

ORIGINAL RESEARCH

Open access

Before Deployment, What Counts as Enough Evidence? A Benchmarking Framework for Trustworthy and Explainable AI Without Field Experiments

Sofia Lindqvist¹, Marcus Weber^{2*}, Priya Nair³, Thomas Andersen⁴

Abstract

Trustworthy and explainable artificial intelligence is frequently evaluated through model-performance benchmarks, laboratory experiments, robustness tests, fairness measures, or explanation metrics. Although these forms of evidence can reveal important technical properties, they do not independently establish that an AI system is suitable for consequential decision use. This article proposes a conceptual benchmarking framework for assessing what may count as sufficient pre-deployment evidence when field experiments, operational-performance observations, and real-world user-outcome data are unavailable. The framework distinguishes benchmark validity and coverage from technical readiness, separates robustness evidence from operational safety, and treats explanation usefulness as different from explanation availability or plausibility. It further integrates evidence quality, relevance, completeness, uncertainty, traceability, failure testing, boundary-condition analysis, human evaluation, appropriate reliance, residual-risk documentation, and governance review. Non-empirical decision-readiness criteria are introduced to support structured reasoning about whether an AI system has been characterized sufficiently to proceed to further evaluation, rather than to certify deployment. The framework emphasizes that benchmark scores must be interpreted in relation to the intended use case, decision context, stakeholder context, and risk environment. It also requires explicit identification of evidence gaps, unsupported inferences, monitoring needs, and claims that remain dependent on field evaluation. The principal contribution is an evidence-to-decision structure that helps researchers and decision makers organize pre-deployment trustworthy-AI evidence without collapsing technical performance, human-centered evaluation, decision readiness, and deployment approval into a single score or conclusion.

Keywords Explainable AI, Trustworthy AI, AI benchmarking, Evidence sufficiency, Technical readiness, Robustness testing

*Correspondence:

Marcus Weber

marcus.weber@lmu.de

¹ Department of AI and Society, Stockholm University, Stockholm, Sweden

² Department of AI Ethics and Reliability Engineering, Faculty of Informatics, LMU Munich, Munich, Germany

³ Department of Responsible AI and Human-Centered Computing, Indian Institute of Technology Bombay, Mumbai, India

⁴ Department of Explainable AI and Human-Computer Interaction, Faculty of Engineering, University of Copenhagen, Copenhagen, Denmark

Introduction

Pre-deployment evaluation has become a central problem in trustworthy artificial intelligence because broad agreement on principles does not automatically determine what evidence is sufficient for a particular system, use case, or decision. International guidance repeatedly invokes transparency, fairness, accountability, privacy, safety, and human oversight, yet these concepts are operationalized unevenly across technical and organizational settings [1]. A related ethical framework

identifies beneficence, non-maleficence, autonomy, justice, and explicability as distinct requirements, reinforcing that trustworthiness cannot be represented adequately by one benchmark score or undifferentiated readiness label [2].

Explainable AI adds another evidentiary challenge. Explanation methods vary in their purpose, technical assumptions, intended audience, and relationship to the underlying model. Some explanations are intended to support debugging, some to communicate influential features, and others to assist human decisions. Consequently, the existence of an explanation does not show that it is faithful, stable, understandable, relevant, or useful for the intended stakeholder. Explainability must therefore be evaluated as a multidimensional evidence domain rather than treated as a binary model property [3].

The problem is especially important where AI informs high-stakes decisions. Post-hoc explanations may appear persuasive while obscuring the behaviour of a complex model, creating a risk that explanatory presentation will be mistaken for genuine interpretability. In such settings, interpretable modelling and transparent decision logic may sometimes offer stronger evidentiary foundations than explanations attached to otherwise opaque systems [4]. Nevertheless, neither interpretable design nor explanation quality alone resolves questions of robustness, fairness, uncertainty, human reliance, organizational accountability, or field validity.

This article develops an original conceptual framework for benchmarking trustworthy and explainable AI before deployment and without field experiments. Its purpose is not to prescribe universal benchmarks, metrics, thresholds, or approval rules. Instead, it organizes pre-deployment reasoning around benchmark validity, evidence sufficiency, technical readiness, failure testing, robustness, boundary conditions, human evaluation, explanation usefulness, non-empirical decision-readiness criteria, residual risk, monitoring requirements, and future field-evaluation needs. The framework's central boundary is that structured evidence assessment may support a decision to continue evaluation, but cannot independently establish operational safety, legal compliance, ethical sufficiency, or deployment readiness.

Benchmarking Challenges in Trustworthy and Explainable AI

Benchmark validity can be undermined even when reported model performance appears strong. Data leakage, inappropriate separation between training and evaluation information, and analytical practices that unintentionally expose test information can produce optimistic results that are difficult to reproduce. Such failures are not minor procedural defects: they weaken the connection between the reported score and the capability the benchmark is assumed to measure [5]. A trustworthy-AI benchmarking process must therefore examine experimental separation, reproducibility, construct definition, and the evidentiary meaning of the selected task before interpreting comparative performance.

Benchmark coverage and representativeness depend substantially on the provenance, composition, collection process, intended use, and limitations of the evaluation data. Dataset documentation can make these assumptions visible by recording why data were created, which populations and variables are represented, how observations were collected, and which uses may be inappropriate [6]. Without this information, benchmark results may be transferred to populations, environments, decision workflows, or temporal conditions that were not represented during evaluation. Representativeness should therefore be treated as a contextual argument rather than a permanent characteristic of a dataset.

Benchmark bias also arises from the definitions used to operationalize fairness and harm. Machine-learning fairness includes multiple, sometimes incompatible, statistical and normative formulations, while bias may originate in sampling, measurement, labels, model design, evaluation, or institutional use [7]. A system may satisfy one fairness measure while failing another that is more relevant to the affected decision context. Fairness benchmarking must consequently state which disparity is being assessed, why the chosen formulation is appropriate, which groups are included, and what limitations remain.

Benchmark selection is only one component of a wider assurance problem. Evidence must be assembled across data management, model development, verification, validation, integration, human interaction, governance, and lifecycle change. Assurance scholarship therefore treats machine-learning dependability as a lifecycle concern rather than a result produced by one test or leaderboard [8]. This wider view exposes evaluation-context mismatch, use-case mismatch, risk-context mismatch, stakeholder mismatch, and the laboratory-to-field inference gap. It also establishes two essential boundaries: no benchmark is sufficient to prove overall trustworthiness, and no benchmark result alone establishes deployment readiness.

