

ORIGINAL RESEARCH

Open access

Building Trustworthiness into Intelligent Systems: From Data Provenance and Model Reliability to Human Control and Institutional Accountability

James Whitfield¹, Aisha Rahman^{2*}, Kenji Sato³

Abstract

Intelligent systems increasingly influence decisions, allocate attention, generate recommendations, and mediate interactions across technical, organizational, and social environments. Yet trustworthiness is often approached through isolated interventions, such as improving model accuracy, adding post-hoc explanations, documenting datasets, or assigning nominal human oversight. This article proposes a trustworthy-by-design conceptual framework that treats trustworthiness as a system-level property requiring coordinated design and evaluation across the full intelligent-system lifecycle. The framework begins with system-purpose and risk-context definition, followed by stakeholder and responsibility mapping, data-integrity assessment, provenance and lineage review, data documentation and governance, model-reliability evaluation, robustness and uncertainty assessment, and bias and fairness review where relevant. It then connects explanation quality with human interpretability and decision usefulness, rather than assuming that technically generated explanations will necessarily support understanding. Human oversight, meaningful control, contestability, recourse, institutional accountability, traceability, auditability, monitoring, and feedback updating form additional interdependent layers. The framework's principal contribution is to clarify how technical evidence must be connected to human decision conditions and institutional responsibility without treating any individual safeguard as sufficient. Data provenance does not establish data appropriateness, robustness does not prove safety, explanations do not guarantee interpretability, human presence does not ensure meaningful control, and governance arrangements do not independently establish accountability. The framework may support conceptual analysis, assurance-oriented evaluation, and lifecycle governance planning, but requires empirical validation, human-centered assessment, domain-specific interpretation, and continuing institutional review.

Keywords Explainability, Trustworthy AI, Model reliability, Uncertainty, Trustworthy-by-design, Data provenance

*Correspondence:

Aisha Rahman
aisha.rahman@lums.edu.pk

¹ Department of Trustworthy Systems Engineering, Warwick Business School, Coventry, United Kingdom

² Department of AI Governance and Policy, Lahore University of Management Sciences, Lahore, Pakistan

³ Department of Intelligent Systems and Informatics, Keio University, Tokyo, Japan

Introduction

Trustworthy intelligent systems have become a central concern because technical performance alone does not determine whether a system deserves reliance. Human trust is a psychological and relational state shaped by perceived competence, predictability, transparency, familiarity, and contextual expectations, whereas trustworthiness concerns whether reliance is warranted by the system's actual qualities and operating conditions. A system may be trusted despite being unreliable, or distrusted despite strong evidence of reliability. Trustworthy-by-design analysis must therefore avoid

using reported confidence, user acceptance, or continued use as substitutes for evidence that data, models, explanations, controls, and institutional arrangements are appropriate for the intended purpose [1].

Responsible-AI and ethical-AI discussions have articulated principles such as transparency, fairness, privacy, safety, responsibility, and human agency. Their recurrence across institutional and disciplinary settings demonstrates broad agreement that intelligent systems raise interconnected technical and social concerns. However, apparent convergence at the level of principle can conceal differences in definitions, priorities, implementation responsibilities, and methods of evaluation. Trustworthy-by-design thinking must consequently translate broad commitments into traceable design questions, evidence requirements, decision boundaries, and review processes rather than assume that declaring adherence to commonly accepted principles establishes trustworthy operation [2].

A useful conceptual foundation connects human autonomy, harm prevention, fairness, and explicability while recognizing that these values interact rather than operate independently. Explicability is especially important because it contains both an epistemic dimension—whether relevant actors can understand a system—and an accountability dimension—whether decisions and consequences can be traced to responsible human and institutional actors. Nevertheless, normative principles remain orientation devices rather than validation results. Their contribution lies in identifying what should receive attention and why, while empirical, technical, human-centered, and organizational assessment must determine how those concerns manifest in a particular intelligent system [3].

Principles cannot independently guarantee ethical conduct because intelligent systems are developed and governed through organizations with uneven expertise, incentives, authority, resources, and accountability structures. A trustworthy-by-design approach must therefore address the mechanisms through which principles influence data acquisition, system architecture, validation, explanation design, operational control, documentation, incident handling, and institutional review. This article develops an original conceptual framework for that purpose. It organizes trustworthiness across data, model, explanation, human-control, and institutional-accountability layers and clarifies their lifecycle dependencies without presenting the framework as an assurance certificate, legal determination, universal governance method, or indication of deployment readiness [4].

Trustworthiness as a System-Level Design Challenge

Trustworthiness is best understood as a multidimensional property of an intelligent system embedded within a socio-technical environment. Relevant qualities include data integrity, privacy, security, reliability, robustness, fairness, transparency, explainability, human agency, accountability, and organizational competence. These qualities cannot be reduced to one score because they answer different questions. Reliability concerns consistent performance, provenance concerns the origins and transformations of data, explanation quality concerns whether a representation accurately and usefully communicates relevant system behaviour, and accountability concerns whether identifiable actors can answer for decisions and consequences. Trustworthy-by-design reasoning must preserve these distinctions while examining their interaction [5].

The trustworthy-AI literature remains distributed across technical, ethical, human-centered, and governance traditions. Robustness research may focus on perturbations and adversarial behaviour; fairness research may analyse subgroup disparities; explainability research may evaluate feature attribution or interpretable representations; privacy and security research may address access, leakage, or manipulation; and governance research may focus on organizational responsibility and oversight. Each area contributes necessary evidence, but fragmentation can encourage dimensional substitution, in which success on one measurable property is treated as evidence of overall trustworthiness. A system-level framework is needed to coordinate evidence without collapsing distinct concerns into a generic checklist [6].

