

REVIEW

Open access

Making Machine Reasoning Open to Question: An Integrative Review of Causal, Neuro-Symbolic, Knowledge-Grounded, and Uncertainty-Aware Explainability

Ana Silva^{1*}, Mohamed Al-Farsi², Luca Moretti³

Abstract

Explainable artificial intelligence increasingly seeks to provide more than visualisations of feature importance or simplified descriptions of model outputs. For intelligent systems used in consequential settings, machine reasoning should be open to questioning, inspection, contextualisation, correction, and evaluation under uncertainty. This integrative review examines emerging approaches that may support this broader objective. It connects causal explainability and counterfactual reasoning with neuro-symbolic intelligence, knowledge-grounded inference, uncertainty-aware explanation, reliability evaluation, and human-centred interpretability. The review uses a structured integrative approach to compare how these paradigms represent causes, alternatives, rules, concepts, knowledge sources, inferential steps, confidence, and limitations. Causal and counterfactual approaches can support questions about mechanisms, contrasts, interventions, actionability, and recourse, but their conclusions depend on explicit assumptions, feasible alternatives, and appropriate causal evidence. Neuro-symbolic and knowledge-grounded approaches can make concepts, relations, rules, provenance, and intermediate reasoning more traceable, although symbolic structure or retrieved knowledge does not automatically ensure faithful, valid, or truthful explanation. Uncertainty-aware approaches can qualify predictions and explanations through uncertainty characterisation, calibration, confidence communication, robustness assessment, and human interpretation, but uncertainty estimates alone do not establish reliability. The synthesis indicates that no individual paradigm is sufficient. More defensible machine reasoning requires complementary evaluation of explanation fidelity, causal plausibility, knowledge validity, reasoning traceability, uncertainty communication, cognitive accessibility, contestability, and lifecycle accountability. Important research opportunities concern causal validation, feasible recourse, provenance governance, explanation uncertainty, hybrid-system evaluation, human-centred testing, and cross-paradigm assessment. The review contributes an integrated account of how machine reasoning may become not merely visible, but meaningfully open to question.

Keywords Explainable AI, Machine reasoning, Causal explainability, Counterfactual reasoning, Neuro-symbolic AI, Knowledge grounding

*Correspondence:

Ana Silva

ana.silva@ulisboa.pt

¹ Department of AI and Reasoning Systems, University of Lisbon, Lisbon, Portugal

² Department of Computer Science, Sultan Qaboos University, Muscat, Oman

³ Department of Information Technology, University of Milan, Milan, Italy

Introduction

Explainable artificial intelligence has developed into a broad field encompassing interpretable modelling, post-hoc explanation, local and global analysis, visualisation, rule extraction, example-based explanation, and responsible-AI considerations. These approaches differ in what they expose, whom they address, and whether they aim to describe model behaviour, support understanding, enable oversight, or justify decisions. Consequently, explanation availability cannot be treated as a uniform property or as sufficient evidence that an intelligent system is trustworthy [1].

Many widely used explanation techniques illuminate correlations between inputs and outputs without revealing a complete inferential process. Feature attribution, surrogate modelling, decision rules, prototypes, and counterfactual examples may provide valuable access to model behaviour, yet each represents only a particular explanatory perspective. A system may therefore be locally explainable while its knowledge sources, causal assumptions, uncertainty, internal dependencies, or decision boundaries remain difficult to inspect [2].

Human explanation also differs from unrestricted technical disclosure. People commonly seek selective and contrastive answers: why one outcome occurred rather than another, what information mattered, what alternatives were possible, and how a decision could be challenged or changed. Explanations are consequently shaped by the recipient's question, prior knowledge, practical objective, and social context. Machine reasoning becomes meaningfully open to question only when explanatory information supports such situated inquiry [3].

This review integrates emerging approaches that move beyond surface-level visibility toward reasoning that can be interrogated, contextualised, qualified, and criticised. Its central contribution is to connect causal, counterfactual, neuro-symbolic, knowledge-grounded, and uncertainty-aware explainability through shared concerns about reasoning transparency, explanation fidelity, traceability, human critique, and reliability. The synthesis remains evaluative rather than prescriptive: explanation design must be aligned with the intended domain, users, decisions, and validation requirements before it can contribute to trustworthy practice [4].

Emerging Paradigms for Reliable Machine Reasoning

Established explainability research provides an essential foundation for reasoning-oriented approaches. Model-specific and model-agnostic methods can describe global behaviour, individual predictions, feature influence, decision boundaries, representative cases, and extracted rules. Emerging paradigms should therefore not be understood as replacements for conventional explainability. Rather, they expand the object of explanation from predictive behaviour toward mechanisms, alternatives, structured knowledge, inferential relations, and limitations [5].