Evidence Sufficiency and Technical Readiness

Evidence sufficiency concerns whether the available evidence is adequate for the claims being considered, not whether a predetermined number of evaluations has been completed. Operationalizing responsible-AI principles requires methods that translate abstract commitments into inspectable assessments, documentation practices, impact analyses, and governance processes [9]. A sufficient evidence set must therefore be claim-specific and context-sensitive. It should identify which properties have been assessed, how directly each assessment relates to the intended use, and which questions remain unresolved.

Evidence completeness must also be distinguished from the presence of an ethics checklist. Reviews of AI-ethics guidance show that principles can receive uneven attention and may omit implementation, enforcement, political, social, or accountability concerns [10]. A technically detailed evidence package may consequently remain incomplete if it does not address affected stakeholders, decision authority, contestability, organizational responsibility, or foreseeable misuse. Completeness should be evaluated through an explicit coverage analysis rather than inferred from the prestige or apparent comprehensiveness of a framework.

Uncertainty is a further condition of technical readiness. In consequential decision support, point predictions and aggregate performance estimates are insufficient when the system cannot communicate when its outputs are unreliable or unsupported [11]. Pre-deployment evidence should examine uncertainty sources, confidence interpretation, calibration assumptions, and conditions requiring abstention, escalation, or human review. Uncertainty characterization does not eliminate risk, but it can expose where apparently precise outputs exceed the strength of available evidence.

Deep-learning uncertainty can arise from irreducible variability, limited knowledge, model assumptions, distribution differences, or insufficient data. Existing methods provide different approximations and may behave differently across applications and model architectures [12]. Consequently, uncertainty evidence should be evaluated for relevance to the intended decision rather than included as a generic technical feature. Calibration under development conditions should not be assumed to transfer automatically to new populations, temporal periods, workflows, or operational environments.

A mature evidence-sufficiency assessment must distinguish evidence quality, relevance, completeness, uncertainty, traceability, and identified gaps. It should also separate model readiness from broader technical and decision readiness. Surveys of neural-network uncertainty emphasize that unresolved uncertainty, out-of-distribution behaviour, and methodological variation remain important limitations of dependable AI evaluation [13]. Performance, reliability, fairness, explainability, privacy, security, and accountability evidence may all contribute to readiness reasoning where relevant, but none creates a universal sufficiency guarantee. Technical readiness is best interpreted as a qualified stage of characterization that may justify additional evaluation, not deployment.

Table 1 summarises benchmarking challenges, evidence-sufficiency requirements, and technical-readiness conditions that shape pre-deployment trustworthy AI evaluation.

Table 1. Benchmarking Challenges, Evidence Sufficiency, and Technical Readiness in Trustworthy and Explainable AI: Benchmark Validity, Coverage, Representativeness, Task Alignment, Evidence Quality, Evidence Relevance, Evidence

Completeness, Evidence Uncertainty, Technical Readiness, Performance Evidence, Reliability Evidence, Explainability Evidence, and Benchmark-Sufficiency Boundaries

Benchmarking or evidence requirement	Core readiness question	Required evidence or assessment	Benchmarking or readiness function	Potential output	Failure or overclaiming risk	Valid governance
Benchmark validity	Does the benchmark measure the intended construct?	Construct definition, task rationale, leakage review, reproducibility evidence	Tests whether results support the claimed inference	Validity-qualified benchmark statement	Strong scores may reflect invalid experimental design	Independent methodology
Benchmark coverage	Which relevant properties and scenarios are represented?	Coverage map across technical, human, risk, and governance dimensions	Identifies evaluated and unevaluated areas	Coverage and gap register	Unmeasured dimensions may be treated as satisfactory	Multidimensional representation
Representativeness	Do data and conditions reflect the intended context?	Population, environment, temporal, and workflow comparison	Assesses contextual similarity	Representativeness statement	Results may be generalized to absent groups or conditions	Domain stakeholder representation
Task alignment	Does the evaluation task reflect the proposed decision-support function?	Task-to-use-case mapping	Prevents proxy-task performance from becoming a use-case claim	Task-alignment assessment	Capability may be mistaken for decision usefulness	Use-case alignment
Evidence quality	Is the evidence methodologically credible?	Design, reproducibility, provenance, and limitation assessment	Filters weak or invalid evidence	Quality-qualified evidence register	Quantity of evidence may conceal poor quality	Independent validation
Evidence relevance	Does each item support the exact claim being considered?	Claim-to-evidence mapping	Prevents indirect evidence from being overstated	Relevance matrix	Broad literature may be used to support specific claims	Expertise
Evidence completeness	Are all material dimensions represented?	Multidimensional coverage audit	Identifies missing evidence families	Completeness profile	A checklist may appear complete while omitting context	Governance representation
Evidence uncertainty	How uncertain are the evidence	Statistical, model, data,	Qualifies confidence	Uncertainty statement	Point estimates	Statistical domain

	and resulting inferences?	and interpretive uncertainty review			may imply false precision	
Evidence traceability	Can every conclusion be linked to its source and limitation?	Provenance, documentation, and change history	Supports reviewability	Claim-evidence ledger	Evidence may be reused outside its original scope	Docu gove
Technical readiness	Is the model sufficiently characterized for further assessment?	Performance, reliability, uncertainty, fairness, and testing evidence	Supports an intermediate readiness judgment	Qualified technical-readiness profile	Technical maturity may be equated with operational readiness	Multid stage
Performance evidence	What capability is demonstrated under defined conditions?	Task-relevant estimates and subgroup analysis	Establishes bounded capability	Context-qualified performance profile	Average performance may conceal failures	Inde eva
Reliability evidence	Is behaviour consistent under relevant repetition and variation?	Repeatability, stability, and sensitivity assessment	Examines consistency	Reliability profile	Narrow repeatability may be described as robustness	Scena re
Explainability evidence	Which explanation properties were evaluated for whom and why?	Fidelity, stability, usefulness, uncertainty, and accessibility evidence	Evaluates explanatory support	Explanation-evidence profile	Explanation presence may be equated with understanding	Tech humar re
Benchmark-sufficiency boundary	What relevant evidence remains outside the benchmark portfolio?	Limitation and evidence-gap statement	Sets the ceiling of permissible inference	Non-sufficiency declaration	Benchmark completion may resemble certification	Gove acknov

Failure Testing, Robustness, and Boundary Conditions

Machine-learning testing requires more than confirming ordinary predictive performance. Testing research distinguishes properties such as correctness, fairness, interpretability, security, privacy, efficiency, and robustness, while also examining data, learning programs, model behaviour, and system integration [14]. A trustworthy-AI test strategy should therefore begin with explicit failure hypotheses: what may fail, under which conditions, for which stakeholders, and with what consequence. Passing a general test suite is less informative than understanding the failure classes that the suite was designed to expose.