Trust must also be distinguished from reliability and moral responsibility. Reliable operation may support rational reliance, but an intelligent system is not necessarily a moral agent capable of accepting responsibility, justifying institutional choices, or responding to affected persons. Anthropomorphic language can obscure this distinction by attributing

intentions, commitments, or accountability to the system itself. Trustworthy-by-design analysis instead locates responsibility among those who define the purpose, supply and govern data, select and validate models, design explanations and interfaces, determine control arrangements, authorize operation, monitor outcomes, and respond to failures [7].

The system-level perspective extends beyond algorithmic properties to the wider effects of intelligent systems on human autonomy, rights, institutional practices, social relationships, and distributions of benefit and burden. AI assurance and governance should therefore consider not only whether components function as intended, but also whether the purpose, decision environment, stakeholder participation, monitoring arrangements, and institutional responses remain defensible over time. Lifecycle governance is necessary because datasets, models, organizations, user practices, and environments change. Even a carefully designed framework cannot establish deployment readiness without domain-specific evidence, operational evaluation, accountable authorization, and continued reassessment [8].

Data Integrity and Provenance Requirements

Data documentation is a foundational trustworthy-by-design requirement because datasets embody decisions about populations, sources, collection procedures, annotation, filtering, categorization, and intended use. Dataset records should make these decisions visible, including relevant limitations, excluded populations, annotation conditions, and contexts in which generalization may be unsupported. Documentation can help downstream actors evaluate representativeness, bias, and appropriate reuse, but its value depends on accuracy, maintenance, accessibility, and integration into decision processes. A completed template is not itself evidence that the underlying data are accurate, fair, suitable, or responsibly governed [9].

Data provenance and lineage provide related but distinct evidence. Provenance concerns where data originated, how they were acquired, which actors or processes contributed to them, and what derivations or transformations occurred. Lineage follows movement and dependency through collection, preparation, integration, labelling, feature construction, model training, evaluation, and monitoring pipelines. Together, they can support reproducibility, impact analysis, error investigation, attribution, and auditability. However, a complete lineage record can faithfully describe a flawed process. Provenance therefore supports traceability but cannot independently establish data quality, representativeness, authorization, privacy, security, or fitness for purpose [10].

Data integrity and quality require active examination of accuracy, consistency, completeness, timeliness, missingness, duplication, labelling quality, and contextual relevance. Machine-learning data pipelines frequently involve iterative acquisition, cleaning, transformation, augmentation, and relabelling, making quality a continuing process rather than a one-time gate. Completeness must be evaluated against the intended decision context, because a dataset may contain few missing cells while omitting important populations, environments, time periods, operating states, or edge cases. Bias assessment must similarly consider historical, sampling, measurement, representation, aggregation, and annotation pathways rather than treat bias as a single statistical defect [11].

Data governance connects technical records to institutional authority. Ownership and stewardship concerns determine who may collect, modify, share, retain, restrict, or withdraw data; access control determines who can view or manipulate particular assets; and privacy and security practices address informational harm, unauthorized disclosure, extraction, poisoning, and loss. Traceability, auditability, and evidence trails depend on assigned responsibilities, maintained records, reviewable decisions, and procedures for handling exceptions and changes. Governance cannot convert unsuitable data into trustworthy data, but it can make responsibilities and unresolved limitations more visible and contestable [12]. **Table 1** summarises data integrity, provenance, documentation, governance, traceability, and auditability requirements for trustworthy-by-design intelligent systems.

Table 1. Data Integrity and Provenance Requirements for Trustworthy-by-Design Intelligent Systems: Data Lineage, Data Quality, Completeness, Representativeness, Bias, Documentation, Dataset Documentation, Governance, Ownership

Concerns, Access Control, Privacy, Security, Traceability, Auditability, Evidence Trails, and Data-Provenance Boundaries

Data requirement	Core trustworthiness question	Required evidence or practice	Trustworthiness function	Potential output	Failure or misuse risk	Governance or value
Data integrity	Have relevant data remained accurate and protected from unintended alteration?	Validation checks, version histories, anomaly review, authorized-change logs	Preserves evidential reliability	Integrity profile	Corrupted or silently altered inputs	Stewardship and accountability
Data provenance	Can origin and acquisition context be reconstructed?	Source, collection, authority, and derivation records	Establishes origin evidence	Provenance record	Unverifiable or inappropriate reuse	Requirement for reconstruction
Data lineage	Can transformations and dependencies be traced end to end?	Pipeline maps, versions, timestamps, transformation records	Supports reproduction and impact analysis	Lineage map	Hidden transformations	Confidence management
Data quality	Are data fit for the defined purpose?	Accuracy, consistency, validity, timeliness, and duplication assessment	Supports contextual validity	Quality profile	Generic quality claims masking poor fit	Domain-specific criteria
Data completeness	Are relevant cases, variables, periods, and states represented?	Missingness and coverage analysis	Identifies evidential gaps	Completeness statement	Systematic omission	User requirements
Data representativeness	Do data reflect relevant populations and environments?	Sampling and distribution analysis	Supports bounded generalization	Representativeness map	Exclusion and unequal performance	Stakeholder representation
Data bias	What bias pathways may be embedded?	Historical, sampling, measurement, proxy, and annotation analysis	Makes patterned distortion visible	Bias register	Reinforcement of inequity or error	Normative and ethical considerations

