

REVIEW

Open access

What Makes Artificial Intelligence Worthy of Trust? A Systematic Review of Transparency, Robustness, Fairness, Privacy, Accountability, and Human Oversight

Elena Petrova¹, Marcus Weber^{2*}, Priya Nair³

Abstract

Artificial intelligence increasingly informs consequential decisions, yet high predictive performance alone does not establish that an intelligent system is worthy of trust. This systematic review examines trustworthy AI as a multidimensional, socio-technical, and lifecycle-dependent research domain. It synthesises selected peer-reviewed journal evidence addressing transparency, interpretability, explainability, robustness, safety, privacy, security, fairness, accountability, governance, and meaningful human oversight. The review applies predefined questions, eligibility principles, trustworthiness-dimension mapping, claim-level evidence extraction, and cross-dimensional synthesis without undertaking meta-analysis or inventing quantitative screening results. The evidence indicates that transparency and explanation must be assessed in relation to fidelity, stability, completeness, usefulness, stakeholder needs, contestability, recourse, and the practical authority of human overseers. Robustness and safety require evaluation under adversarial conditions, uncertainty, context shift, system failure, and residual risk, while privacy-preserving approaches must still be examined for leakage, cybersecurity vulnerabilities, and governance limitations. Fairness depends on context-sensitive definitions, bias identification, distributional consequences, documentation, responsibility allocation, auditability, and institutional mechanisms for correction. Cross-dimensional analysis reveals persistent tensions between transparency and privacy, explainability and security, automation and human control, and performance and fairness. Important gaps remain in multidimensional evaluation, realistic human-oversight studies, explanation-usefulness validation, lifecycle monitoring, accountability operationalisation, stakeholder participation, and foundation-model assurance. The review contributes an integrated evidence synthesis showing that AI worthiness of trust should be supported through coordinated technical validation, human-centred assessment, institutional governance, and continuing monitoring rather than inferred from any isolated property.

Keywords Explainable AI, Trustworthy AI, Fairness, Transparency, Robustness, Privacy

*Correspondence:

Marcus Weber

marcus.weber@lmu.de

¹ Department of Trustworthy AI and Governance, Faculty of Computer Science, Sofia University, Bulgaria

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

Introduction

Trustworthy artificial intelligence has become a central concern because intelligent systems increasingly influence access to information, services, employment, healthcare, finance, public administration, and other consequential activities. The

literature nevertheless describes trustworthiness through partially overlapping ethical principles, technical properties, institutional duties, and human-centred requirements. These accounts collectively indicate that trustworthy AI is a multidimensional socio-technical construct rather than a synonym for accuracy, explainability, or rule compliance [1-3].

Trust must also be distinguished from trustworthiness. Trust concerns a person's willingness to accept vulnerability or rely on an intelligent system, whereas trustworthiness concerns whether that reliance is justified by relevant evidence. Empirical research shows that human trust can be shaped by performance, appearance, transparency, task characteristics, and interaction context, but favourable attitudes do not prove that a system is robust, fair, private, accountable, or safe [4]. Excessive trust may therefore be as problematic as unjustified distrust.

The need for systematic synthesis arises because major trustworthiness dimensions are commonly examined in separate research traditions. Explainable-AI studies investigate how model behaviour can be communicated, adversarial-machine-learning research examines vulnerability to strategic manipulation, privacy research studies information exposure, fairness research analyses distributional harms, and governance scholarship considers responsibility and institutional control. Evaluating these dimensions independently can obscure their dependencies, conflicts, and shared lifecycle requirements.

This review asks how trustworthy AI is defined; how transparency, explainability, and human oversight are conceptualised; how robustness, safety, privacy, and security are operationalised; how fairness, accountability, and governance are connected; which tensions occur among dimensions; and which research gaps remain. Its contribution is a structured systematic synthesis that links technical evidence, human-centred evaluation, and institutional governance while maintaining clear boundaries against certification, legal-compliance, universal-metric, and deployment-readiness claims.

Trustworthy AI as a Multidimensional Research Domain

Trustworthiness is broader than perceived trust and more demanding than satisfactory model performance. Reliability may provide reasons for calibrated reliance, but it does not independently establish ethical acceptability, institutional accountability, privacy protection, or safety across changing contexts [5]. Trustworthy AI should therefore be treated as an evidence-based assessment of whether a system possesses appropriate properties, operates within justified limits, and remains subject to effective human and organisational control.

Human-centred AI adds an important design perspective by rejecting the assumption that increasing automation necessarily requires diminishing human authority. High levels of computational capability can coexist with strong human control when systems support inspection, correction, escalation, override, and responsibility rather than merely inserting a nominal person into an automated workflow [6]. Human-in-the-loop, human-on-the-loop, and human-control arrangements must consequently be distinguished according to when intervention occurs, what information is available, and whether the person has practical authority.

Responsible AI and ethical AI concern the principles, practices, and institutional arrangements through which AI-related values and risks are addressed. Explainable, interpretable, robust, safe, secure, privacy-preserving, fair, accountable, and auditable AI describe more specific properties or governance functions. Evidence connecting principles, technical requirements, organisational processes, and regulation shows that these concepts overlap but are not interchangeable [7]. Governance determines who sets objectives, approves risk, monitors performance, records changes, investigates incidents, and provides remedy.