Deep-learning explanations illustrate why this expansion is necessary. Attribution maps, relevance propagation, activation analysis, concept methods, and visualisation techniques can reveal how neural models respond to particular inputs. Nevertheless, sensitivity to a feature does not demonstrate that the feature constitutes a causal reason, that the explanation reproduces the complete decision process, or that the same rationale remains stable under meaningful changes. Reasoning transparency requires additional evidentiary and structural support [6].

Reliable machine reasoning also raises a design-level question: whether an opaque model should be explained after prediction or replaced by an interpretable model when a sufficiently accurate transparent alternative is available. In high-stakes contexts, post-hoc explanation can introduce approximation errors and may provide an appearance of understanding without exposing the actual decision logic. Interpretability, explanation fidelity, and model selection must therefore be considered together rather than treated as interchangeable remedies for opacity [7].

Causal explanation, counterfactual questioning, neuro-symbolic reasoning, knowledge grounding, and uncertainty communication represent complementary responses to these limitations. They seek to expose mechanisms, possible alternatives, symbolic relationships, evidence sources, inferential steps, or limits on confidence. However, none automatically establishes dependable reasoning. Their contribution depends on validated assumptions, faithful traces, reliable knowledge, comprehensible communication, and opportunities for human challenge, correction, and oversight [8].

Review Scope and Integrative Analytical Approach

The review addresses six integrative questions: how emerging paradigms make reasoning open to question; how causal and counterfactual approaches support explanation, action, and critique; how neuro-symbolic and knowledge-grounded approaches support traceability; how uncertainty-aware explanation contributes to reliable interpretation; where these approaches can converge; and which research gaps prevent accountable use. Evaluation is treated as multidimensional because existing XAI assessment remains fragmented across fidelity, robustness, plausibility, usefulness, stability, and human-centred criteria [9].

The analytical scope includes peer-reviewed journal scholarship relevant to explainable machine reasoning, causal explanation, counterfactual reasoning, neuro-symbolic intelligence, structured knowledge, uncertainty communication, and trustworthy AI. Sources were retained when they supplied conceptual distinctions, methodological comparisons, human-centred findings, evaluation principles, or integrative implications. Narrow performance studies without a clear reasoning, explanation, reliability, or accountability contribution were excluded. This approach distinguishes explanation objectives and evaluation targets rather than assuming that interpretability has a single accepted meaning [10].

Evidence extraction focused on the reasoning question addressed, the explanatory function performed, the relevant assumptions, the potential human use, the major failure or overclaiming risk, and the required validation or governance process. The synthesis also distinguished fidelity from plausibility, explanation from justification, causal reasoning from associative prediction, structured knowledge from validated knowledge, and confidence from demonstrated reliability. These distinctions reflect the pragmatic principle that explanatory value depends partly on the user’s epistemic situation and practical purpose [11].

The synthesis proceeded through cross-paradigm comparison rather than quantitative aggregation. Each approach was examined according to what can be questioned: model behaviour, causal assumptions, alternatives, rules, semantic relations, knowledge sources, inferential steps, uncertainty, or institutional responsibility. Human–agent explainability was therefore treated as an interaction involving an identifiable recipient, task, timing, and decision context rather than as static information presented independently of use [12].

Table 1 summarises the integrative review logic used to connect causal, neuro-symbolic, knowledge-grounded, and uncertainty-aware explainability as emerging paradigms for reliable machine reasoning.

Table 1. Integrative Review Logic for Reliable Machine Reasoning: Review Questions, Analytical Scope, Inclusion Criteria, Exclusion Criteria, Evidence Extraction Categories, Paradigm Mapping, Reasoning-Transparency Dimensions, Citation Planning, and Review Boundaries

Review component	Core integrative question or purpose	Inclusion logic	Exclusion logic	Evidence mapped	Reasoning or explainability dimension addressed	Bias or limitation risk	In re
Problem framing	What makes machine reasoning open to question?	Work addressing explanation, reasoning, reliability, or critique	Work limited to predictive accuracy	Definitions and conceptual boundaries	Reasoning transparency and human critique	Overly broad use of “explanation”	Es th
Causal stream	How are causes,	Causal explanation,	Associational interpretation	Causal claims,	Causal plausibility	Causal overclaiming	