Data documentation	Are purpose, sources, preparation, ownership, and limitations recorded?	Maintained documentation with named owners	Supplies context for downstream actors	Data record	Stale or ceremonial documentation	Re-sch
Dataset documentation	Are population, annotation, use, and limitation details recorded?	Dataset statement or equivalent record	Supports appropriate interpretation	Dataset documentation	Unsupported generalization	Ch trig re
Data governance	Are decision rights and stewardship responsibilities defined?	Roles, approvals, escalation, and change procedures	Converts records into accountable practice	Governance map	Diffuse responsibility	Insti ove
Ownership concern	Who has authority over collection, reuse, retention, and access?	Stewardship and authority rationale	Clarifies duties and decision rights	Ownership register	Unauthorized use or disputes	Co sp re
Access control	Who may access or modify data, and why?	Permissions, access logs, revocation procedures	Limits misuse and supports tracing	Access record	Unauthorized disclosure or alteration	Se re
Privacy	Could collection or use create informational harm?	Purpose limitation, minimization, access restriction	Reduces avoidable privacy risk	Privacy-risk statement	Re-identification or secondary use	Do sp asse
Security	Are data protected from manipulation, extraction, loss, or denial of access?	Threat analysis, safeguards, logging, recovery planning	Protects confidentiality, integrity, and availability	Security evidence	Leakage, poisoning, or evidence loss	Con mor
Traceability	Can changes, approvals, and dependencies be followed?	Persistent identifiers and linked logs	Supports reconstruction	Traceability chain	Inability to investigate failures	Evi ret r
Auditability	Can relevant claims and records be scrutinized?	Accessible evidence, review scope, exception histories	Enables structured challenge	Auditable evidence package	Selective disclosure or audit theatre	Con appi sep re

Evidence trail	Are claims linked to supporting records and decisions?	Attributable, timestamped, versioned evidence	Preserves accountability evidence	Evidence map	Unsupported retrospective claims	Evidence quality requirements
Data-provenance boundary	Is traceability being mistaken for data trustworthiness?	Separate quality, bias, privacy, security, and governance assessments	Prevents overclaiming	Qualified provenance statement	Traceable but unsuitable data	Confidence dimensions requirements

Model Reliability, Robustness, and Uncertainty

Model reliability concerns whether an intelligent system performs with sufficient consistency for a defined purpose, population, environment, and decision process. Validity is broader: it asks whether the evaluation evidence supports the interpretation and use claimed for the model. Uncertainty is central to both because predictive outputs can reflect irreducible variability in observations as well as reducible limitations in model knowledge. Distinguishing aleatoric from epistemic uncertainty helps clarify whether uncertainty primarily concerns noisy or inherently variable data, limited knowledge, sparse coverage, model assumptions, or unfamiliar operating conditions. The distinction does not eliminate ambiguity, but it discourages treating every confidence value as the same type of evidence [13].

Uncertainty estimation in deep learning includes Bayesian approximations, ensembles, sampling-based methods, probabilistic outputs, and other techniques with different assumptions and computational demands. Their usefulness depends on whether the resulting estimates are empirically evaluated, appropriately calibrated, stable under relevant conditions, and understandable to the people using them. Uncertainty communication must connect numerical or qualitative representations to decision consequences, escalation pathways, and abstention conditions. A technically sophisticated uncertainty method may still mislead when users interpret it as a guarantee, overlook distribution shift, or cannot determine what action should follow from a high-uncertainty output [14].

Robustness concerns how model behaviour changes under perturbations, noise, missing information, adversarial manipulation, or environmental variation. Meaningful robustness testing requires explicit threat models, plausible disturbance scenarios, reproducible procedures, and transparent scope limitations. Out-of-distribution evaluation should examine whether the system detects unfamiliar inputs, communicates reduced confidence, defers action, or fails confidently. Model safety cannot be inferred from robustness alone because a robust model may remain invalid for the intended decision, produce unfair effects, provide misleading explanations, or operate within an institution unable to detect and respond to harm. Robustness is therefore a bounded evidential component rather than a synonym for trustworthiness [15].

Error analysis should identify not only how often a system is wrong, but also which errors occur, under what conditions, for whom, and with what consequences. Bias and fairness assessment can expose systematic disparities that aggregate performance obscures, although no fairness metric can resolve all normative questions or represent every affected group [16]. Reliability is also time-dependent. Data drift and concept drift can change input distributions, relationships between variables and outcomes, or the meaning of operational categories, thereby weakening previously established validity. Monitoring and feedback updating should therefore be linked to accountable investigation, controlled modification, revalidation, versioning, and rollback rather than unrestricted automatic retraining [17]. Finite validation evidence cannot guarantee future reliability, making lifecycle monitoring and qualified use boundaries indispensable.

Explanation Quality and Human Interpretability

Explainability concerns the capacity to provide information about a model, process, or output, whereas interpretability concerns whether a person can form an appropriately accurate understanding of relevant system behaviour. Explanations are not merely technical artefacts: they are selective, contrastive, audience-dependent, and shaped by the question being asked. An explanation useful to a model developer investigating an error may be unsuitable for an operator deciding whether to rely on a recommendation or for an affected person seeking to challenge an outcome. Trustworthy-by-design systems must therefore define the explanation's audience, purpose, timing, and decision context before selecting an explanatory method [18].

Explanation methods also differ in scope and evidential function. They may be local or global, intrinsic or post-hoc, model-specific or model-agnostic, descriptive or contrastive. Transparency may concern datasets, model architecture, development processes, limitations, responsible actors, or decision rules rather than a single technical disclosure. These distinctions matter because an explanation of feature influence does not necessarily reveal causal relationships, model validity, data suitability, or institutional reasons for using the system. Explainability should consequently be treated as a structured set of design and evaluation questions rather than one binary property [19].