A multidimensional perspective also prevents single-property sufficiency claims. Explainability does not eliminate bias, robustness does not prove ethical acceptability, privacy protection does not establish security, fairness metrics do not create accountability, and human oversight does not guarantee effective intervention. Technical surveys of trustworthy AI similarly demonstrate that distinct dimensions require different methods, assumptions, stakeholders, and evaluation

criteria [8]. Trustworthiness is therefore better represented as a structured portfolio of bounded evidence maintained across the AI lifecycle.

Review Questions, Search Strategy, and Study Selection

The review was organised around six questions: how trustworthy AI is defined; how transparency, explainability, and human oversight are conceptualised; how robustness, safety, privacy, and security are evaluated; how fairness, accountability, and governance are connected; which dependencies and trade-offs occur across dimensions; and which research gaps remain. These questions guided search-concept design, eligibility assessment, evidence extraction, synthesis, and citation planning.

Search concepts combined trustworthy AI, responsible AI, ethical AI, explainable AI, interpretability, transparency, human oversight, robustness, safety, uncertainty, privacy, security, fairness, accountability, auditability, governance, assurance, and intelligent systems. Earlier XAI surveys were used to identify broad method families, while later conceptual reviews helped distinguish explanation types, responsible-AI connections, and evaluation approaches [9-11]. Search terms were expanded iteratively when synonymous or adjacent concepts were identified.

Eligibility required a peer-reviewed journal article with a verified DOI, relevant journal standing, and substantive evidence for at least one review question. Records were excluded when they were non-peer-reviewed, conference-only, preprint-only, outside the defined scholarly scope, focused exclusively on predictive performance, or unable to support a planned claim. Explanation evaluation was extracted separately from explanation generation because quantitative XAI research distinguishes properties such as fidelity, stability, completeness, consistency, and human usefulness [12].

Evidence extraction recorded each article’s aim, intelligent-system context, trustworthiness dimensions, technical or socio-technical functions, evaluation approach, principal limitation, synthesis role, and first-use citation position. Selection was based on relevance and evidentiary fit rather than invented search totals, screening counts, quality scores, or meta-analytic outcomes. **Table 1** summarises the systematic review logic used to identify, screen, select, and extract evidence on trustworthy and explainable intelligent systems. **Figure 1** illustrates the systematic review workflow used to move from review questions and search concepts to study screening, evidence extraction, trustworthiness-dimension mapping, synthesis, and research-gap identification.

Table 1. Systematic Review Logic for Trustworthy and Explainable Intelligent Systems: Review Questions, Search Concepts, Inclusion Criteria, Exclusion Criteria, Screening Logic, Evidence Extraction Categories, Trustworthiness Dimension Mapping, Citation Planning, and Review Boundaries

Review component	Core review question or purpose	Inclusion logic	Exclusion logic	Evidence extracted	Trustworthiness dimension addressed	Bias or limitation risk
Review questions	Define the concepts, operationalisations, relationships, and gaps examined	Direct relevance to at least one review question	No identifiable contribution to a question	Definitions, mechanisms, evaluations, tensions, gaps	Cross-cutting	Questions may privilege established terminology
Search concepts	Capture technical, human-centred, ethical, and governance scholarship	Trustworthy AI, XAI, robustness, privacy, fairness,	Performance-only or unrelated AI research	Keywords, concepts, intelligent-system context	All dimensions	Terminology and database bias

		accountability, oversight, or related terms				
Source eligibility	Restrict evidence to appropriate scholarly publications	Peer-reviewed journal article, DOI, relevant journal standing	Preprint, conference paper, book, website, report, standard, or policy document	Publication type, journal, DOI	Cross-cutting	Ranking categories may vary
Relevance screening	Confirm substantive trustworthiness content	Supports a definition, method, evaluation, limitation, trade-off, or research gap	Superficial mention without usable evidence	Abstract and full-text relevance	Dimension-specific	Reviewer interpretation
Evidence extraction	Capture comparable article-level information	Clearly reported aim, feature, evaluation, limitation, or implication	Insufficient detail for a planned claim	Context, dimension, function, evaluation, limitation	Technical and socio-technical	Selective extraction
Dimension mapping	Connect each source to one or more review domains	Clear relationship to transparency, oversight, robustness, privacy, fairness, governance, or synthesis	Cannot be mapped reliably	Primary and secondary dimensions	All dimensions	Overrepresentation of mature fields
Citation planning	Preserve claim-level support and first-appearance order	Precise claim–source fit and approved citation position	Decorative, random, or redundant citation	First-use section, paragraph, claim, citation form	Cross-cutting	Citation convenience
Review boundaries	Prevent overclaiming	Evidence supports synthesis and future research design	Claims of certification, legal compliance, universal scoring, or	Assumptions and residual limitations	No-guarantee boundaries	Readers may overgeneralise