	interventions, and assumptions represented?	inference, graphs, pathways, and interventions	presented as causality	assumptions, and validation needs	and intervention reasoning		qu bo
Counterfactual stream	How are alternatives, actionability, and recourse addressed?	Contrastive, counterfactual, what-if, and recourse research	Alternatives lacking explanatory relevance	Feasibility, proximity, actionability, and consequences	Alternative-based explanation	Unrealistic or ineffective alternatives	S ex gu r
Neuro-symbolic stream	How are learned and structured reasoning integrated?	Hybrid, symbolic, relational, compositional, and logic-based reasoning	Architecture descriptions without explanatory relevance	Integration points and reasoning traces	Structured reasoning and traceability	Assuming symbols ensure interpretability	C re f
Knowledge-grounded stream	What knowledge supports the inference?	Knowledge graphs, ontologies, grounding, provenance, and validity	Retrieval without reasoning or explanation relevance	Source lineage, semantic relations, and knowledge quality	Knowledge traceability	Treating retrieval as factual proof	C e so
Uncertainty stream	How are limitations and confidence communicated?	UQ, calibration, confidence, robustness, and uncertainty communication	Confidence values without reliability relevance	Uncertainty type, estimation, and interpretation	Qualified explanation and reliability	Miscalibration or user misunderstanding	Ir ur ex
Human-centred stream	Who questions the reasoning and for what purpose?	Human understanding, interaction, contestability, and trust calibration	Explanation evaluated only through technical availability	Audience, task, usefulness, and cognitive demands	Cognitive accessibility	Generalising across stakeholder groups	C t ex v
Synthesis and boundaries	Where can approaches converge without overclaiming?	Evidence supporting complementary functions or identified gaps	Claims of universal sufficiency or readiness	Trade-offs, convergence opportunities, and research needs	Cross-cutting accountability	Normative inference beyond evidence	F in co

Figure 1 illustrates the integrative analytical approach used to connect causal, counterfactual, neuro-symbolic, knowledge-grounded, and uncertainty-aware explainability as complementary routes toward machine reasoning that can be questioned.

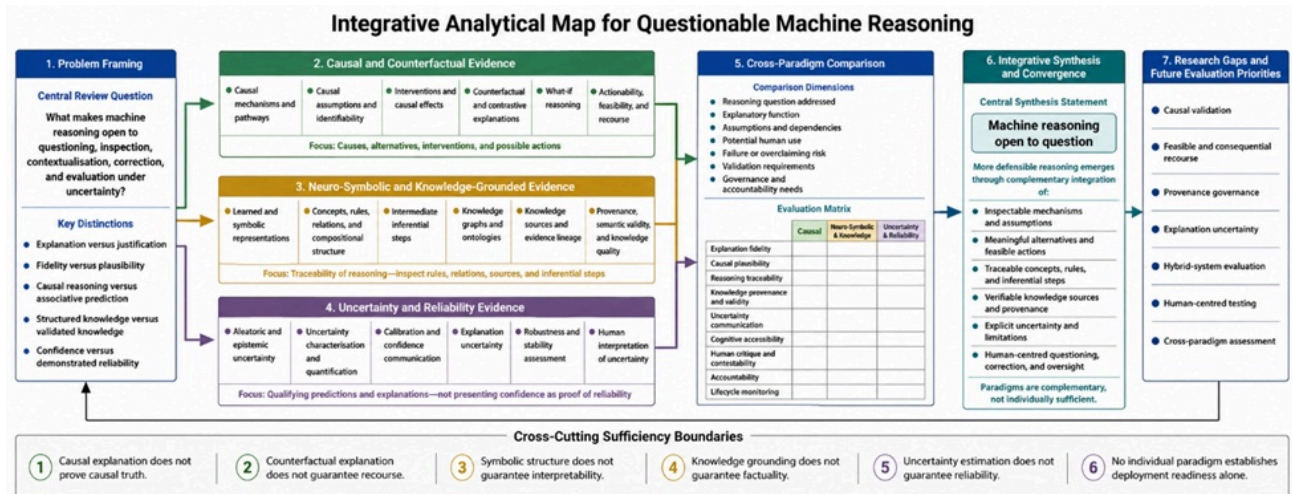


Figure 1. Integrative Analytical Map for Questionable Machine Reasoning

Alt-text description:

An integrative review map showing causal explainability, counterfactual reasoning, neuro-symbolic intelligence, knowledge grounding, uncertainty-aware explanations, reliability, human critique, accountability, synthesis, and future research gaps.

Causal Explainability and Counterfactual Reasoning

Causal explainability seeks to move from statistical association toward accounts of variables, mechanisms, and interventions that may influence an outcome. Causal reasoning asks what would occur if a condition were actively changed, whereas causal inference concerns whether such effects can be identified from data under stated assumptions. Causal representation learning may support mechanism-sensitive reasoning and robustness across changing environments, but its conclusions remain dependent on identifiability, invariance, data adequacy, and domain knowledge [13].

A causal explanation should therefore make its assumptions open to challenge. It may refer to a causal graph, pathway, mediator, intervention, or structural relationship, but the presence of causal terminology does not prove that the represented relation is correct. Counterfactual explanation can strengthen questioning by showing how an alternative state relates to a different outcome, yet a generated alternative can remain associational, infeasible, or inconsistent with the underlying data-generating process [14].