Explanation quality requires explicit assessment of fidelity, usefulness, stability, completeness, clarity, and cognitive accessibility where relevant. Fidelity asks whether the explanation reflects the system's actual behaviour; usefulness asks whether it supports the intended human task; stability considers whether materially similar cases receive appropriately consistent explanations; and completeness concerns whether sufficient information is provided for the stated purpose without implying total disclosure. Human-centered evaluation should assess comprehension, mental-model accuracy, error recognition, decision quality, and reliance. Technical plausibility, readability, or user satisfaction alone cannot establish that an explanation supports sound understanding [20].

Human interpretability also depends on expertise, workload, interface design, uncertainty communication, and the consequences of misunderstanding. Explanations may increase confidence while failing to improve judgment, particularly when persuasive post-hoc accounts are mistaken for faithful representations of black-box reasoning. In high-stakes contexts, inherently interpretable models may sometimes offer advantages because relevant reasoning can be inspected more directly, although interpretability does not remove the need for validation and governance. Explanations should also support contestability by helping reviewers or affected persons identify relevant factors, limitations, and routes for challenge. Explanation availability does not independently establish reliability, fairness, accountability, or appropriate reliance [21].

Human Control and Institutional Accountability

Human control concerns the allocation of decision authority between intelligent systems and people. Human-in-the-loop arrangements require human input or approval before specified actions, whereas human-on-the-loop arrangements permit automated activity under human supervision. Meaningful human control demands more than nominal presence. Relevant actors must possess sufficient authority, competence, information, time, and intervention capacity, and system behaviour must remain responsive to relevant human reasons. Control must also be traceable to proximal actors operating the system and distal actors who shape its purpose, design, constraints, and institutional environment [22].

High automation and strong human control need not be opposites. A system may automate complex analytical functions while preserving human authority over goals, exceptional cases, irreversible actions, or decisions carrying substantial consequences. The appropriate arrangement depends on task characteristics, system speed, uncertainty, workload, reversibility, and the operator's ability to understand system state. Decision boundaries should specify what the system may recommend, decide, execute, defer, or refuse. Override mechanisms require operational testing because an inaccessible, delayed, confusing, or unsafe override cannot be treated as effective control merely because it formally exists [23].

Contestability and recourse extend human control beyond operators. Contestability enables affected persons or authorized reviewers to question outputs, evidence, assumptions, or procedures. Recourse provides a meaningful possibility of reconsideration, correction, or remedy. These mechanisms require identifiable responsible actors, usable explanations, evidence access, response duties, and authority to change outcomes. Responsibility must therefore be mapped across data collection, model development, validation, interface design, operational use, monitoring, and institutional authorization rather than displaced onto the algorithm or concentrated artificially in the final human decision-maker [24].

Institutional accountability requires organizations capable of demanding explanations, assigning responsibilities, reviewing conduct, preserving evidence, and responding to failures. Technical evidence from provenance records, reliability testing, uncertainty assessment, explanation evaluation, and monitoring acquires governance significance only when competent actors can interpret it and take consequential action. Traceability and documentation support this process by linking system states, decisions, changes, approvals, interventions, and outcomes. **Figure 1** illustrates the system-level trustworthiness chain linking data provenance, model reliability, explanation quality, human control, institutional accountability, auditability, monitoring, and feedback updating [25].

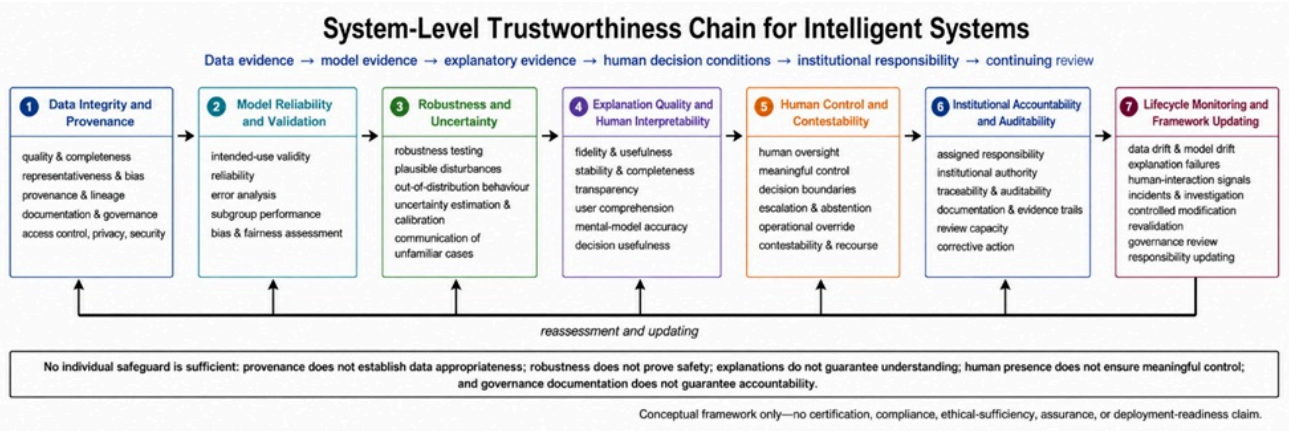


Figure 1. System-Level Trustworthiness Chain for Intelligent Systems

Alt-text description:

A conceptual chain diagram showing data provenance, data quality, model reliability, robustness, uncertainty, explanation quality, human interpretability, human control, accountability, traceability, auditability, monitoring, and feedback updating.