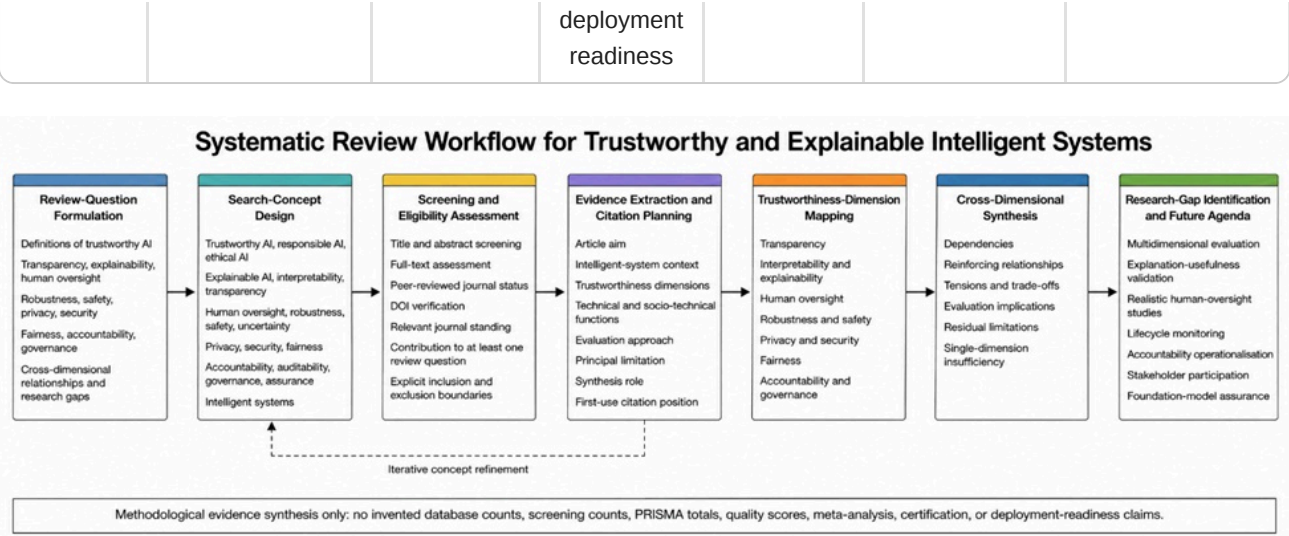


Figure 1. Systematic Review Workflow for Trustworthy and Explainable Intelligent Systems

Alt-text description. A systematic review workflow diagram showing review questions, search concepts, peer-reviewed journal screening, eligibility checks, evidence extraction, trustworthiness-dimension mapping, cross-dimensional synthesis, and research-gap identification.

Transparency, Explainability, and Human Oversight

Transparency concerns the availability of relevant information about system purpose, data, modelling choices, limitations, deployment context, and responsible actors. Interpretability concerns whether model logic can be understood directly, while explainability concerns representations intended to clarify outputs or behaviour. Black-box explanation methods include feature-attribution, example-based, counterfactual, rule-based, local, and global approaches, each of which differs in model access, scope, format, and fidelity assumptions [13].

Explanation quality cannot be reduced to visual clarity or user satisfaction. Explanations may need to be faithful to model behaviour, stable under immaterial changes, sufficiently complete for their stated purpose, and useful to a particular stakeholder. Social-science evidence indicates that effective explanations are often selective, contrastive, audience-dependent, and situated within dialogue rather than delivered as exhaustive technical accounts [14]. Causal wording should therefore be reserved for explanations supported by defensible causal assumptions.

The distinction between intrinsic interpretability and post-hoc explanation is especially important in consequential settings. A persuasive explanation of an opaque model may create an impression of understanding without accurately representing the mechanism that produced the decision. Where comparably effective interpretable models are feasible, post-hoc explanation should not automatically be treated as an adequate substitute [15]. Explainability must also be evaluated against privacy, security, fairness, and safety requirements rather than presented as an independent trust mechanism.

Human-centred explanation requires alignment among stakeholder knowledge, decision responsibilities, explanation purpose, and evaluation method. Developers may need diagnostic explanations, auditors may require traceable evidence, decision makers may need uncertainty and limitation information, and affected individuals may need understandable reasons, contestability, and feasible recourse. Stakeholder-oriented XAI frameworks consequently support evaluation designs that connect explanation form to real tasks rather than assuming that one interface can satisfy all users [16].

Human-in-the-loop oversight involves direct participation in selected decisions or learning stages, whereas human-on-the-loop oversight involves supervision of substantially automated processes with authority to intervene. Both can fail through automation bias, excessive workload, inadequate information, delayed intervention, or nominal authority. User-study evidence therefore supports measuring error detection, decision quality, calibrated reliance, intervention success, and practical understanding instead of perceived clarity alone [17]. **Table 2** maps transparency, explainability, interpretability, explanation quality, and human oversight requirements across trustworthy intelligent systems.

Table 2. Transparency, Explainability, Interpretability, Explanation Quality, and Human Oversight in Trustworthy Intelligent Systems: Explanation Usefulness, Fidelity, Stability, Completeness, Counterfactuals, Feature Attribution, Human-in-the-Loop Oversight, Human-on-the-Loop Oversight, Contestability, Recourse, and Explainability Boundaries