Existing testing research remains fragmented across model types, system components, quality properties, and test-generation strategies [15]. This fragmentation makes coverage analysis essential. Stress testing, sensitivity testing, data-quality testing, edge-case analysis, and scenario testing should be linked to the intended use and risk context rather than

selected solely because they are common in the literature. A test portfolio should show which assumptions were challenged, which operating regions were explored, and which important conditions remain outside the assessment.

Robustness evidence must also be separated from security, interpretability, and safety evidence. Reviews of deep-neural-network trustworthiness treat verification, testing, adversarial attack, defence, and interpretability as related but distinct research areas [16]. Resistance to a specified perturbation does not establish resistance to other perturbations, intentional attacks, distribution shifts, or harmful system interactions. Similarly, adversarial testing is meaningful only relative to a stated threat model and cannot demonstrate protection against all possible attackers or manipulation strategies.

Conventional software-testing assumptions are difficult to transfer directly to machine-learning programs because learned behaviour depends on training data, model structure, optimization, and statistical generalization [17]. Testing may therefore incorporate metamorphic relations, differential comparisons, mutation-based strategies, controlled data transformations, and sensitivity analysis. Within the proposed framework, these methods are not prescribed universally; they are examples of test forms that may be justified by the system and risk context. Simulation, synthetic cases, or red-team exercises may likewise extend coverage when their assumptions and limitations are documented.

Failure testing ultimately produces bounded evidence, not proof that all harmful behaviour has been eliminated. Reviews of safety-critical machine learning identify continuing challenges in integrating robustness, uncertainty, verification, explainability, safe learning, and broader assurance arguments [18]. Residual risk should therefore be documented after mitigation, alongside untested conditions, uncertain interactions, and requirements for monitoring or field evaluation. Risk communication must explain both what was tested and what cannot be inferred. The framework consequently includes a no robustness-safety guarantee boundary and a no failure-testing sufficiency boundary.

Table 2 maps failure testing, robustness testing, boundary-condition testing, out-of-distribution testing, and residual-risk requirements for trustworthy AI benchmarking before deployment.

Table 2. Failure Testing, Robustness, Boundary Conditions, and Residual Risk in Pre-Deployment AI Benchmarking: Stress Testing, Adversarial Testing, Edge Cases, Out-of-Distribution Testing, Distribution Shift, Data Quality Testing, Sensitivity Testing, Scenario Testing, Failure Mode Analysis, Risk Mitigation, Risk Communication, and Robustness-Safety Boundaries

Testing or boundary requirement	Core failure question	Required evidence or test condition	Risk-control function	Potential output	Failure or residual risk	Review or monitoring need	F
Failure testing	How may the system fail under foreseeable conditions?	Explicit failure hypotheses and targeted tests	Makes known failure pathways inspectable	Failure register	Unknown and interacting failures remain	Periodic reassessment	C
Stress testing	How does behaviour change near or beyond expected limits?	Escalating noise, load, scarcity, or complexity	Reveals degradation patterns	Stress-response profile	Tested stressors may not reflect real combinations	Scenario revision	Li
Adversarial testing	How does the system respond	Defined threat model and	Identifies security and	Threat-bounded findings	Unmodelled adversaries remain	Threat-model updating	Se

	to intentional manipulation?	attack conditions	robustness vulnerabilities				
Robustness testing	Is behaviour stable under justified perturbations?	Context-relevant perturbation assessment	Examines sensitivity to variation	Robustness profile	Artificial perturbations may lack contextual relevance	Domain review	
Boundary-condition testing	Where does dependable behaviour cease?	Tests near data, task, and operating boundaries	Defines the supported region	Boundary map	Boundaries may shift over time	Monitoring and re-evaluation	Sc
Edge-case analysis	Which rare but plausible cases create disproportionate risk?	Expert-derived and data-derived edge cases	Exposes low-frequency hazards	Edge-case register	Unknown rare events remain	Incident-informed updating	F
Out-of-distribution testing	How does the model respond to unfamiliar inputs?	Defined shift scenarios and detection assessment	Reveals generalization limits	Out-of-distribution profile	Novelty cannot be exhaustively represented	Drift monitoring	val
Distribution-shift testing	How sensitive is performance to population, temporal, or environmental change?	Plausible held-out domains or simulated shifts	Tests change vulnerability	Shift-sensitivity statement	Simulated shifts may not match field change	Longitudinal evaluation	
Data-quality testing	How do error, missingness, noise, or coding differences affect behaviour?	Controlled data degradation	Identifies data-dependence failures	Data-quality sensitivity map	New pipeline defects may emerge	Data-pipeline monitoring	a
Sensitivity testing	Which inputs or assumptions materially alter outputs?	Structured variation of influential factors	Reveals fragile dependencies	Sensitivity profile	Variable interactions may remain hidden	Expert review	A tes
Scenario testing	How does the system behave across normal, adverse, and ambiguous situations?	Use-case and risk-linked scenarios	Connects evaluation to decision context	Scenario evidence map	Scenario design may omit stakeholder concerns	Participatory review	Co
Simulation or synthetic testing	Can controlled representations	Documented simulation or	Expands rare-condition testing	Qualified simulated evidence	Simulator-to-reality and	Comparison with authentic data	C tes

	extend coverage safely?	generation assumptions			synthetic-data gaps		
Failure-mode analysis	Which causal pathways may produce invalid or harmful outcomes?	Structured failure decomposition	Connects tests to risks and mitigations	Failure-mode map	Complex interactions may remain uncertain	Cross-functional review	o
Risk mitigation	Do proposed controls address identified failures?	Control-to-failure mapping and retesting	Evaluates mitigation relevance	Mitigation evidence profile	Controls may introduce new failure modes	Change review	Co
Risk communication	Are limitations and remaining risks understandable?	Audience-appropriate disclosure and comprehension review	Supports informed governance	Risk-communication record	Disclosure may be misunderstood or ignored	Human-centered assessment	Co
Residual risk	What remains unresolved after available testing and mitigation?	Residual-risk and unknown-condition register	Prevents zero-risk conclusions	Residual-risk statement	Unknown hazards cannot be eliminated	Governance review and monitoring	F
Robustness-safety boundary	Why does robustness evidence not establish safety?	Explicit distinction between perturbation resistance and harm prevention	Limits safety inference	Non-guarantee statement	Narrow robustness may be framed as general safety	Independent governance review	f
Failure-testing sufficiency boundary	Which possible failure classes remain untested?	Coverage-limit and unknown-failure statement	Prevents exhaustive-test claims	Open-failure register	Passing known tests may conceal unknown hazards	Lifecycle updating	f