Auditability can support structured scrutiny of data, models, explanations, controls, and organizational processes, but it does not independently prove ethical sufficiency or effective accountability. Audit evidence is shaped by scope, access, criteria, reviewer competence, independence, and institutional willingness to act on findings. Lifecycle governance should therefore connect monitoring and incident evidence to investigation, corrective action, revalidation, and responsibility updating. Human-control arrangements likewise require continuing evaluation as systems, workloads, users, and operating environments change. Neither human presence nor formal governance documentation guarantees that control and accountability will function effectively in practice [26]. **Table 2** maps model reliability, robustness, uncertainty, explanation quality, human interpretability, human control, and institutional accountability requirements for trustworthy intelligent systems.

Table 2. Model Reliability, Explanation Quality, Human Control, and Institutional Accountability Requirements for Trustworthy Intelligent Systems: Robustness, Uncertainty, Calibration, Monitoring, Explanation Fidelity, Human Interpretability, Decision Usefulness, Oversight, Contestability, Recourse, Traceability, Auditability, Lifecycle Governance, and Accountability Boundaries

Requirement domain	Core design or governance question	Required evidence or design condition	Trustworthiness function	Potential output	Failure or responsibility risk	Validation, evaluation, or governance need
Model reliability	Does performance remain sufficiently consistent for the intended use?	Repeated evaluation, subgroup results, error distributions	Characterizes dependable performance	Reliability profile	Aggregate metrics conceal consequential failures	Domain-specific validation and monitoring
Model validity	Does evidence support the claimed interpretation and use?	Construct, external, decision-context, and consequence evidence	Connects performance to intended purpose	Validity argument	Correct metric applied to the wrong problem	Independent contextual review
Robustness	How does behaviour change under disturbance or manipulation?	Threat models, stress tests, perturbation and sensitivity analysis	Identifies fragility	Robustness envelope	Narrow testing presented as universal resilience	Scenario review and repeated testing
Out-of-distribution behaviour	Can unfamiliar cases be recognized or safely deferred?	Shift tests, novelty detection, abstention evaluation	Limits unsupported extrapolation	OOD response profile	Confident output outside validated conditions	Distribution monitoring
Uncertainty	What limitations accompany an output?	Method specification and empirical uncertainty evaluation	Represents predictive and knowledge limitations	Uncertainty profile	False precision or undifferentiated confidence	Reassessment under change
Calibration	Do probabilities correspond to outcomes within the evaluated context?	Calibration curves and subgroup assessment	Supports probabilistic interpretation	Calibration profile	Good average calibration hides local error	Periodic recalibration
Error and bias assessment	Which errors occur, for whom, and under what conditions?	Error taxonomy, subgroup and proxy analysis	Reveals patterned failures	Error and bias register	Harmful minority failures remain hidden	Stakeholder and domain review

Drift monitoring	Have data or model–outcome relationships changed?	Baselines, live distributions, outcomes, alerts	Detects degradation	Drift report	False alarms or delayed detection	Human investigation and controlled updating
Explanation fidelity	Does the explanation reflect relevant model behaviour?	Faithfulness and perturbation testing	Reduces misleading rationalization	Fidelity evidence	Persuasive but false explanation	Technical evaluation
Explanation usefulness	Does the explanation support the intended task?	Task-based and decision-quality evaluation	Links explanation to use	Usefulness profile	Confidence rises without better decisions	Human-centered studies
Explanation stability and completeness	Is the explanation sufficiently consistent and informative?	Sensitivity, coverage, and omission analysis	Identifies fragility and scope	Stability and completeness statement	Partial account treated as exhaustive	Technical and user review
Human interpretability	Do intended users understand the system and its limitations?	Comprehension, mental-model, and reliance assessment	Tests actual understanding	Interpretability evidence	Readability mistaken for comprehension	Evaluation with target users
Decision usefulness	Does information support appropriate reliance and error detection?	Comparative task evaluation	Connects understanding to decisions	Decision-usefulness profile	Automation bias or avoidable rejection	Longitudinal human-centered evaluation
Human oversight and control	Can capable humans review and influence relevant actions?	Authority, competence, information, timing, workload	Preserves human agency	Control architecture	Symbolic or ineffective oversight	Operational testing
Decision boundaries and override	What may the system do, and can action be contained?	Scope, abstention, escalation, override and recovery tests	Limits delegated authority	Boundary specification	Authority expansion or unusable override	Governance approval and exercises
Contestability and recourse	Can decisions be challenged	Notice, reasons, review channel,	Enables answerability	Contestation process	Appeal without practical	Affected-stakeholder

	and corrected?	corrective authority	and remedy		consequence	evaluation
Institutional accountability	Who answers for design, use, effects, and failures?	Responsibility, authority, review and consequence mechanisms	Prevents responsibility displacement	Accountability map	"The algorithm decided" gap	Institutional capability review
Traceability and auditability	Can evidence, decisions, and changes be reconstructed and scrutinized?	Linked logs, versions, documentation and review access	Supports investigation and challenge	Auditable evidence package	Missing context or audit theatre	Competent review and follow-up
Lifecycle governance	How are monitoring, incidents, changes, and responsibilities managed over time?	Review triggers, incident process, versioning, revalidation	Maintains continuing oversight	Lifecycle-governance record	One-time approval logic	Periodic and event-triggered review

Proposed Trustworthy-by-Design Framework

The proposed framework organizes trustworthiness as a lifecycle of connected claims, evidence, decisions, and feedback. It begins with system-purpose and risk-context definition because data and model quality cannot be assessed independently of what the system is intended to do, whose decisions it affects, and which harms or limitations matter. Stakeholder and responsibility mapping then identifies system owners, data stewards, developers, validators, operators, affected groups, reviewers, and decision authorities. This structure translates ethical commitments into lifecycle questions and evidence requirements rather than treating responsible-AI tools as isolated interventions [27].