Dimension or requirement	Core trustworthiness question	Technical or socio-technical function	Potential evaluation approach	Primary stakeholder	Failure or misuse risk	Human oversight n
Transparency	Is relevant information about purpose, data, limits, and responsibility available?	Supports scrutiny and informed use	Documentation and comprehension review	Users, auditors, deployers	Information dumping or strategic opacity	Confirm th disclosur enables act
Interpretability	Can operative logic be understood directly?	Enables inspection without a surrogate	Simulatability, complexity, rule inspection	Experts and decision makers	Apparent simplicity may conceal system complexity	Verify that t inspected representatio operative
Explainability	Can a system provide intelligible reasons for an output or behaviour?	Communicates selected aspects of model operation	Fidelity, usefulness, stability, and user testing	Users, experts, affected people	Plausible but misleading explanations	Match explanation task and audience
Explanation usefulness	Does explanation improve appropriate understanding or action?	Connects explanation to human outcomes	Decision quality, error detection, calibrated reliance	Decision makers and affected users	Confidence may increase without accuracy	Observe downstrea human performan
Explanation fidelity	Does explanation reflect relevant model behaviour?	Limits divergence from the operative model	Perturbation tests and surrogate agreement	Auditors and developers	Rationalisation of opaque behaviour	Restrict relia when fidelit unknown
Explanation stability and completeness	Is explanation consistent and sufficiently	Supports reproducibility and scope awareness	Sensitivity, consistency, coverage, residual analysis	Evaluators and users	Arbitrary changes or partial accounts	Investigat unstable o incomplet outputs

	comprehensive for its purpose?					
Feature attribution	Which inputs appear to influence the output?	Provides local or global contribution estimates	Sensitivity, baseline, and sanity checks	Experts and users	Association may be presented as causation	Domain review of actionability
Counterfactual explanation and recourse	What feasible changes could produce another outcome?	Supports decision-boundary understanding and action	Validity, proximity, feasibility, diversity, stability	Affected individuals	Impossible, inequitable, or privacy-invasive advice	Review feasibility and downstream consequences
Example- and rule-based explanation	Can representative cases or understandable rules clarify behaviour?	Provides prototypes, influential examples, or decision structures	Representativeness, privacy, rule fidelity, and coverage	Experts, users, auditors	Stereotyping, leakage, or oversimplification	Validate examples and rule scope
Causal explanation	Are claimed causes supported rather than inferred from association?	Supports intervention-oriented reasoning	Assumption review and intervention testing	Domain experts	Unsupported causal claims may cause harm	Require domain and causal validation
Human-in-the-loop oversight	Does a person participate meaningfully in a decision stage?	Enables review, correction, escalation, or approval	Override quality, workload, and disagreement analysis	Operators and professionals	Rubber-stamping and automation bias	Provide authority, competence, time, and evidence
Human-on-the-loop oversight and control	Can a supervisor observe and intervene effectively?	Supports monitoring, exception handling, and shutdown	Detection latency and intervention success	Supervisors and safety operators	Poor awareness or intervention too late	Ensure observability and reliable control path
Contestability	Can an affected decision be challenged effectively?	Supports correction and answerability	Appeal accessibility and resolution quality	Affected individuals	Symbolic or inaccessible appeal	Human review must have corrective authority
No explainability-sufficiency boundary	What does explanation fail to establish independently?	Prevents explanation from becoming a trust certificate	Cross-check with safety, fairness, privacy, and governance evidence	All stakeholders	Explainability theatre and unjustified reliance	Require multidimensional assessment

Robustness, Safety, Privacy, and Security

Robustness concerns whether an intelligent system maintains acceptable behaviour under plausible variation, perturbation, and stress. It is distinct from reliability, which concerns consistency under intended conditions, and resilience, which includes detection, recovery, and adaptation after disruption. Adversarial-machine-learning research demonstrates that satisfactory ordinary test performance does not establish resistance to deliberately manipulated inputs, poisoned training data, model extraction, or adaptive attacks [18].

Safety requires a broader analysis of hazards, consequences, operating assumptions, foreseeable misuse, and residual risk. Verification may examine whether specified properties hold, while validation asks whether the system is appropriate for its intended context. Testing, adversarial assessment, interpretability analysis, fault injection, and scenario evaluation can provide complementary evidence, but none independently proves that a deep-learning system will remain harmless under all real-world conditions [19].

Uncertainty is also central to dependable decision support. Aleatoric uncertainty reflects variability that may persist even with additional data, whereas epistemic uncertainty concerns limitations in knowledge or model representation that may sometimes be reduced. These forms should not be treated interchangeably because they can justify different actions, including collecting evidence, abstaining, escalating to a human reviewer, or communicating irreducible uncertainty [20]. Calibration must additionally be assessed under relevant populations and changing conditions.

Privacy-preserving learning aims to reduce unnecessary exposure of personal or sensitive information, but architectural labels should not be interpreted as privacy guarantees. Federated learning can reduce centralised movement of raw data while introducing challenges involving heterogeneous participants, communication, reliability, and distributed control [21]. It can also remain vulnerable to inference, poisoning, backdoor, reconstruction, and communication attacks, meaning that decentralised training does not eliminate cybersecurity or privacy-leakage risks [22].

Security extends beyond model robustness to infrastructure, interfaces, credentials, data pipelines, software dependencies, logs, and organisational access practices. Domain-specific reviews show that technical defences must be interpreted in relation to the consequences of error, available expertise, data constraints, and continuing human oversight [23]. Trustworthiness assessment should therefore document threat models, protection assumptions, failure responses, monitoring arrangements, and residual risks without presenting finite testing as a safety or privacy certificate.