Human Evaluation and Explanation Usefulness

Human evaluation is necessary because explanations are interpreted through cognitive, social, and decision contexts rather than received as purely technical objects. Research drawing on the social sciences shows that people commonly prefer explanations that are selective, contrastive, responsive to particular questions, and situated within existing beliefs or expectations [19]. A technically generated explanation may therefore be accurate in one formal sense yet fail to address what an intended user needs to know. Pre-deployment evaluation should distinguish whether an explanation is intended to support comprehension, error detection, justification, contestation, learning, or decision support.

Explanation usefulness also depends on pragmatic context. The adequacy of an explanation cannot be judged solely by its internal structure because users approach AI outputs with different questions, expertise, responsibilities, and practical goals [20]. An explanation designed for a model developer may be unsuitable for a professional decision maker or

affected person. Evaluation should therefore identify the intended audience, the decision being supported, the time available for interpretation, and the consequences of misunderstanding or ignoring the explanation.

A human-centered evaluation plan should connect stakeholder groups and system objectives to appropriate evaluation methods. Different users may need global model information, case-specific explanations, uncertainty communication, actionable recourse, or evidence of model limitations. Evaluation frameworks for explainable AI consequently recommend matching user groups, design goals, and evaluation approaches rather than assuming that one usability test can establish general explanatory value [21]. Domain expertise should also be distinguished from technical expertise because users may understand the subject matter while lacking detailed knowledge of model construction.

Explanation evidence must identify the property being evaluated. Fidelity concerns whether an explanation reflects relevant model behaviour; stability concerns whether materially similar cases receive appropriately consistent explanations; consistency concerns whether repeated or alternative explanation methods produce conflicting accounts; and completeness concerns whether relevant factors or limitations have been omitted. Reviews of explainability evaluation demonstrate that these properties are conceptually distinct and require different forms of assessment [22]. An explanation can be clear but unfaithful, stable but incomplete, or technically plausible but inaccessible to its intended audience.

Human evaluation should ultimately examine whether explanations support accurate understanding, appropriate reliance, and valid decision processes. A systematic review of explainable-AI evaluation found substantial variation in quantitative, qualitative, and anecdotal assessment practices, with many evaluations addressing only a limited subset of explanation properties [23]. Pre-deployment evidence should therefore document what users understood, which tasks were evaluated, and what remains unknown. Neither positive user reactions nor technically favourable explanation measures establish universal usefulness, prevent overreliance or underreliance, or prove that laboratory behaviour will transfer to real decision environments.

Non-Empirical Criteria for Decision Readiness

Non-empirical decision-readiness criteria can help determine whether an AI system has been characterized sufficiently to justify further controlled evaluation. Human-centered AI emphasizes that reliable automation should coexist with meaningful human control, responsibility, and the ability to intervene when system behaviour becomes uncertain or inappropriate [24]. Decision-readiness reasoning should consequently define what the AI may support, which decisions remain under human authority, when escalation is required, and which uses are excluded.

Trust should not be treated as a substitute for trustworthiness. Empirical research on human trust in AI shows that trust is influenced by system representation, observed performance, process transparency, user expectations, and contextual factors [25]. High trust may produce overreliance, whereas low trust may lead users to reject useful assistance. A pre-deployment framework should therefore seek evidence of calibrated reliance: users should accept outputs when confidence is justified, question them under uncertainty, and retain the ability to override or disregard inappropriate recommendations.

Decision-readiness criteria must also specify the intended relationship between human expertise and computational capability. Human-AI symbiosis is particularly relevant where AI can process large volumes of information while people contribute contextual judgment, social understanding, ethical reasoning, and the ability to interpret ambiguous situations [26]. The framework should document which capabilities are delegated, which remain human responsibilities, and how disagreements between users and AI outputs are resolved. Such role definition is a conceptual prerequisite for responsible evaluation, not evidence that the arrangement will function effectively in practice.

Organizational decision structures determine who can authorize, review, contest, or override AI-supported decisions. Different configurations may assign decisions to an AI system, a human decision maker, or a combined structure

depending on task characteristics and organizational conditions [27]. Pre-deployment readiness reasoning should therefore include accountability evidence, decision rights, review procedures, recordkeeping expectations, and escalation routes. Formal responsibility assignments remain provisional until their practical operation is examined within real workflows.

Automation and augmentation should not be treated as mutually exclusive or inherently beneficial alternatives. Their interaction may create tensions involving expertise, control, responsibility, work design, and organizational learning [28]. Pre-deployment benchmarking can support decision-readiness reasoning only when use-case alignment, evidence sufficiency, technical readiness, failure coverage, explanation usefulness, human evaluation, residual risk, and field-evaluation needs remain distinct from deployment approval. The framework may identify a provisional basis for additional testing, but it cannot establish that an AI system is operationally safe or ready for real-world use.

Figure 1 illustrates the pre-deployment evidence-to-decision pathway, showing how benchmark evidence, technical readiness, failure testing, human evaluation, non-empirical criteria, residual risk, and field-evaluation needs should be separated before any deployment-oriented claim is made.