The second stage addresses data integrity, provenance, lineage, documentation, governance, ownership concerns, access, privacy, security, representativeness, and bias. The third stage evaluates model reliability, validity, robustness, safety, out-of-distribution behaviour, uncertainty, calibration, error patterns, fairness where relevant, and monitoring requirements. These stages should be revisited iteratively because design choices alter data needs, model behaviour, stakeholder exposure, and possible harms. Responsible design therefore requires anticipatory assessment, reflection on assumptions, stakeholder participation, and responsiveness to emerging evidence throughout development and use [28].

The fourth and fifth stages connect explanation quality to human interpretability and human control. Explanation assessment examines fidelity, usefulness, stability, completeness, cognitive accessibility, uncertainty communication, and decision usefulness for specified audiences. Human-control design defines authority, supervision, approval points, abstention, escalation, override, contestability, and recourse. Contextual ethical interpretation may benefit from multidisciplinary expertise, but institutions cannot outsource responsibility for objectives, trade-offs, or consequences. Specialist processes should support rather than replace accountable institutional judgment [29].

The sixth and seventh stages establish institutional accountability, traceability, auditability, lifecycle monitoring, incident response where relevant, and governance updating. Monitoring should capture data drift, model drift, uncertainty, errors, explanation failures, human-interaction signals, contestation, and incidents. Evidence must then be interpreted by responsible actors who can restrict operation, require investigation, authorize controlled change, or initiate revalidation.

The framework consequently links technical, human, and institutional dimensions without assuming that any one property can stand in for overall trustworthiness [30].

The framework includes explicit non-sufficiency and decision boundaries. Provenance does not establish data appropriateness; robustness does not prove validity or safety; explanation does not guarantee understanding; human presence does not establish meaningful control; and governance arrangements do not guarantee accountability. Monitoring evidence does not self-validate, and framework completion does not establish legal compliance, ethical sufficiency, certification, or deployment readiness. The framework may help organize assurance-oriented arguments, but those arguments require domain-specific empirical evidence, human-centered evaluation, accountable institutional review, and continuing challenge. **Table 3** presents the proposed trustworthy-by-design conceptual framework linking data provenance, model reliability, explanation quality, human control, accountability, governance, monitoring, and framework updating [26].

Table 3. Proposed Trustworthy-by-Design Conceptual Framework for Intelligent Systems: System Purpose, Risk Context, Stakeholder Mapping, Data Integrity, Data Provenance, Model Reliability, Robustness, Uncertainty, Explanation Quality, Human Interpretability, Human Control, Accountability, Traceability, Auditability, Lifecycle Monitoring, Incident Response, Governance Review, and Deployment Boundaries

Framework stage	Core purpose	Required input	Trustworthiness function	Potential output	Validation or governance need	Failure risk
System purpose and risk context	Define the system's purpose, environment, affected decisions, and relevant harms	Intended use, domain conditions, stakeholders, misuse scenarios	Anchors assessment to context	Purpose-and-risk statement	Domain and stakeholder review	Vague or expanding purpose
Stakeholder and responsibility mapping	Identify actors, interests, authority, and lifecycle duties	Organizational roles and affected-party analysis	Prevents responsibility gaps	Stakeholder and responsibility map	Institutional approval	Missing actors or token participation
Data-integrity assessment	Examine accuracy, coherence, completeness, and change control	Data assets, quality evidence, version records	Establishes input reliability	Integrity profile	Steward review	Corrupted or incomplete data
Provenance and lineage review	Reconstruct origin, transformations, and dependencies	Source and pipeline records	Supports tracing and reproduction	Provenance-lineage map	Coverage review	Broken evidence chain
Data documentation and	Examine documentation, ownership, access, privacy,	Dataset records and governance arrangements	Links data evidence to institutional authority	Data-governance package	Organizational and domain review	Documentation without stewardship

governance review	security, bias, and representation					
Model-reliability assessment	Evaluate consistency, validity, and errors	Evaluation data, metrics, subgroup and case results	Establishes bounded performance evidence	Reliability profile	Separated validation where appropriate	Aggregate metrics hide severe errors
Robustness and uncertainty assessment	Examine disturbance, OOD behaviour, uncertainty, and calibration	Threat models, shifted data, UQ methods	Defines fragility and epistemic limits	Robustness and uncertainty profile	Method and scenario review	False precision or narrow testing
Bias and fairness review	Identify patterned disparities and normative choices	Group, impact, and stakeholder evidence	Makes distributional effects visible	Fairness argument	Normative and empirical review	Metric-based fairness theatre
Explanation-quality assessment	Test fidelity, usefulness, stability, and completeness	Model, explanation method, audience and task	Evaluates explanatory evidence	Explanation-quality profile	Technical and human-centered assessment	Persuasive but misleading accounts
Human-interpretability review	Determine whether intended users understand outputs and limits	User profiles, interface, task scenarios	Validates comprehension and reliance	Interpretability evidence	Target-user evaluation	Readability mistaken for understanding
Human-control and oversight design	Allocate authority, supervision, intervention, and fallback	Decision boundaries, capabilities, workload, roles	Preserves effective human agency	Control architecture	Operational testing	Nominal oversight
Contestability and recourse design	Enable challenge, correction, reconsideration, and remedy	Decision process, explanations, review authority	Makes accountability accessible	Contestation process	Affected-stakeholder review	Formal but ineffective appeal
Institutional-accountability mapping	Identify the institution and actors answerable for outcomes	Responsibility, authority, evidence, escalation	Prevents responsibility displacement	Accountability map	Senior institutional ownership	Policies without capacity