Fairness, Accountability, and Governance

Fairness concerns how benefits, burdens, errors, opportunities, and representations are distributed across affected populations. Bias may arise through data collection, sampling, measurement, labels, feature construction, model objectives, evaluation procedures, or deployment practices. It may produce allocation harms, such as unjustified denial of opportunities, or representational harms, such as stereotyping, exclusion, and systematic misrepresentation [24].

Formal fairness measures operationalise different normative ideas and may yield incompatible assessments of the same system. Equality of error rates, calibration, demographic parity, individual similarity, and other criteria rely on different assumptions about groups, outcomes, and acceptable trade-offs. Selecting a metric is therefore not a neutral technical step, and satisfying one criterion does not demonstrate that a decision process is substantively fair [25].

Accountability asks who must explain, justify, correct, and answer for decisions involving AI. Responsibility can be distributed among data creators, developers, vendors, procurers, deployers, operators, managers, and public institutions, but distribution must not become diffusion. Algorithmic systems embed choices concerning objectives, categories, thresholds, evidence, and acceptable error, making those organisational decisions relevant to accountability even when model outputs are technically reproducible [26].

Governance provides the institutional structures through which objectives, authority, risk acceptance, monitoring, escalation, and remedy are organised. General principles can orient action but cannot ensure ethical practice when

incentives, responsibilities, implementation mechanisms, and enforcement remain unspecified [27]. Operational tools—including impact assessments, documentation processes, audit procedures, risk registers, review gates, and stakeholder engagement—may translate principles into practice, although their effectiveness requires empirical evaluation [28].

Traceability and auditability depend on accurate, current, and accessible records of data provenance, model development, intended use, limitations, changes, incidents, and responsible actors. Dataset documentation can support deliberation about composition, collection, processing, appropriate use, distribution, and maintenance, but completed documentation does not itself validate fairness, privacy, or safety [29]. **Table 3** synthesises robustness, safety, privacy, security, fairness, accountability, and governance requirements for trustworthy AI systems.

Table 3. Robustness, Safety, Privacy, Security, Fairness, Accountability, and Governance Requirements for Trustworthy AI: Reliability, Resilience, Adversarial Robustness, Uncertainty, Calibration, Privacy Protection, Cybersecurity, Bias, Discrimination, Traceability, Auditability, Documentation, Lifecycle Governance, and Compliance Boundaries

Trustworthiness dimension	Core evaluation or governance question	Technical or institutional function	Potential evaluation approach	Failure, harm, or attack risk	Residual limitation	Monitoring or governance need	F
Robustness	Does behaviour remain acceptable under plausible variation?	Tests sensitivity beyond nominal conditions	Perturbation, stress, subgroup, and context-shift testing	Failure under changed inputs or environments	Tests cannot cover every future condition	Drift monitoring and revalidation	Se
Reliability and resilience	Is operation consistent, recoverable, and maintainable?	Supports continuity and recovery	Repeated testing, fault injection, recovery analysis	Intermittent failure or cascading disruption	Recovery mechanisms may fail under correlated events	Maintenance, fallback, and incident review	te
Safety	Are unacceptable harms identified and controlled?	Links hazards to safeguards and operating limits	Hazard analysis, scenario testing, assurance arguments	Physical, social, economic, or institutional harm	Unknown hazards and misuse remain	Lifecycle review and escalation	Es
Adversarial robustness	Can adaptive manipulation be resisted?	Protects model integrity	Evasion, poisoning, extraction, and red-team testing	Backdoors, adversarial inputs, and model theft	Threat models become outdated	Repeated adversarial assessment	Dis
Uncertainty and calibration	Are confidence and knowledge limits	Supports abstention and risk-sensitive decisions	Calibration, coverage, subgroup, and shift analysis	False certainty or unusable uncertainty	Estimates may themselves be misspecified	Recalibration and escalation thresholds	Lin

	represented appropriately?						
Privacy protection	Are sensitive data protected throughout the lifecycle?	Limits unauthorised exposure and secondary use	Privacy-impact review, leakage attacks, access audit	Re-identification, inference, and misuse	Composition and auxiliary data may weaken protection	Data minimisation and incident response	Es t p gu bo
Security and cybersecurity	Are models, infrastructure, interfaces, and dependencies protected?	Preserves confidentiality, integrity, and availability	Penetration testing, dependency audit, red teaming	Tampering, theft, disruption, and compromised pipelines	New vulnerabilities remain possible	Continuous vulnerability management	E ev be
Fairness and bias	Are errors, opportunities, and harms distributed justifiably?	Shapes data, objectives, metrics, and remedies	Subgroup, intersectional, process, and impact analysis	Discrimination, exclusion, and unequal burden	Formal metrics cannot resolve every normative dispute	Stakeholder review and documented rationale	Es t fa su bo
Accountability and responsibility	Who must justify, correct, and answer for outcomes?	Assigns duties and corrective authority	Responsibility mapping, decision logs, remedy review	Responsibility gaps and blame shifting	Complex supply chains obscure control	Named owners and escalation routes	C te ch ans
Traceability and auditability	Can evidence and decisions be reconstructed and inspected?	Supports investigation and independent review	Provenance, version, access, and audit checks	Missing lineage or performative audits	Records may be incomplete or inaccessible	Independent access and remediation authority	C cla rev e
Documentation	Are data, models, intended uses, limits, and changes recorded?	Preserves context across the lifecycle	Completeness, accuracy, and update assessment	Stale records and inappropriate reuse	Documentation depends on truthful maintenance	Ownership and change triggers	Ne ins as infra
Lifecycle governance	Are controls maintained from formulation to retirement?	Coordinates technical and institutional risk management	Stage gates, change control, monitoring, incident review	Drift, repurposing, and control decay	Governance maturity varies by organisation	Continuing reassessment	P o a be pe as
Compliance boundary	What cannot be established	Separates evidence synthesis	Jurisdiction-specific assessment by	False claims of regulatory approval	Legal requirements	Documented legal and	P co ove

	by technical review alone?	from legal determination	qualified actors		vary by context and time	institutional review	
--	----------------------------	--------------------------	------------------	--	--------------------------	----------------------	--