Counterfactual methods commonly optimise several competing properties, including proximity to the observed case, sparsity of change, plausibility, diversity, stability, and actionability. Improvements in one property may weaken another. A very close counterfactual can depend on an immutable characteristic, while a plausible alternative can require several costly changes. Counterfactual quality must consequently be evaluated against the intended explanatory and decision context rather than through a single generic criterion [15].

Contrastive explanation addresses why one outcome occurred rather than a specified alternative, while what-if reasoning explores how predictions respond to hypothetical changes. These approaches can be cognitively useful because they correspond to common human questions, but they should not be conflated with intervention-based causal reasoning. Their validity depends on meaningful foil selection, realistic feature dependencies, temporal order, and an explicit distinction between model sensitivity and changes that would occur in the world [16].

Actionability concerns whether a proposed change can be performed, whereas recourse asks whether feasible action is likely to produce a favourable future result. A nearby counterfactual does not guarantee either. Recommendations may fail because of causal dependencies, institutional constraints, model updates, delayed effects, or environmental uncertainty. Responsible recourse therefore requires feasibility assessment, causal and consequential validation, monitoring, and an accountable process through which affected people can question ineffective recommendations [17].

Table 2 maps causal explainability and counterfactual reasoning approaches according to reasoning questions, explanation functions, assumptions, validation needs, accountability roles, and boundary conditions.

Table 2. Causal Explainability and Counterfactual Reasoning for Questionable Machine Reasoning: Causal Explanation, Counterfactual Explanation, Contrastive Explanation, What-If Reasoning, Intervention Reasoning, Actionability, Recourse, Causal Graphs, Causal Assumptions, Feasibility Constraints, and Causal-Truth Boundaries

Approach or concept	Core reasoning question	Explanation function	Potential use	Failure or overclaiming risk	Validation or assumption need	Human critique or accountability role	Intervention
Causal explainability	Which mechanisms are claimed to produce the outcome?	Connects explanation with causal structure	Mechanism-oriented interpretation	Association may be presented as causation	Explicit model, assumptions, and domain evidence	Enables challenges to the proposed mechanism	Explains behaviour
Causal inference	What causal effect is identifiable?	Estimates effects under assumptions	Intervention and policy comparison	Confounding or misspecification	Identification strategy and sensitivity analysis	Exposes evidentiary premises	Disseminates causal knowledge
Causal explanation	What caused this outcome under the stated account?	Gives an outcome-specific causal justification	Decision and incident review	May imply excessive certainty	Causal definition, contrast, and evidence	Supports dispute over causes	Assesses mechanism
Counterfactual reasoning	What would happen under an alternative condition?	Compares actual and alternative states	Scenario exploration	Alternative world may be incoherent	Structural dependencies and temporal validity	Tests hidden dependencies	Suggests alternative behaviour
Counterfactual explanation	What must differ for the output to change?	Produces an output-changing alternative	Local explanation	Minimal change may be infeasible	Plausibility, stability, and causal consistency	Provides a concrete basis for challenge	Local explanation and action
Contrastive explanation	Why this outcome rather than the foil?	Selects differences relevant to a comparison	Human-centred explanation	Misleading foil selection	Explicit and meaningful comparison	Lets users define the relevant alternative	Assesses explanation with question

What-if reasoning	How does the model respond to hypothetical variation?	Reveals prediction sensitivity	Interactive exploration	Hypothetical values may violate dependencies	Realistic ranges and constraints	Supports probing of model behaviour	C expl witho p
Intervention reasoning	What follows from actively changing a variable?	Represents action within a causal system	Planning and decision support	Downstream effects may be omitted	Valid intervention semantics	Enables scrutiny of recommended actions	Bridge expl and
Actionability and feasibility	Can the proposed change realistically be made?	Filters alternatives by controllability	Practical guidance	Social or institutional barriers may be ignored	User- and context-specific constraints	Prevents impossible burdens being shifted to users	Gr count in cor
Recourse	Will feasible action improve a future decision?	Converts explanation into consequential guidance	Response to adverse decisions	Outcome improvement may not persist	Prospective validation and institutional support	Supports remedy and appeal	Seq expl t gua bo
Causal graphs and pathways	Which variables, edges, and mediators structure the account?	Makes relationships inspectable	Process analysis	Graphs may be incomplete or disputed	Structure validation and expert review	Allows challenges to nodes, edges, and omissions	Sup tra c repre
Causal assumptions	Which premises must hold?	Makes causal conditions explicit	Audit and sensitivity analysis	Hidden assumptions create false authority	Documentation and stress testing	Provides specific propositions to contest	C con ques rea
No causal-truth or recourse guarantee	What cannot be concluded from the explanation alone?	Defines interpretive boundaries	Responsible communication	Explanation may be mistaken for proof or remedy	Independent evidence and monitoring	Protects users from overclaiming	Preser suff bou

Neuro-Symbolic Intelligence and Knowledge-Grounded Inference

Neuro-symbolic intelligence combines statistical learning with structured representations or inference. Its architectures may connect neural perception with relational learning, constraints, rules, logic, or symbolic outputs. Because these combinations differ substantially, neuro-symbolic AI should be treated as a family of hybrid approaches rather than a single reasoning model. Its explanatory value depends on which component is exposed and whether the resulting trace corresponds faithfully to the system's operation [18].