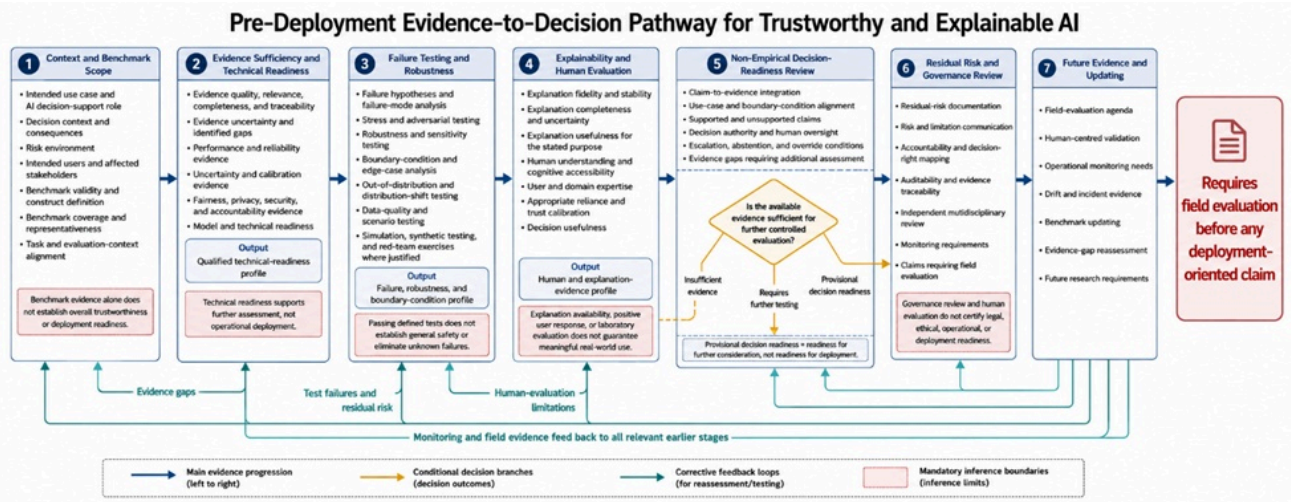


Figure 1. Pre-Deployment Evidence-to-Decision Pathway for Trustworthy and Explainable AI

Alt-text description. A conceptual pathway showing pre-deployment benchmarking evidence moving through validity review, evidence sufficiency, technical readiness, failure testing, human evaluation, non-empirical decision-readiness criteria, residual risk, monitoring needs, and field-evaluation needs.

Table 3 connects human evaluation, explanation usefulness, non-empirical decision-readiness criteria, residual-risk documentation, field-evaluation needs, and pre-deployment decision boundaries.

Table 3. Human Evaluation, Explanation Usefulness, and Non-Empirical Decision-Readiness Criteria for Trustworthy and Explainable AI: Human Understanding, Cognitive Accessibility, User Expertise, Explanation Fidelity, Explanation Stability, Appropriate Reliance, Decision Usefulness, Evidence Sufficiency, Boundary Conditions, Residual Risk, Field Evaluation Needs, and Deployment-Readiness Boundaries

Human, explanation, or readiness criterion	Core evaluation or decision question	Required evidence or condition	Evaluation or readiness function	Potential output	Failure or overclaiming risk	Human-centered or field-evaluation need

Human understanding	Can intended users form an adequately accurate understanding?	Comprehension and mental-model assessment	Evaluates understanding rather than exposure	Understanding profile	Self-reported clarity may not reflect accurate comprehension	Representative-user evaluation
Cognitive accessibility	Can information be interpreted under realistic literacy, time, and workload conditions?	Accessibility and cognitive-load assessment	Tests practical usability	Accessibility profile	Detailed explanations may overwhelm users	Inclusive evaluation
User expertise	How do explanation requirements vary across user roles?	Role- and expertise-sensitive assessment	Prevents one-user-fits-all assumptions	Audience-specific requirements	Expert findings may be generalized to novices	Stratified user evaluation
Domain expertise	Can substantive correctness and relevance be assessed?	Independent domain-expert review	Connects explanation to subject-matter validity	Domain-validity statement	Technically plausible explanations may be substantively misleading	Domain validation
Explanation fidelity	Does the explanation reflect relevant model behaviour?	Model-grounded fidelity assessment	Tests explanatory correctness	Fidelity profile	Persuasive but unfaithful explanations may increase reliance	Technical validation
Explanation stability	Are similar cases explained consistently where expected?	Perturbation and repeatability assessment	Identifies explanation fragility	Stability profile	Stable explanations may still be inaccurate	Joint technical and human review
Explanation uncertainty	Are explanatory limitations communicated?	Uncertainty characterization and communication review	Prevents false explanatory certainty	Uncertainty statement	Explanations may appear more definitive than evidence permits	Comprehension testing
Explanation completeness	Are relevant factors and limitations represented?	Omission and scope analysis	Identifies explanatory gaps	Completeness statement	Concise explanations may omit decisive information	Domain and user review
Explanation usefulness	Does the explanation	Purpose-specific task	Evaluates practical	Usefulness profile	Preference may be	Task-based user evaluation

	support its stated purpose?	evaluation	contribution		mistaken for decision benefit		
Appropriate reliance	Do users accept or reject outputs under appropriate conditions?	Evaluation with reliable, uncertain, and incorrect outputs	Tests calibrated use	Reliance profile	Acceptance may be misclassified as successful trust	Error-sensitive assessment	
Trust calibration	Does trust correspond to demonstrated capability and limitation?	Trust assessment across varied system conditions	Distinguishes calibrated trust from trust maximization	Calibration profile	Interface presentation may inflate trust	Longitudinal evaluation	
Decision usefulness	Does the AI contribution improve the defined decision process?	Decision-task rather than prediction-only assessment	Links technical output to decision value	Decision-support profile	Speed or confidence may be mistaken for decision quality	Outcome-relevant field evaluation	
Use-case alignment	Is the evidence relevant to the intended system role?	Intended-use, excluded-use, and workflow mapping	Establishes scope	Use-case fit statement	Evidence may be transferred to unsupported uses	Domain and governance review	
Evidence sufficiency	Does the evidence address the defined readiness questions?	Claim-evidence mapping and gap analysis	Supports bounded readiness reasoning	Qualified sufficiency narrative	Sufficiency may become an arbitrary score	Multidisciplinary review	
Failure-mode coverage	Have material known failure pathways been examined?	Failure and test-coverage map	Evaluates known-risk coverage	Coverage statement	Unknown failures remain	Lifecycle reassessment	
Boundary-condition clarity	Are unsupported conditions explicitly identified?	Data, model, task, and operational boundary map	Prevents out-of-scope use	Boundary statement	Boundaries may change or be ignored	Monitoring and governance control	
Fairness and accountability evidence	Are relevant disparities and responsibilities addressed?	Contextual fairness assessment and decision-right map	Connects impacts to responsibility	Governance-evidence profile	Formal criteria may hide unexamined harms	Stakeholder and organizational evaluation	
Residual-risk documentation	Which risks remain after	Residual-risk and uncertainty	Prevents zero-risk	Residual-risk statement	Documentation may be	Governance acknowledgment	

	testing and mitigation?	register	conclusions		mistaken for risk removal	
Monitoring requirement	What evidence must be collected if later use is considered?	Drift, incident, performance, and interaction monitoring plan	Connects current gaps to lifecycle learning	Monitoring specification	Monitoring may be treated as a substitute for prior evidence	Operational validation
Field-evaluation need	Which claims cannot be established without real-world evidence?	External-validity gap analysis	Defines future empirical requirements	Field-evaluation agenda	Field testing may be postponed despite unresolved risk	Risk-proportionate field study
Decision-readiness boundary	Is evidence adequate for further consideration rather than deployment?	Integrated evidence and limitation review	Separates readiness levels	Provisional decision-readiness narrative	"Ready for review" may become "ready for operation"	Independent governance process
Deployment-readiness boundary	What prevents the framework from authorizing deployment?	Explicit empirical, legal, ethical, organizational, and operational exclusions	Limits final inference	Non-authorization declaration	Framework completion may resemble certification	Separate deployment decision and field evidence