Cross-Dimensional Evidence Synthesis

The evidence supports a view of trustworthy AI as an interdependent portfolio of technical, human-centred, and institutional evidence. Multidimensional reviews identify distinct requirements, lifecycle-assurance research connects claims to evidence across development and operation, and ethics-based auditing provides an institutional mechanism for examining whether principles are reflected in practice [30-32]. These approaches are complementary, but they do not create a universal trust score or automatic certification.

Governance and technical validation address different questions. Technical evaluation examines system behaviour under specified assumptions, whereas governance determines who defines acceptable performance, approves residual risk, oversees changes, investigates incidents, and provides remedy. Governance frameworks can connect these responsibilities across organisations and institutions, but their use cannot by itself establish jurisdiction-specific legal compliance [33].

Human trust adds another layer because confidence, acceptance, and reliance emerge from interactions among system performance, explanations, institutional signals, user experience, and context. The literature remains fragmented across technical, psychological, ethical, and governance perspectives, limiting direct comparison among trust measures and trustworthiness evidence [34]. High reported trust may therefore reflect persuasive design or limited understanding rather than justified reliance.

Several relationships are reinforcing. Robustness evidence can support safety arguments when tests correspond to meaningful hazards; uncertainty communication can improve oversight when humans understand and can act upon it; documentation can enable traceability and auditability; and fairness analysis can inform accountability when responsibility for metric selection and remediation is explicit. Stakeholder participation may similarly improve contextual understanding when participants possess genuine influence rather than symbolic involvement.

Other relationships create tensions. Detailed transparency may expose confidential data, explanation interfaces may reveal attackable information, privacy-preserving records may reduce audit detail, and extensive automation may weaken situational awareness. Predictive-performance improvements may also distribute errors unequally. **Table 4** presents the cross-dimensional synthesis of trustworthy AI evidence, including tensions, trade-offs, design implications, evidence gaps, and future research needs. **Figure 2** presents the cross-dimensional synthesis model of trustworthy AI, linking transparency, explainability, robustness, safety, privacy, security, fairness, accountability, governance, and human oversight as interdependent requirements rather than isolated system properties.

Table 4. Cross-Dimensional Evidence Synthesis for Trustworthy and Explainable Intelligent Systems: Transparency–Privacy Tensions, Explainability–Fairness Links, Robustness–Safety Dependencies, Privacy–Accountability Trade-Offs, Human Oversight Limits, Governance Requirements, Evidence Gaps, and Future Research Priorities

Cross-dimensional relationship	Dimensions involved	Core synthesis question	Evidence pattern	Tension or trade-off	Design or evaluation implication	Remaining gap

Transparency–privacy	Transparency, documentation, privacy	How much disclosure supports scrutiny without exposing protected information?	Disclosure aids review but may reveal sensitive data or model details	Inspectability versus confidentiality	Use role-based, layered, purpose-limited disclosure	Limited end-end evaluation
Explainability–security	Explainability, model vulnerability, security	Can explanations reveal attackable information?	Explanations support diagnosis but may expose boundaries or influential features	Fidelity and detail versus attack surface	Threat-model explanation interfaces	Security effects of XAI remain unevenly studied
Explainability–fairness	Explainability, bias, fairness, recourse	Do explanations reveal unfairness or rationalise it?	XAI may support detection and contestability but can faithfully explain an unfair model	Clarity may legitimise harmful logic	Evaluate model and explanation fairness together	Limited affected-user evidence
Fairness–accountability	Fairness, responsibility, auditability	Who selects fairness definitions and owns trade-offs?	Metric choice embeds normative and organisational decisions	Technical optimisation versus social responsibility	Record metric rationale, authority, and corrective duties	Responsibility is often diffused
Robustness–safety	Robustness, reliability, resilience, safety	When does robustness support a safety claim?	Stress testing can reduce known failure risk	Benchmark strength versus operational relevance	Connect tests to hazards and consequences	Weak benchmark-to-harm linkages
Privacy–accountability	Privacy, traceability, auditability	How can reviewable evidence be retained without excessive personal data?	Accountability needs records; privacy favours minimisation	Investigative detail versus data protection	Use secure provenance and controlled access	Limited privacy-preserving audit evidence
Human oversight–automation	Oversight, control, automation	When is human participation effective?	Human review can identify errors but may fail under overload or bias	Efficiency versus meaningful intervention	Match oversight mode to risk and response time	Few realistic comparative studies