Traceability and auditability design	Preserve inspectable links among evidence, decisions, and changes	Logs, versions, records, approvals	Supports scrutiny and investigation	Evidence map	Access, competence and follow-up review	Missing or selective evidence
Lifecycle-monitoring plan	Define indicators, responsibilities, alerts, and responses	Risk profile and operating conditions	Detects emerging problems	Monitoring plan	Actionability review	Metrics without response
Drift monitoring	Detect changing inputs and model–outcome relationships	Baselines, live data, outcomes	Maintains validity awareness	Drift assessment	Context and label review	False alarms or missed degradation
Incident-response planning	Prepare containment, investigation, communication, and learning	Hazard scenarios, roles, evidence access	Limits harm and supports accountability	Incident plan	Exercises and post-incident review	Evidence loss or repeated failure
Governance review and updating	Reassess assumptions, responsibilities, evidence, and boundaries	Monitoring, incidents, audits, stakeholder feedback	Keeps framework aligned with change	Updated governance record	Authorized, version-controlled review	Static or ceremonial governance
Sufficiency and readiness boundaries	Prevent isolated evidence from being treated as complete assurance	Cross-layer evidence and residual limitations	Constrains unsupported claims	Qualified framework statement	Independent contextual review	Premature compliance or readiness claims

Figure 2 presents the proposed trustworthy-by-design framework for intelligent systems, integrating system purpose, risk context, stakeholder mapping, data provenance, model reliability, explanation quality, human control, institutional accountability, lifecycle monitoring, and governance updating.

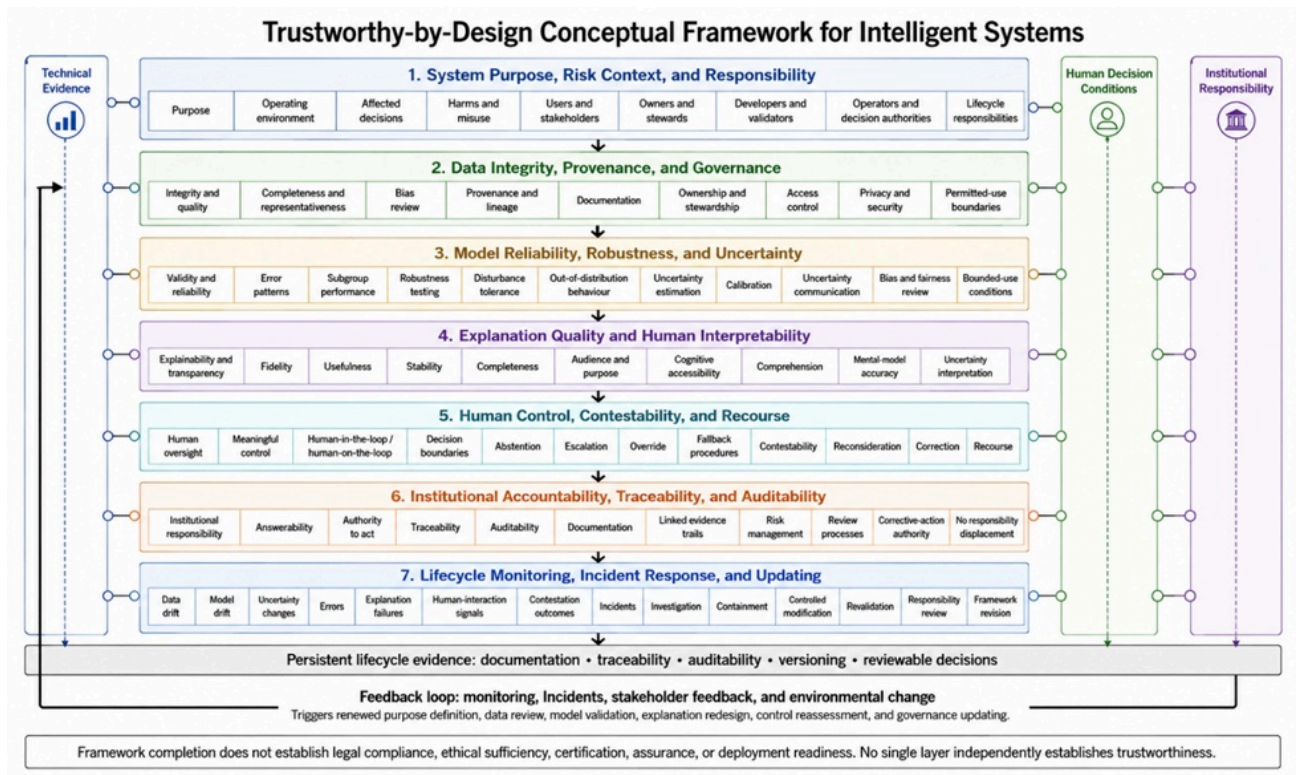


Figure 2. Trustworthy-by-Design Conceptual Framework for Intelligent Systems

Alt-text description:

A conceptual trustworthy-by-design framework showing intelligent system purpose, risk context, stakeholder mapping, data provenance, model reliability, uncertainty, explanation quality, human control, accountability, traceability, auditability, lifecycle monitoring, governance review, and framework updating.

Practical, Governance, and Research Implications

For system designers, data stewards, model developers, and platform providers, the framework highlights the need to preserve connections among purpose, evidence, responsibility, and change. Data documentation should remain linked to provenance, lineage, access, representation, and downstream use rather than exist as a static record. Reliability assessment should combine validity, errors, uncertainty, robustness, bias, and drift instead of optimizing an isolated metric. Platform-level research should examine how shared infrastructure affects evidence portability, logging, versioning, access, and responsibility handoffs across organizations [10].