Proposed Trustworthy AI Benchmarking Framework

The proposed framework begins by defining the use case, decision context, affected stakeholders, risk environment, and intended allocation of human and AI responsibilities. Trustworthy AI is multidimensional and includes technical, ethical, organizational, and human-centered characteristics that cannot be reduced to conventional predictive performance [29]. Context definition therefore precedes benchmark selection. It establishes which claims require evidence, which stakeholders must be considered, which harms are material, and which system uses lie outside the framework's scope.

The second stage evaluates benchmark validity, coverage, representativeness, limitations, and task alignment. The third stage assesses evidence sufficiency and technical readiness through performance, reliability, uncertainty, calibration, fairness, privacy, security, accountability, and explainability evidence where relevant. Trustworthy AI requires attention to robustness, fairness, accountability, explainability, and other properties beyond traditional measures of accuracy [30]. Evidence is retained as a structured profile rather than aggregated into a universal readiness score.

The fourth stage reviews failure testing, robustness, adversarial conditions, data-quality sensitivity, edge cases, distribution shift, out-of-distribution behaviour, and boundary conditions. The fifth stage evaluates explanation fidelity, stability, uncertainty, usefulness, cognitive accessibility, human understanding, appropriate reliance, and trust calibration. Contemporary trustworthy-AI research links principles to technical and lifecycle practices across fairness, privacy, robustness, security, and explainability domains [31]. The framework therefore treats these evidence families as connected but non-interchangeable.

The sixth stage integrates non-empirical decision-readiness criteria. It examines whether the evidence is sufficiently relevant, traceable, complete, and risk-sensitive to justify further controlled evaluation. Holistic trustworthy-AI assessment requires coordinated consideration of reliability, explainability, fairness, accountability, privacy, security, and human acceptance rather than a pass result in one dimension [32]. The output is a qualified narrative identifying supported claims, unsupported claims, unresolved risks, decision boundaries, and the conditions under which additional testing is required.

The final stage documents residual risk, field-evaluation requirements, monitoring needs, governance review, and feedback updating. Explainable-AI research continues to identify unresolved interdisciplinary challenges involving evaluation, human needs, uncertainty, robustness, and societal context [33]. The framework must therefore remain revisable as new evidence emerges. Its explicit ceilings prohibit benchmark-sufficiency claims, technical-readiness sufficiency claims, robustness-safety guarantees, explanation-usefulness guarantees, human-evaluation sufficiency claims, non-empirical deployment guarantees, and deployment-readiness certification.

Table 4 presents the proposed trustworthy AI benchmarking framework linking use-case context, evidence sufficiency, technical readiness, robustness, failure testing, explanation usefulness, human evaluation, decision readiness, residual risk, and future field-evaluation needs.

Table 4. Proposed Trustworthy AI Benchmarking Framework Without Field Experiments: Use-Case Context, Risk Context, Stakeholder Context, Benchmark Validity, Evidence Sufficiency, Technical Readiness, Performance Evidence, Robustness Evidence, Explainability Evidence, Failure Testing, Human Evaluation, Explanation Usefulness, Decision Readiness, Residual Risk, Monitoring Requirements, Field Evaluation Needs, and Deployment Boundaries

Framework stage	Core purpose	Required input	Benchmarking function	Potential output	Validation or governance need	Failure risk	
Use-case and decision-context definition	Define the system's intended decision-support role	Intended use, users, workflow, excluded uses	Anchors evidence to a stated purpose	Context specification	Domain and organizational review	Vague scope enables unsupported inference	R us
Risk and stakeholder mapping	Identify harms, affected interests, and evidence needs	Risk scenarios and stakeholder map	Sets proportionate evidence expectations	Risk–stakeholder profile	Indirectly affected groups may be omitted	Ethics and stakeholder review	c
Benchmark-validity review	Determine whether benchmarks measure relevant constructs	Benchmark documentation and task rationale	Tests construct and task validity	Validity-qualified benchmark set	Independent methodological review	Convenient benchmarks may be accepted without fit	b
Benchmark-coverage review	Identify represented and omitted evidence dimensions	Benchmark inventory and context map	Reveals coverage gaps	Coverage matrix	High benchmark volume may conceal narrow scope	Multidisciplinary review	e

Evidence-sufficiency assessment	Determine whether claims have adequate support	Claim-evidence map and gap register	Integrates evidence without collapsing dimensions	Qualified sufficiency narrative	Sufficiency may become a rigid score	Expert and governance review	
Technical-readiness assessment	Characterize model-level maturity	Performance, reliability, uncertainty, and documentation	Supports an intermediate readiness judgment	Technical-readiness profile	Model maturity may be equated with operational safety	Independent validation	R
Performance and reliability review	Establish bounded capability and consistency	Task, subgroup, repetition, and sensitivity evidence	Reviews model behaviour under stated conditions	Capability profile	Aggregate results may conceal scenario failures	Methodological review	co s
Uncertainty and calibration review	Qualify confidence and unknowns	Uncertainty-source and calibration evidence	Evaluates confidence interpretation	Uncertainty profile	Poor uncertainty estimates may create false reassurance	Statistical and domain review	u
Fairness, accountability, and governance review	Examine disparities, responsibility, and contestability	Fairness evidence and decision-right map	Connects evaluation to affected interests	Governance-evidence profile	Formal criteria may substitute for substantive review	Stakeholder and organizational evaluation	de res
Explainability-evidence review	Evaluate explanations by property, audience, and purpose	Fidelity, stability, usefulness, and accessibility evidence	Separates explanatory dimensions	Explanation-evidence profile	Persuasive explanations may be unfaithful	Technical and human-centered review	Re
Failure-testing and robustness review	Examine known failure hypotheses and perturbations	Test portfolio, threat model, and scenarios	Assesses failure coverage	Failure and robustness profile	Passing narrow tests may imply general safety	Independent challenge	Ad ne
Boundary and out-of-distribution review	Identify unsupported regions and shifts	Boundary, edge-case, and shift evidence	Defines where inference should stop	Boundary map	Boundaries may remain incomplete	Monitoring and domain review	U ne
Human and explanation-usefulness review	Test interaction and comprehension assumptions	Relevant-user or justified-proxy evidence	Evaluates understanding and reliance	Human-evidence profile	Laboratory acceptance may be generalized	Human-factors and field review	rep

