness, and explainability [6] [7]. Although these concerns remain critically important, an additional challenge involves the integrity of AI reasoning processes. Ensuring that AI-generated conclusions are produced through evidence-based and logically coherent reasoning pathways is essential for the safe use of AI in clinical environments.

The DEEP SEE™-MIRROR architecture addresses this challenge by introducing structured mechanisms for examining and auditing the reasoning processes through which AI systems generate conclusions.

10.1. Transparency and Interpretability

Many machine learning systems function as "black boxes," producing predictions without clearly interpretable reasoning pathways. In healthcare, such opacity complicates accountability and may reduce clinician trust in AI-generated recommendations.

MIRROR enhances transparency by reconstructing and systematically auditing the reasoning processes underlying AI-generated conclusions. Rather than focusing solely on output explanations, the framework requires explicit examination of the evidence, logic, and assumptions supporting each inference. In this way, MIRROR complements existing explainable AI approaches by shifting the focus from interpretability alone to the evaluation of reasoning integrity.

10.2. Accountability in AI-Assisted Decision-Making

The integration of AI into clinical workflows raises complex questions regarding responsibility when AI recommendations influence patient care. When adverse outcomes occur, determining whether the decision pathway was appropriately supported becomes essential.

Structured reasoning audit may support clearer accountability by documenting how AI-generated conclusions are derived and whether those conclusions are ad-

equately supported. In the event of adverse outcomes, this enables more precise evaluation of whether reasoning pathways were evidence-based, whether alternative explanations were considered, and whether uncertainty was appropriately recognised. Such transparency is essential for defining responsibility within AI-assisted clinical decision-making.

10.3. Mitigating Artificial Cognitive Bias

Although AI systems do not possess cognition in the human sense, they may exhibit reasoning patterns analogous to human cognitive bias, including overconfidence, premature analytical closure, and unsupported inference. Such patterns may arise from limitations in training data, probabilistic inference processes, or incomplete contextual information.

MIRROR addresses these risks by explicitly requiring evaluation of evidential completeness, logical consistency, alternative explanations, and appropriate confidence calibration. By introducing structured reasoning audit, the framework provides a systematic approach to identifying and mitigating forms of artificial cognitive bias that may arise from limitations in data, model design, or contextual interpretation.

10.4. Alignment with Patient Safety Principles

Healthcare ethics is strongly shaped by the principle of non-maleficence, which emphasises the obligation to avoid harm. When AI systems influence clinical decisions, this obligation extends to the safe design, evaluation, and governance of those systems.

The DEEP SEE™-MIRROR architecture supports this objective by introducing a structured mechanism for evaluating reasoning before it influences clinical decisions. By requiring explicit linkage between evidence and conclusions, consideration of alternative interpretations, and appropriate expression of uncertainty, the framework aligns with core patient safety principles and may help reduce the risk of unsafe AI-assisted recommendations.

However, the effectiveness of reasoning audit is dependent on the availability and quality of underlying data, as well as the transparency of AI systems. This highlights the need to integrate technical capability, ethical oversight, and organisational governance when implementing AI-assisted clinical decision-making frameworks.

11. Limitations

While the DEEP SEE™-MIRROR architecture provides a structured approach to improving transparency and evaluation of artificial intelligence reasoning, several limitations should be acknowledged.

First, the framework is dependent on the availability and quality of underlying clinical data and reasoning artifacts. In systems where intermediate reasoning steps, evidence links, or explanatory outputs are not accessible, the ability to per-

form comprehensive reasoning audit is limited. In such cases, MIRROR may only be applied partially, focusing on available outputs rather than full reasoning pathways.

Second, many artificial intelligence systems operate as proprietary or "black-box" models, restricting access to internal processes and limiting the depth of reasoning reconstruction. This may constrain the applicability of the framework in certain real-world implementations, particularly where transparency is restricted by system design or regulatory considerations.

Third, there is a potential risk of post-hoc rationalisation, particularly in systems that generate explanations after producing outputs. In such cases, reconstructed reasoning pathways may not fully reflect the internal processes that led to the original conclusion, which may affect the validity of reasoning audit.