The prospect of combining adaptive learning with explicit reasoning has motivated descriptions of neuro-symbolic AI as a further wave of intelligent-system research. Symbols, concepts, and rules may make some intermediate operations easier to inspect, but their presence does not make an entire system interpretable. Neural representations, interfaces between components, and conflict-resolution mechanisms can remain opaque and require independent evaluation [19].

Neuro-symbolic designs also differ according to where structured knowledge enters the system. Logic may constrain learning, supervise representation formation, guide inference, validate outputs, or appear only in the final explanation. Compositional and relational reasoning can potentially support reusable concepts and explicit relationships, yet systematic generalisation and trace fidelity must be demonstrated rather than inferred from architectural labels [20].

Knowledge-grounded explanation introduces identifiable domain concepts, entities, relations, or evidence into predictive reasoning. Knowledge graphs can contextualise outputs through semantic paths and structured associations, making it possible to inspect which concepts or relations supported an explanation. Their usefulness nevertheless depends on knowledge coverage, relevance, integration quality, and the extent to which graph-derived paths accurately represent the model's reasoning [21].

Ontologies, schemas, rules, and provenance records can strengthen traceability by identifying how knowledge was represented, sourced, and transformed. However, structured knowledge may remain incomplete, outdated, disputed, or incorrectly connected to a conclusion. Knowledge grounding should therefore expose source lineage, semantic commitments, and validity limitations rather than functioning as an automatic factuality guarantee [22].

Uncertainty-Aware Explanations and Reliability

Uncertainty-aware explainability qualifies what a model or explanation claims to know. Epistemic uncertainty concerns limitations in data, knowledge, or model specification and may sometimes be reduced through additional evidence. Aleatoric uncertainty concerns variability associated with the observed process or measurement conditions. These categories are conceptually distinct, although their practical separation can depend on modelling assumptions [23].

Deep-learning uncertainty can be estimated through Bayesian approximations, ensembles, predictive distributions, evidential methods, or related techniques. These approaches differ in computational demands, assumptions, calibration behaviour, and sensitivity to unfamiliar inputs. An uncertainty estimate should consequently be interpreted as a method-dependent output requiring validation rather than as a complete representation of all possible failure [24].

Interpretability and uncertainty can provide complementary information. An explanation may indicate which evidence influenced a prediction, while uncertainty may communicate how cautiously the prediction should be treated. Presenting these signals together may support faster calibration of reliance, but the effect depends on explanation fidelity, uncertainty quality, interface design, user expertise, and decision stakes [25].

Uncertainty communication should help users decide whether to accept, question, defer, corroborate, or seek another opinion. Raw probabilities or confidence labels may be insufficient when users cannot identify what is uncertain or what response is appropriate. Human-centred assessment should therefore test comprehension and decision behaviour rather than assuming that displaying uncertainty automatically improves judgement [26].

Reliability extends beyond confidence estimation. It includes dependable performance under intended conditions, calibration, robustness, external validity, monitoring, and appropriate responses to distribution change. Uncertainty quantification can support responsible decision assistance, but it cannot independently establish that a system is accurate, safe, fair, or operationally dependable [27].

Table 3 synthesises neuro-symbolic, knowledge-grounded, and uncertainty-aware explanation approaches, including reasoning traceability, knowledge provenance, uncertainty communication, reliability, and validation needs.

Table 3. Neuro-Symbolic, Knowledge-Grounded, and Uncertainty-Aware Explainability for Reliable Machine Reasoning: Symbolic Reasoning, Sub-Symbolic Learning, Knowledge Representation, Knowledge Graphs, Ontologies, Rule-Based Inference, Reasoning Traces, Knowledge Provenance, Uncertainty Quantification, Calibration, Confidence Communication, and Reliability Boundaries