Non-empirical decision-readiness review	Integrate available evidence for bounded consideration	All evidence profiles and gaps	Supports structured readiness reasoning	Provisional decision-readiness narrative	Narrative may resemble certification	Multidisciplinary governance panel	
Residual-risk documentation	Preserve unresolved risks and unknowns	Failure findings, gaps, and uncertainty	Prevents zero-risk claims	Residual-risk register	Risks may be documented but not controlled	Governance acknowledgment	U
Field-evaluation need identification	Specify claims requiring real-world evidence	External-validity gaps	Defines future empirical requirements	Field-evaluation agenda	Required field evidence may be deferred	Independent study-design review	L
Monitoring-requirement definition	Specify future lifecycle evidence needs	Anticipated drift, incidents, and interaction risks	Connects readiness to continued learning	Monitoring specification	Monitoring may create false reassurance	Operational governance	re
Governance review and feedback updating	Maintain accountability and framework revision	Complete evidence and limitation record	Supports challenge, revision, and traceability	Reviewed framework dossier	Review may become procedural	Feedback to every earlier stage	G
Deployment-boundary declaration	State the framework's final inference limit	All unresolved empirical and contextual requirements	Prevents assessment from becoming authorization	Explicit non-deployment statement	Framework completion may resemble certification	Separate legal, ethical, operational, and field processes	M

Figure 2 presents the proposed trustworthy AI benchmarking framework, integrating benchmark validity, evidence sufficiency, technical readiness, failure testing, robustness, explanation usefulness, human evaluation, decision-readiness criteria, residual-risk documentation, monitoring requirements, and field-evaluation needs.

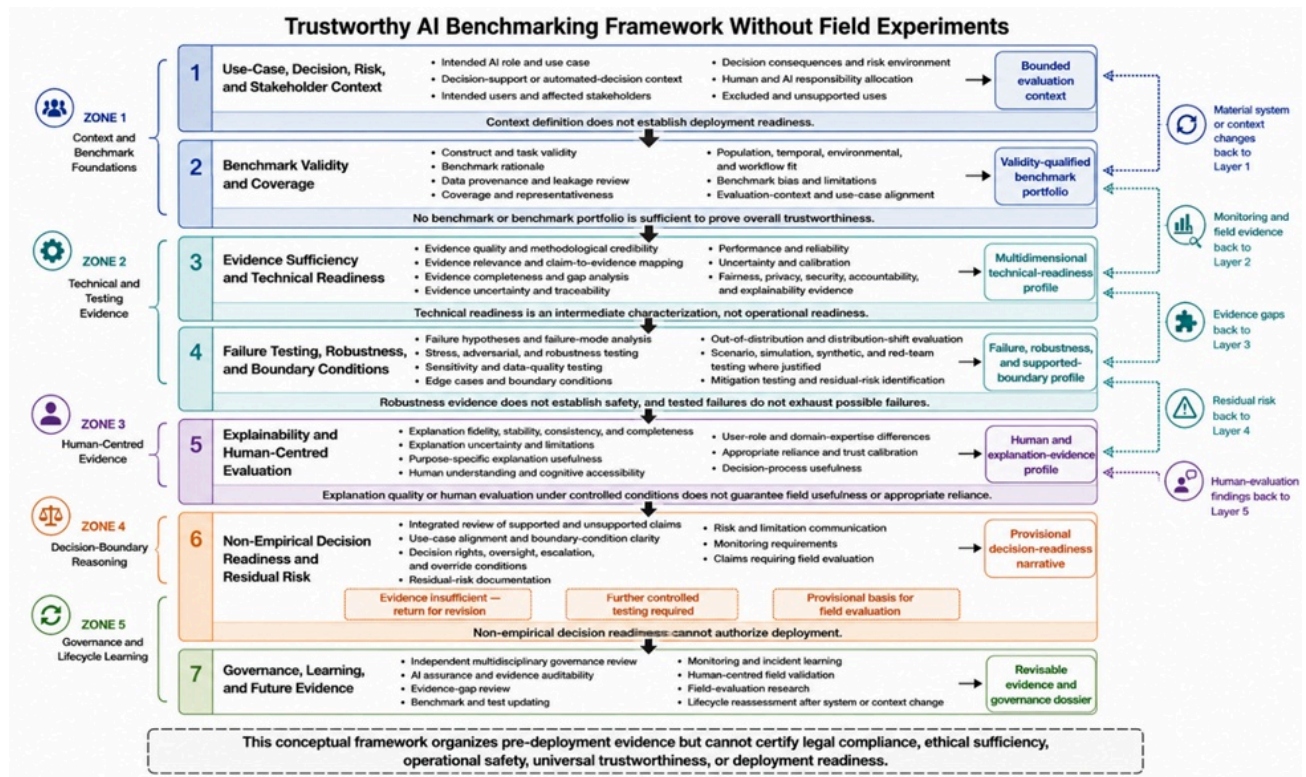


Figure 2. Trustworthy AI Benchmarking Framework Without Field Experiments

Alt-text description. A conceptual trustworthy AI benchmarking framework showing use-case context, benchmark validity, evidence sufficiency, technical readiness, robustness testing, failure testing, explanation usefulness, human evaluation, decision readiness, residual risk, monitoring requirements, governance review, and field-evaluation needs.

Practical Implications and Future Research

The framework implies that benchmark design should be studied as a validity and governance problem rather than only as a technical comparison exercise. Future research should examine how construct validity, task alignment, data provenance, representativeness, and leakage influence the strength of trustworthy-AI claims. Reproducibility failures show why apparently competitive performance can remain evidentially weak when evaluation design is compromised [5]. Research should also investigate how benchmark choices embed assumptions about relevant populations, harms, workflows, and decision objectives.