Governance–technical validation	Governance, assurance, verification	How should bounded technical evidence inform institutional decisions?	Tests support claims; governance sets thresholds and responsibility	Narrow metrics may be overinterpreted	Use claim–evidence traceability	Integrated processes la validation
Uncertainty–human decisions	Uncertainty, calibration, explanation, oversight	Does uncertainty communication improve reliance?	It may support escalation but can increase cognitive burden	Information value versus complexity	Test representations through decision outcomes	Limited domain-realistic evidence
Documentation–auditability	Documentation, traceability, accountability	Can auditors use the recorded evidence effectively?	Documentation supports review but may become stale or performative	Reporting burden versus evidentiary quality	Link records to audit questions and update triggers	Limited evidence o downstream use
Stakeholder participation–trustworthiness	Participation, fairness, governance	Does participation materially change design and risk decisions?	Contextual knowledge may improve fit and legitimacy	Inclusion versus tokenism	Specify representation and decision influence	Participation outcomes at under-evaluated
Performance–trustworthiness	Accuracy, fairness, robustness, safety	How should predictive performance relate to broader trustworthiness?	High performance can coexist with opacity, bias, or insecurity	Metric improvement may harm another dimension	Use multidimensional evidence profiles	No validate universal integration ru
Single-dimension sufficiency boundary	All dimensions	Can one property establish worthiness of trust?	Evidence consistently shows contextual interdependence	Strength in one dimension may conceal severe weakness elsewhere	Preserve veto-level risks and explicit assumptions	No universa trustworthine metric

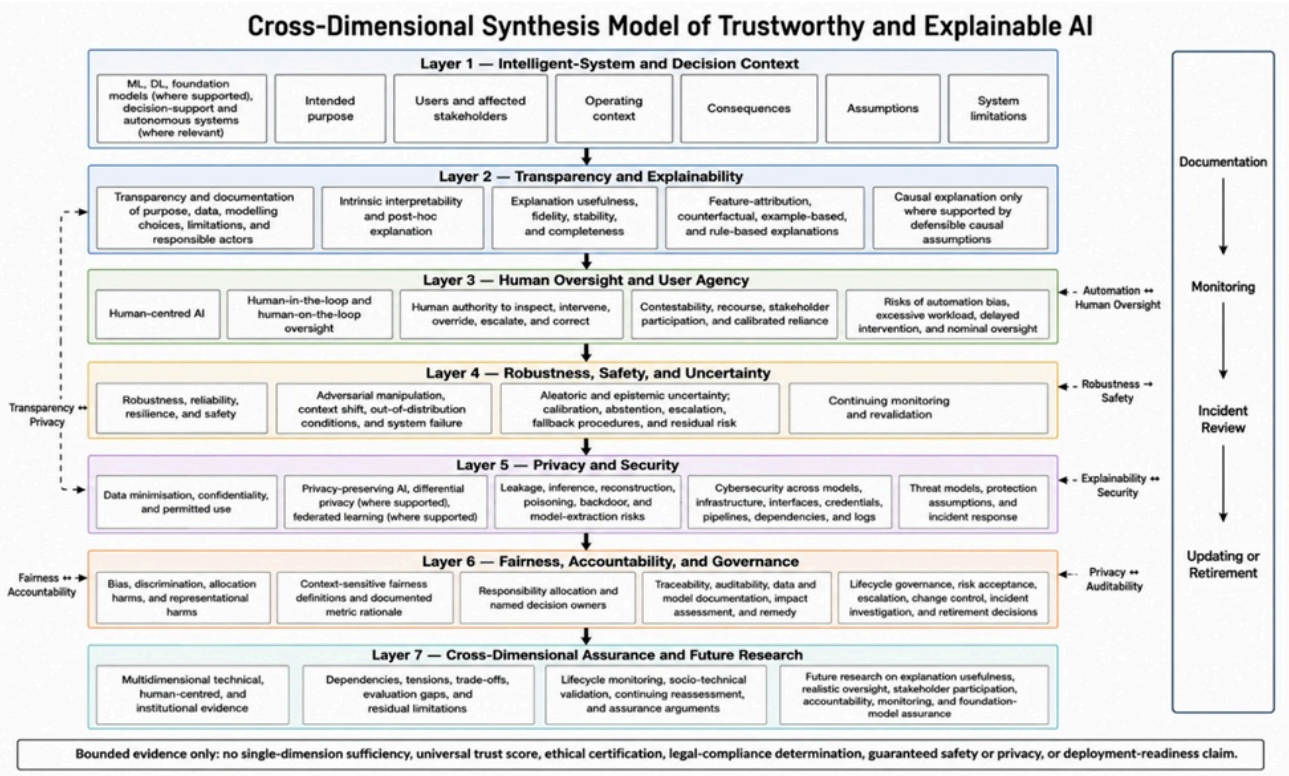


Figure 2. Cross-Dimensional Synthesis Model of Trustworthy and Explainable AI

Alt-text description. A conceptual synthesis model showing trustworthy AI as a multidimensional system connecting transparency, explainability, human oversight, robustness, safety, privacy, security, fairness, accountability, governance, monitoring, documentation, and lifecycle assurance.