Fourth, the introduction of structured reasoning and audit layers may increase computational and workflow complexity. Additional processing steps may introduce latency, and the presentation of structured reasoning outputs must be carefully designed to avoid cognitive overload for clinicians.

Fifth, the current study presents a conceptual framework supported by an illustrative proof-of-concept rather than empirical validation in live clinical systems. While the framework is grounded in established principles from patient safety science and cognitive psychology, further research is required to evaluate its effectiveness, feasibility, and impact in real-world healthcare environments.

Finally, the framework assumes that structured reasoning and audit processes will improve transparency and reliability; however, the extent to which this translates into measurable improvements in clinical outcomes, clinician trust, or patient safety remains to be empirically validated.

12. Future Research Directions

The DEEP SEE™-MIRROR architecture is proposed as a structured conceptual framework for improving the transparency and reliability of artificial intelligence reasoning in healthcare. While the present study provides an illustrative proof-of-concept, further research is required to evaluate its implementation, feasibility, and impact in real-world clinical environments.

Future studies may investigate the integration of the framework into clinical decision support systems and predictive monitoring platforms. In particular, comparative studies may evaluate AI systems with and without DEEP SEE™-MIRROR augmentation to assess differences in reasoning transparency, detection of unsupported inferences, clinical interpretability, and usability. Such work may help determine whether structured reasoning and reasoning audit improve clinician trust and decision-making quality.

Further research is needed to operationalise and validate metrics of reasoning integrity, including evidence traceability, logical coherence, alternative explanation coverage, and uncertainty calibration. Quantitative and qualitative evaluation approaches may be developed to assess how these dimensions relate to diagnostic

accuracy, clinical outcomes, and patient safety indicators.

Future work may also explore the development of standardised evaluation frameworks and benchmarking protocols to enable consistent assessment of reasoning integrity across different AI systems and clinical contexts.

Another important direction involves the development of automated or semi-automated implementations of the DEEP SEE™-MIRROR architecture. Advances in natural language processing, explainable AI, and reasoning trace extraction may support the creation of systems capable of structuring reasoning processes and performing meta-cognitive audit in real time within clinical workflows.

Additional research may examine the impact of structured reasoning and reasoning audit on early detection of patient deterioration, reduction of false alerts in monitoring systems, and consistency of clinical decision-making. The influence of the framework on clinician trust, adoption of AI systems, and interdisciplinary communication within healthcare teams also warrants further investigation.

Finally, the principles underlying the DEEP SEE™-MIRROR architecture may extend beyond healthcare to other high-stakes domains where reliable reasoning processes are critical, including aviation safety, financial risk analysis, and public policy decision-making. Future research may explore the adaptability and effectiveness of structured reasoning and audit frameworks in these contexts.

13. Conclusions

Artificial intelligence is rapidly transforming healthcare and other high-stakes domains in which complex decisions must be made under conditions of uncertainty. While AI technologies offer substantial opportunities to improve diagnostic accuracy, predictive monitoring, and clinical decision support, they also introduce challenges related to transparency, reliability, and accountability. In particular, current approaches often focus on what AI systems predict or explain, with comparatively less attention given to how those conclusions are derived.

This paper proposed the DEEP SEE™-MIRROR architecture, a dual-layer framework designed to structure and audit artificial intelligence reasoning in healthcare. The DEEP SEE™ layer provides a structured reasoning protocol that guides how AI systems analyse clinical data, identify emerging safety signals, and generate interpretable reasoning outputs. The MIRROR layer introduces a meta-cognitive audit process that evaluates the integrity of those reasoning pathways by examining evidential support, logical consistency, alternative explanations, and uncertainty calibration.

By combining structured reasoning with systematic audit, the DEEP SEE™-MIRROR architecture operates as a complementary layer alongside existing AI systems, enhancing transparency, accountability, and reliability without replacing underlying models. The framework shifts the focus of AI evaluation from outputs alone to the integrity of the reasoning processes that produce those outputs.