Approach or concept	Core reasoning or reliability question	Knowledge, reasoning, or uncertainty function	Potential explanation role	Failure or overclaiming risk	Validation, governance, or monitoring need	Human critique or traceability role	
Neuro-symbolic AI	How are learning and structured reasoning connected?	Integrates neural and symbolic components	Exposes concepts or intermediate operations	Hybridisation may be mistaken for interpretability	Component and end-to-end evaluation	Enables cross-component inspection	le
Symbolic and rule-based reasoning	Which premises and rules produce the conclusion?	Applies explicit logic or conditions	Produces inspectable derivations	Valid logic may use invalid premises	Rule verification and domain review	Supports rule-level contestation	S t
Sub-symbolic learning	Which statistical patterns support the output?	Learns distributed representations	Supplies predictive inputs to structured reasoning	Internal representations remain opaque	Robustness and representation testing	Requires complementary explanatory access	r
Hybrid reasoning	How are conflicting reasoning modes reconciled?	Combines statistical and structured inference	Shows complementary evidence and constraints	Interfaces may conceal failure propagation	Integration and conflict-resolution testing	Supports scrutiny of component disagreement	En cc
Knowledge representation and ontology	Which concepts and relations define the domain?	Encodes entities, categories, properties, and constraints	Clarifies semantic assumptions	Categories may be incomplete or contested	Ontology governance and version control	Enables challenges to definitions	re s
Knowledge graph	Which entities and paths support the conclusion?	Organises relational knowledge	Provides contextual evidence paths	Graph structure may be inaccurate or incomplete	Quality, coverage, and relevance assessment	Supports path-level inspection	e wit
Knowledge provenance and validity	Where did knowledge originate, and is it appropriate?	Records source and transformation history	Enables evidence attribution	A cited source may not support the inference	Source authentication and correction workflows	Supports audit and correction	gro ac

Reasoning trace	Which steps led to the output?	Records intermediate operations	Makes inference inspectable	Generated traces may be incomplete or post-hoc	Fidelity and reproducibility testing	Provides concrete points for challenge	Op
Uncertainty quantification	What uncertainty can the model estimate?	Characterises predictive limitations	Qualifies conclusions	Estimates may be miscalibrated	External validation and recalibration	Supports cautious interpretation	Ad
Calibration	Does expressed confidence match observed performance?	Relates probability to empirical frequency	Makes confidence more interpretable	Aggregate calibration can conceal subgroup error	Contextual and subgroup monitoring	Enables scrutiny of confidence claims	cor
Confidence and uncertainty communication	How should limitations be presented to users?	Translates estimates into decision information	Supports deferral and corroboration	Displays may confuse or falsely reassure	Human-centred interface testing	Enables appropriate questioning	unc
Reliability and robustness	Does the system remain dependable under intended and changing conditions?	Evaluates performance and stability	Communicates operating boundaries	Explanation may substitute for reliability evidence	Lifecycle testing and shift monitoring	Supports escalation and oversight	Es

Integrative Synthesis of Emerging Approaches

An integrative account begins with stakeholder needs. Different users may seek causal understanding, actionable alternatives, evidence provenance, reasoning traces, uncertainty, or opportunities to appeal. No explanation paradigm can satisfy these requirements without specifying the recipient, decision context, explanatory objective, and property being evaluated [28].

Trustworthy machine reasoning is therefore multidimensional. Causal plausibility cannot substitute for knowledge validity, and traceability cannot substitute for reliability. Similarly, calibrated confidence does not establish fairness or accountability. Emerging approaches are most defensible when treated as complementary resources within broader responsible-AI evaluation rather than as independent proofs of trustworthiness [29].

Convergence also requires lifecycle governance. Causal assumptions, symbolic rules, knowledge sources, calibration, and explanation interfaces can become outdated as data, environments, institutions, or users change. Validation should therefore continue through monitoring, incident review, provenance management, reassessment of uncertainty, and clearly assigned responsibility for correction [30].

Human trust provides an important but insufficient evaluation outcome. Explanations may increase confidence or acceptance even when they are incomplete, persuasive, or poorly calibrated. The objective should be warranted reliance:

users should know when to accept system support, when to question it, and when to seek independent evidence [31].

Perceived explainability and causability can influence trust and acceptance, but subjective approval does not establish causal truth, fidelity, or technical reliability [32]. Visibility alone also fails to create accountability when users lack authority, expertise, or institutional mechanisms for challenge and remedy [33]. Explanatory access should therefore be proportionate to decision stakes and designed to support meaningful contestability rather than exhaustive disclosure for its own sake [34].

Table 4 presents the integrative synthesis of emerging explainability approaches, including cross-paradigm tensions, convergence opportunities, remaining gaps, and future research needs.