Research Gaps and Future Research Agenda

Conceptual fragmentation remains a foundational limitation. Trust, trustworthiness, responsible AI, ethical AI, explainability, assurance, compliance, and safety are frequently used at different analytical levels, complicating comparison. Future research should report explicit construct definitions and distinguish system properties from human attitudes and institutional judgements. It should also avoid universal trustworthiness metrics, because aggregation may conceal assumptions, incompatible criteria, and severe weaknesses that should not be compensated by strengths elsewhere.

Evaluation research should move beyond isolated dimensions. Explanation studies need stronger validation of usefulness, fidelity, stability, recourse, and effects on human decisions rather than reliance on satisfaction alone [12]. Fairness research requires transparent justification of selected metrics and closer examination of intersecting harms [25]. Robustness, privacy, security, and fairness should increasingly be evaluated together because interventions designed for one property may alter performance or risk across others.

Human oversight and socio-technical validation require more realistic evidence. Future studies should compare human-in-the-loop and human-on-the-loop arrangements under representative workload, time pressure, uncertainty, and organisational authority. Outcomes should include error detection, disagreement resolution, intervention quality, calibrated reliance, and recovery from automation failure [17]. Participatory research should additionally examine whose knowledge is included, whether participation changes design decisions, and how conflicts among stakeholder values are resolved.

Lifecycle evidence remains limited for monitoring, accountability, changing contexts, and foundation-model use. Prospective research should examine distribution shift, incident reporting, documentation maintenance, audit effectiveness, update governance, and retirement decisions. Lifecycle assurance can structure these investigations, but its claims must remain context-specific [31]. Foundation-model research should address open-ended behaviour, data provenance, tool use, security exposure, uncertainty, and oversight without presuming that evaluation methods developed for narrower models transfer unchanged [34].

Conclusion

This systematic review synthesises trustworthy and explainable AI as an interdependent, socio-technical, and lifecycle-aware domain linking transparency, interpretability, explanation quality, meaningful human oversight, robustness, safety, privacy, security, fairness, accountability, documentation, auditability, governance, and monitoring. The evidence shows that no isolated model property can establish worthiness of trust: explanations may be unfaithful, robust systems may remain unfair, privacy-oriented designs may remain insecure, and human participation may be merely nominal. Appropriate assessment therefore requires bounded technical validation, stakeholder-sensitive human evaluation, explicit responsibility, institutional governance, and continuing reassessment under changing conditions. This synthesis supports future research and assessment design but does not establish universal measurement, legal compliance, ethical certification, guaranteed safety, or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Aug 2024

Revised: 23 Sep 2024

Accepted: 06 Nov 2024

Published online: 15 March 2025

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

- Floridi L, Cowls 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* (Dordr). 2018;28(4):689-707.
<https://doi.org/10.1007/s11023-018-9482-5>.
- 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>.
- 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>.
- 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>.
- Ryan M. In AI we trust: ethics, artificial intelligence, and reliability. *Sci Eng Ethics*. 2020;26(5):2749-67.
<https://doi.org/10.1007/s11948-020-00228-y>.
- 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>.
- Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, López de Prado M, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf Fusion*. 2023;99:101896.
<https://doi.org/10.1016/j.inffus.2023.101896>.
- 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>.
- Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138-60.
<https://doi.org/10.1109/ACCESS.2018.2870052>.
- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Benetot 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>.
- 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>.
- 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>.
- 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>.

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>.

Rong Y, Leemann T, Nguyen TT, Fiedler L, Qian P, Unhelkar V, et al. Towards human-centered explainable AI: a survey of user studies for model explanations. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(4):2104-22.

<https://doi.org/10.1109/TPAMI.2023.3331846>.

Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognit*. 2018;84:317-31.

<https://doi.org/10.1016/j.patcog.2018.07.023>.

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>.

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>.

Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag*. 2020;37(3):50-60.

<https://doi.org/10.1109/MSP.2020.2975749>.

Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G. A survey on security and privacy of federated learning. *Future Gener Comput Syst*. 2021;115:619-40.

<https://doi.org/10.1016/j.future.2020.10.007>.

Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: a survey. *IEEE Rev Biomed Eng*. 2021;14:156-80.

<https://doi.org/10.1109/RBME.2020.3013489>.

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):115:1-35.

<https://doi.org/10.1145/3457607>.

Pessach D, Shmueli E. A review on fairness in machine learning. *ACM Comput Surv*. 2022;55(3):1-44.

<https://doi.org/10.1145/3494672>.

Martin K. Ethical implications and accountability of algorithms. *J Bus Ethics*. 2019;160(4):835-50.

<https://doi.org/10.1007/s10551-018-3921-3>.

Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019;1(11):501-7.

<https://doi.org/10.1038/s42256-019-0114-4>.

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>.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. *Commun ACM*. 2021;64(12):86-92.

<https://doi.org/10.1145/3458723>.

Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. *ACM Comput Surv.* 2023;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>.

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>.

Mökander J, Floridi L. Ethics-based auditing to develop trustworthy AI. *Minds Mach (Dordr).* 2021;31(2):323-7.
<https://doi.org/10.1007/s11023-021-09557-8>.

de Almeida PGR, dos Santos CD, Farias JS. Artificial intelligence regulation: a framework for governance. *Ethics Inf Technol.* 2021;23(3):505-25.
<https://doi.org/10.1007/s10676-021-09593-z>.

Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. *Humanit Soc Sci Commun.* 2024;11(1):1568.
<https://doi.org/10.1057/s41599-024-04044-8>.