An illustrative proof-of-concept application demonstrates how the framework can transform AI-generated outputs into structured and auditable reasoning

pathways, while the proposed evaluation and implementation considerations outline how reasoning integrity may be assessed and integrated into real-world systems. At the same time, the study acknowledges important limitations, including dependence on data availability, system transparency, and the need for empirical validation in clinical environments.

Taken together, the DEEP SEE™-MIRROR architecture provides a practical and scalable approach for advancing trustworthy artificial intelligence in healthcare by making reasoning processes both structured and auditable.

From a patient safety and quality management perspective, the DEEP SEE™-MIRROR architecture can be viewed as aligning with principles of quality assurance and quality control, where structured reasoning supports error prevention and meta-cognitive audit supports error detection. In this sense, the framework reflects a quality-oriented approach to AI reasoning, consistent with broader healthcare quality and safety paradigms.

Conceptually, the DEEP SEE™-MIRROR architecture also reflects principles of metacognitive monitoring and regulation, in which reasoning processes are not only generated but also critically evaluated. In this sense, the framework introduces a form of structured "artificial metacognition" within AI-assisted clinical reasoning.

As artificial intelligence continues to expand across healthcare and other safety-critical domains, frameworks such as DEEP SEE™-MIRROR may contribute to the development of systems that are not only powerful and accurate, but also transparent, interpretable, and accountable. More broadly, this work highlights the importance of evaluating not only what artificial intelligence systems conclude, but how they reason, and whether those reasoning processes can be systematically examined and trusted.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al.* (2017) Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. https://doi.org/10.1038/nature21056

[2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P. and Ng, A.Y. (2017) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv: 1711.05225.

[3] Topol, E. (2019) Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[4] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the* 2021 *ACM Conference on Fairness, Accountability, and Transparency*, 3-10 March 2021, 610-623. https://doi.org/10.1145/3442188.3445922

[5] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., *et al.* (2023) Survey of Hallucination

in Natural Language Generation. *ACM Computing Surveys*, **55**, 1-38. https://doi.org/10.1145/3571730

[6] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., *et al.* (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, **28**, 689-707. https://doi.org/10.1007/s11023-018-9482-5

[7] European Commission (2019) Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu

[8] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. andMané, D. (2016) Concrete Problems in AI Safety. arXiv: 1606.06565.

[9] Gunning, D. and Aha, D.W. (2019) Darpa's Explainable Artificial Intelligence Program. *AI Magazine*, **40**, 44-58. https://doi.org/10.1609/aimag.v40i2.2850

[10] Russell, S. (2019) Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

[11] Kahneman, D. (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux.

[12] Croskerry, P. (2003) The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them. *Academic Medicine*, **78**, 775-780. https://doi.org/10.1097/00001888-200308000-00003

[13] Reason, J. (2000) Human Error: Models and Management. *BMJ*, **320**, 768-770. https://doi.org/10.1136/bmj.320.7237.768

[14] Vincent, C. (2010) Patient Safety. 2nd Edition, Wiley. https://doi.org/10.1002/9781444323856

[15] Bataweel, A.O. (2025) DEEP See™—A Seven-Step Framework for Deeper, Bias-Aware Root Cause Analysis in Healthcare. *Health*, **17**, 1081-1086. https://doi.org/10.4236/health.2025.179070

[16] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., *et al.* (2023) Large Language Models Encode Clinical Knowledge. *Nature*, **620**, 172-180. https://doi.org/10.1038/s41586-023-06291-2

[17] Nori, H., King, N., McKinney, S. M., Carignan, D. and Horvitz, E. (2023) Capabilities of GPT-4 on Medical Challenge Problems. arXiv: 2303.13375.

[18] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., *et al.* (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 33-44. https://doi.org/10.1145/3351095.3372873

[19] Sculley, D., *et al.* (2015) Hidden Technical Debt in Machine Learning Systems. *Proceedings of the* 29*th International Conference on Neural Information Processing Systems*, Montreal, 7-12 December 2015, 2503-511.

[20] Doshi-Velez, F. and Kim, B. (2017) Towards a Rigorous Science of Interpretable Machine Learning. arXiv: 1702.08608.