Table 4. Integrative Synthesis of Emerging Approaches for Questionable Machine Reasoning: Causal Reasoning, Counterfactuals, Neuro-Symbolic Intelligence, Knowledge Grounding, Uncertainty-Aware Explanation, Reliability, Human Critique, Accountability, Cross-Paradigm Tensions, Convergence Opportunities, and Future Research Priorities

Synthesis theme	Approaches integrated	Core integrative question	Evidence pattern	Tension or trade-off	Convergence opportunity	Remaining gap	Integ rev contri
Causal reasoning and uncertainty	Causal inference and UQ	How uncertain is the causal account?	Causal claims depend on assumptions	Qualification may reduce apparent simplicity	Communicate assumption-sensitive uncertainty	Limited joint evaluation protocols	Fra cause qual reaso
Counterfactuals and recourse	Counterfactual explanation and actionability	When does an alternative become useful guidance?	Proximity differs from feasibility and effect	Simplicity can conflict with realism	Combine constraints, causal effects, and uncertainty	Limited prospective validation	Sepa expla fro guar rem
Causal and knowledge grounding	Causal models and structured knowledge	Can knowledge expose causal premises?	Knowledge can represent mechanisms and constraints	Structured knowledge may be wrong	Link causal assumptions to validated provenance	Maintenance and consistency gaps	Makes pren mo critici
Neuro-symbolic traceability	Neural, symbolic, and hybrid reasoning	Is the reasoning trace faithful?	Symbolic components can expose steps	Neural dependencies may remain opaque	Validate traces across interfaces	Limited end-to-end fidelity evidence	Tre traceal test
Grounding and factual reliability	Knowledge graphs and provenance	Does the cited knowledge support the conclusion?	Grounding supplies identifiable evidence	More sources may increase noise	Jointly assess source quality and entailment	Weak user-centred evidence	Sepa grou from
Uncertainty and trust calibration	Explanation and uncertainty communication	Does information support	Signals may be complementary	More information may overload users	Adaptive communication by role and stakes	Limited long-term studies	Prior calib relie

		appropriate reliance?					
Hybrid reasoning and critique	Causal, symbolic, grounded, and uncertainty- aware approaches	What can users challenge?	Structured components create critique points	Technical detail can reduce accessibility	Layered and interactive explanations	Limited human testing	Def questio thro chall
Evaluation across paradigms	Fidelity, plausibility, grounding, calibration, and usefulness	Which criteria are shared or paradigm- specific?	Evaluation remains fragmented	Universal metrics erase distinctions	Modular evaluation profiles	Few shared protocols	Sup comp with univ sco
Reliability versus complexity	Multi- component reasoning systems	Do added components justify verification burdens?	Richer systems expose more assumptions	Complexity increases maintenance needs	Measure marginal explanatory value	Limited lifecycle evidence	Fra integra a trac
No single- paradigm sufficiency	All approaches	Can one paradigm establish trustworthy reasoning?	Each addresses only part of the problem	Integration creates new failure interfaces	Complementary validation and governance	Mature integrated evidence remains limited	Estab the re cer bour

Research Gaps and Convergence Opportunities

Causal explanation requires stronger validation of mechanisms, assumptions, and intervention claims. Counterfactual research also needs more realistic treatment of feasibility, temporal effects, institutional constraints, and uncertainty about whether recommended actions will succeed. Future work should connect causal modelling with participatory constraint identification, sensitivity analysis, and prospective recourse evaluation rather than relying primarily on geometric proximity [13].

Neuro-symbolic and knowledge-grounded systems require scalable evaluation that separates architectural structure from demonstrated interpretability. Research should assess whether reasoning traces are faithful, whether symbolic components improve critique, whether knowledge sources support conclusions, and whether provenance permits correction. Human-centred studies are especially needed to determine whether domain experts and affected users can understand and challenge these outputs [21].

Uncertainty research should extend from prediction uncertainty to explanation uncertainty. Attributions, counterfactuals, rules, retrieved evidence, and reasoning traces may vary across methods or model states even when the prediction is stable. Joint evaluation of fidelity, stability, calibration, communication, and downstream reliance could clarify when uncertainty-aware explanation supports appropriate decisions rather than additional complexity [25].

Cross-paradigm evaluation remains fragmented. Future studies should use modular profiles combining shared criteria, such as fidelity, robustness, usefulness, and contestability, with paradigm-specific criteria for causal validity, feasible recourse, trace correspondence, knowledge provenance, and uncertainty calibration. No universal reasoning-explanation

metric is likely to capture all purposes. Foundation-model applications further require cautious study of generated rationales, grounding, source use, abstention, and accountability without assuming that fluent reasoning traces are faithful.

Conclusion

This integrative review connects causal, counterfactual, neuro-symbolic, knowledge-grounded, and uncertainty-aware explainability as emerging routes toward machine reasoning that can be questioned rather than merely observed. Causal assumptions and interventions, alternative outcomes, symbolic structures, knowledge provenance, uncertainty communication, reasoning traces, and human critique provide complementary forms of explanatory access, but none is sufficient alone. Reliable machine reasoning should be assessed through fidelity, causal plausibility, feasibility, knowledge validity, calibration, robustness, cognitive accessibility, contestability, accountability, and lifecycle validation. The resulting synthesis supports disciplined integration and future research rather than a universal architecture, certification mechanism, or deployment-ready prescription.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 13 Nov 2025

Revised: 10 Jan 2026

Accepted: 09 Feb 2026

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.