Decision-support operators and end users require information that supports appropriate reliance, error recognition, escalation, and contestation. Explanation evaluation should therefore examine decision usefulness and misunderstanding rather than only technical properties. Human-control research should test whether people possess effective authority, sufficient time, appropriate competence, and usable intervention mechanisms under realistic workload and uncertainty. Affected stakeholders should be involved where their knowledge or exposure is relevant, but participation should not be represented as universal approval or as a substitute for institutional responsibility [23].

Institutional governance bodies and auditors should treat documentation, traceability, and auditability as components of an evidence process rather than proof of compliance or ethical adequacy. Research is needed on audit scope, evidence quality, reviewer independence, corrective-action follow-through, organizational incentives, and the consequences of

unresolved findings. Accountability mapping should identify who can obtain evidence, demand explanations, authorize restrictions, remedy affected decisions, and revise system arrangements. Assurance-oriented evaluation may structure qualified trustworthiness arguments, but it must preserve residual uncertainty and opportunities for challenge [26].

Future research should investigate cross-layer interactions through longitudinal and socio-technical evaluation. Priorities include maintainable provenance records, joint reliability and fairness assessment, usable uncertainty communication, explanation fidelity and comprehension, meaningful-control evaluation, drift and incident governance, and methods for updating responsibilities as systems change. These are proposed conceptual implications and research directions rather than legal advice, compliance guidance, certification guidance, procurement advice, model-selection advice, platform-selection advice, engineering instructions, institutional-governance advice, or deployment instructions. Their relevance and implementation require contextual empirical assessment [30].

Conclusion

Trustworthiness should be designed and evaluated across the complete lifecycle of intelligent systems rather than inferred from isolated model performance, explanation availability, human presence, or governance language. The proposed trustworthy-by-design framework connects system purpose and risk context with stakeholder responsibility, data integrity and provenance, model reliability, robustness and uncertainty, explanation quality and human interpretability, human control, contestability, institutional accountability, traceability, auditability, monitoring, incident response, and framework updating. Its central contribution is to clarify the dependencies and boundaries among these dimensions: provenance cannot establish data appropriateness, robustness cannot prove safety, explanations cannot guarantee understanding, human involvement cannot automatically create meaningful control, and institutional procedures cannot independently guarantee accountability. The framework may support conceptual analysis and assurance-oriented planning, but trustworthiness remains a context-dependent claim requiring empirical validation, human-centered evaluation, domain knowledge, institutional authority, continuing monitoring, and the capacity to revise or restrict system use when evidence changes.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Jan 2025

Revised: 08 Mar 2025

Accepted: 06 Apr 2025

Published online: 15 September 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

- 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>.
- 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 (Dordr)*. 2018;28(4):689-707. <https://doi.org/10.1007/s11023-018-9482-5>.
- 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>.
- 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>.
- 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>.
- 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>.
- Stahl BC, Andreou A, Brey P, Hatzakis T, Kirichenko A, Macnish K, et al. Artificial intelligence for human flourishing—beyond principles for machine learning. *J Bus Res*. 2021;124:374-88. <https://doi.org/10.1016/j.jbusres.2020.11.030>.
- Bender EM, Friedman B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans Assoc Comput Linguist*. 2018;6:587-604. https://doi.org/10.1162/tacl_a_00041.
- Herschel M, Diestelkämper R, Ben Lahmar H. A survey on provenance: what for? What form? What from? *VLDB J*. 2017;26(6):881-906. <https://doi.org/10.1007/s00778-017-0486-1>.
- Roh Y, Heo G, Whang SE. A survey on data collection for machine learning: a big data–AI integration perspective. *IEEE Trans Knowl Data Eng*. 2021;33(4):1328-47. <https://doi.org/10.1109/TKDE.2019.2946162>.
- Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: organizing data for trustworthy artificial intelligence. *Gov Inf Q*. 2020;37(3):101493. <https://doi.org/10.1016/j.giq.2020.101493>.
- Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506. <https://doi.org/10.1007/s10994-021-05946-3>.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion*. 2021;76:243-97.
<https://doi.org/10.1016/j.inffus.2021.05.008>.

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>.

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>.

Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. *IEEE Trans Knowl Data Eng*. 2019;31(12):2346-63.
<https://doi.org/10.1109/TKDE.2018.2876857>.

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>.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Benetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence: 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>.

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>.

Calvert SC. Principles and framework for the operationalisation of meaningful human control over autonomous systems. *Sci Eng Ethics*. 2025;31(5):27.
<https://doi.org/10.1007/s11948-025-00554-z>.

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>.

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>.

Busuioc M. Accountable artificial intelligence: holding algorithms to account. *Public Adm Rev*. 2021;81(5):825-36.
<https://doi.org/10.1111/puar.13293>.

Mökander J, Morley J, Taddeo M, Floridi L. Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Sci Eng Ethics*. 2021;27(4):44.
<https://doi.org/10.1007/s11948-021-00319-4>.

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>.

Peters D, Vold K, Robinson D, Calvo RA. Responsible AI—two frameworks for ethical design practice. *IEEE Trans Technol Soc*. 2020;1(1):34-47.

<https://doi.org/10.1109/TTS.2020.2974991>.

Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L. Ethics as a service: a pragmatic operationalisation of AI ethics. *Minds Mach (Dordr)*. 2021;31(2):239-56.

<https://doi.org/10.1007/s11023-021-09563-w>.

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>.