Evidence-sufficiency research should develop transparent ways of reasoning about coverage without creating universal numerical thresholds. Dataset documentation and lifecycle assurance provide foundations for traceability, but further work is required to connect documented evidence to context-specific readiness claims [6]. Technical-readiness studies should clarify how performance, reliability, uncertainty, fairness, privacy, security, accountability, and explainability evidence interact while remaining distinct. Domain-specific case analysis could test whether the proposed framework helps identify missing evidence without being converted into a certification checklist.

Failure-testing research should connect test-generation methods more directly to realistic hazards, stakeholder concerns, and operating boundaries. The existing testing landscape demonstrates substantial methodological diversity but incomplete coverage of complex AI-system behaviour [14]. Research priorities include out-of-distribution evaluation, distribution-shift analysis, realistic stress conditions, edge-case construction, red-team methods, residual-risk

communication, and mechanisms for updating tests after material system or context changes. Comparative field research will be needed to determine which pre-deployment tests are informative about later operational failures.

Human-centered research should examine explanation usefulness, calibrated reliance, cognitive accessibility, and decision validity across representative users, roles, and organizational settings. It should also investigate how monitoring requirements, governance review, and field-evaluation plans can be incorporated without creating false reassurance. These implications are conceptual research directions rather than legal advice, compliance guidance, audit instruction, procurement advice, benchmark- or metric-selection advice, model- or platform-selection advice, safety certification, field-deployment instruction, or deployment authorization.

Conclusion

This article proposed a conceptual trustworthy-AI benchmarking framework for reasoning about evidence sufficiency before deployment when field experiments and operational outcome evidence are not yet available. The framework separates benchmark validity and coverage from technical readiness, distinguishes robustness and failure-testing evidence from operational safety, and evaluates explanation usefulness independently from explanation availability. It further integrates human evaluation, appropriate reliance, accountability, residual-risk documentation, monitoring requirements, and field-evaluation needs into a structured evidence-to-decision pathway. Its central contribution is not a readiness score or approval mechanism, but a disciplined way to state what available evidence supports, what it does not support, and which claims remain dependent on empirical field evaluation. Trustworthy and explainable AI cannot be established through benchmark performance, laboratory testing, or explanation metrics alone. Pre-deployment benchmarking can organize evidence and expose gaps, but it cannot certify legal compliance, ethical sufficiency, operational safety, universal trustworthiness, or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Aug 2025

Revised: 03 Oct 2025

Accepted: 11 Dec 2025

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

- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-99.
<https://doi.org/10.1038/s42256-019-0088-2>.
- Floridi L, Cows J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach*. 2018;28(4):689-707.
<https://doi.org/10.1007/s11023-018-9482-5>.
- 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>.
- 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>.
- Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)*. 2023;4(9):100804.
<https://doi.org/10.1016/j.patter.2023.100804>.
- Gebri T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H 3rd, et al. Datasheets for datasets. *Commun ACM*. 2021;64(12):86-92.
<https://doi.org/10.1145/3458723>.
- Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):1-35.
<https://doi.org/10.1145/3457607>.
- Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: desiderata, methods, and challenges. *ACM Comput Surv*. 2021;54(5):1-39.
<https://doi.org/10.1145/3453444>.
- Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics*. 2020;26(4):2141-68.
<https://doi.org/10.1007/s11948-019-00165-5>.
- Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach*. 2020;30(1):99-120.
<https://doi.org/10.1007/s11023-020-09517-8>.
- 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>.
- 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>.

Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep neural networks. *Artif Intell Rev.* 2023;56(Suppl 1):1513-89.
<https://doi.org/10.1007/s10462-023-10562-9>.

Zhang JM, Harman M, Ma L, Liu Y. Machine learning testing: survey, landscapes and horizons. *IEEE Trans Softw Eng.* 2022;48(1):1-36.
<https://doi.org/10.1109/TSE.2019.2962027>.

Riccio V, Jahangirova G, Stocco A, Humbatova N, Weiss M, Tonella P. Testing machine learning based systems: a systematic mapping. *Empir Softw Eng.* 2020;25(6):5193-254.
<https://doi.org/10.1007/s10664-020-09881-0>.

Huang X, Kroening D, Ruan W, Sharp J, Sun Y, Thamo E, et al. A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability. *Comput Sci Rev.* 2020;37:100270.
<https://doi.org/10.1016/j.cosrev.2020.100270>.

Ben Braiek H, Khomh F. On testing machine learning programs. *J Syst Softw.* 2020;164:110542.
<https://doi.org/10.1016/j.jss.2020.110542>.

Tambon F, Laberge G, An L, Nikanjam A, Nouwou Mindom PS, Pequignot Y, et al. How to certify machine learning based safety-critical systems? A systematic literature review. *Autom Softw Eng.* 2022;29(2):38.
<https://doi.org/10.1007/s10515-022-00337-x>.

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>.

Páez A. The pragmatic turn in explainable artificial intelligence. *Minds Mach.* 2019;29(3):441-59.
<https://doi.org/10.1007/s11023-019-09502-w>.

Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst.* 2021;11(3-4):1-45.
<https://doi.org/10.1145/3387166>.

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>.

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>.

Shneiderman B. Human-centered artificial intelligence: reliable, safe and trustworthy. *Int J Hum Comput Interact.* 2020;36(6):495-504.
<https://doi.org/10.1080/10447318.2020.1741118>.

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

Jarrahi MH. Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Bus Horiz.* 2018;61(4):577-86.
<https://doi.org/10.1016/j.bushor.2018.03.007>.

Shrestha YR, Ben-Menahem SM, von Krogh G. Organizational decision-making structures in the age of artificial intelligence. *Calif Manage Rev*. 2019;61(4):66-83.

<https://doi.org/10.1177/0008125619862257>.

Raisch S, Krakowski S. Artificial intelligence and management: the automation-augmentation paradox. *Acad Manag Rev*. 2021;46(1):192-210.

<https://doi.org/10.5465/amr.2018.0072>.

Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. *Electron Mark*. 2021;31(2):447-64.

<https://doi.org/10.1007/s12525-020-00441-4>.

Wing JM. Trustworthy AI. *Commun ACM*. 2021;64(10):64-71.

<https://doi.org/10.1145/3448248>.

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>.

Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. *ACM Comput Surv*. 2022;55(2):1-38.

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

Longo L, Brčić M, Cabitza F, Choi J, Confalonieri R, Del Ser J, et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf Fusion*. 2024;106:102301.

<https://doi.org/10.1016/j.inffus.2024.102301>.