<https://doi.org/10.1016/j.inffus.2019.12.012>.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):1-42.

<https://doi.org/10.1145/3236009>.

Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell*. 2019;267:1-38.

<https://doi.org/10.1016/j.artint.2018.07.007>.

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.

<https://doi.org/10.1016/j.jbi.2020.103655>.

Burkart N, Huber MF. A survey on the explainability of supervised machine learning. *J Artif Intell Res*. 2021;70:245-317.

<https://doi.org/10.1613/jair.1.12228>.

Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: A review of methods and applications. *Proc IEEE*. 2021;109(3):247-78.

<https://doi.org/10.1109/JPROC.2021.3060483>.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.

<https://doi.org/10.1038/s42256-019-0048-x>.

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. *Sci Robot*. 2019;4(37):eaay7120.

<https://doi.org/10.1126/scirobotics.aay7120>.

Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput Surv*. 2023;55(13s):1-42.

<https://doi.org/10.1145/3583558>.

Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion*. 2021;76:89-106.

<https://doi.org/10.1016/j.inffus.2021.05.009>.

Páez A. The pragmatic turn in explainable artificial intelligence (XAI). *Minds Mach*. 2019;29(3):441-59.

<https://doi.org/10.1007/s11023-019-09502-w>.

Rosenfeld A, Richardson A. Explainability in human-agent systems. *Auton Agent Multi-Agent Syst*. 2019;33(6):673-705.

<https://doi.org/10.1007/s10458-019-09408-y>.

Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. *Proc IEEE*. 2021;109(5):612-34.

<https://doi.org/10.1109/JPROC.2021.3058954>.

Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Inf Fusion*. 2022;81:59-83.

<https://doi.org/10.1016/j.inffus.2021.11.003>.

Guidotti R. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Min Knowl Discov*. 2024;38(5):2770-824.

<https://doi.org/10.1007/s10618-022-00831-6>.

Stepin I, Alonso JM, Catalá A, Pereira-Fariña M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*. 2021;9:11974-2001.

<https://doi.org/10.1109/ACCESS.2021.3051315>.

Karimi AH, Barthe G, Schölkopf B, Valera I. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput Surv*. 2023;55(5):1-29.

<https://doi.org/10.1145/3527848>.

Marra G, Dumančić S, Manhaeve R, De Raedt L. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artif Intell*. 2024;328:104062.

<https://doi.org/10.1016/j.artint.2023.104062>.

Garcez AD, Lamb LC. Neurosymbolic AI: The 3rd wave. *Artif Intell Rev*. 2023;56(11):12387-406.

<https://doi.org/10.1007/s10462-023-10448-w>.

Hitzler P, Eberhart A, Ebrahimi M, Sarker MK, Zhou L. Neuro-symbolic approaches in artificial intelligence. *Natl Sci Rev*. 2022;9(6):nwac035.

Tiddi I, Schlobach S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif Intell*. 2022;302:103627.

<https://doi.org/10.1016/j.artint.2021.103627>.

Hogan A, Blomqvist E, Cochez M, d'Amato C, de Melo G, Gutierrez C, et al. Knowledge graphs. *ACM Comput Surv*. 2021;54(4):1-37.

<https://doi.org/10.1145/3447772>.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506.

<https://doi.org/10.1007/s10994-021-05946-3>.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf Fusion*. 2021;76:243-97.

<https://doi.org/10.1016/j.inffus.2021.05.008>.

Tomsett R, Preece A, Braines D, Cerutti F, Chakraborty S, Srivastava M, et al. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns (N Y)*. 2020;1(4):100049.

<https://doi.org/10.1016/j.patter.2020.100049>.

Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digit Med*. 2021;4(1):4.

<https://doi.org/10.1038/s41746-020-00367-3>.

Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell*. 2019;1(1):20-3.

<https://doi.org/10.1038/s42256-018-0004-1>.

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif Intell*. 2021;296:103473.

<https://doi.org/10.1016/j.artint.2021.103473>.

Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: A review. *ACM Comput Surv*. 2023;55(2):1-38.

<https://doi.org/10.1145/3491209>.

Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: From principles to practices. *ACM Comput Surv*. 2023;55(9):1-46.

<https://doi.org/10.1145/3555803>.

Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. *Acad Manage Ann*. 2020;14(2):627-60.
<https://doi.org/10.5465/annals.2018.0057>.

Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int J Hum Comput Stud*. 2021;146:102551.
<https://doi.org/10.1016/j.ijhcs.2020.102551>.

Ananny M, Crawford K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc*. 2018;20(3):973-89.
<https://doi.org/10.1177/1461444816676645>.

Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in algorithmic and human decision-making: Is there a double standard? *Philos Technol*. 2019;32(4):661-83.
<https://doi.org/10.1007/s13347-018-0330-6>.