

ORIGINAL RESEARCH

Open access

When Does an Intelligent System Become Trustworthy? A Maturity Model for Transparency, Reliability, Human Oversight, Governance Integration, and Continuous Assurance

Clara Novak^{1*}, Ahmed El-Sayed², Julia Berg³

Abstract

Trustworthiness is frequently treated as a label attached to an intelligent system once selected technical, ethical, or governance requirements have been addressed. Such a binary interpretation is inadequate because system properties, operating conditions, organizational responsibilities, stakeholder expectations, and available assurance evidence change throughout the lifecycle. This article proposes a conceptual trustworthiness maturity model that interprets trustworthiness as an evolutionary system capability rather than a fixed attribute, certification outcome, or deployment decision. The model organizes progression across six levels: unstructured or ad hoc practices; foundational transparency and documentation; managed reliability, explainability, and human oversight; defined governance integration and accountability; continuous assurance and lifecycle monitoring; and adaptive trustworthiness learning and system updating. Foundational maturity requires accessible system, model, data, dataset, and decision documentation together with provenance, traceability, auditability, and maintained evidence trails. Higher capability requires context-sensitive evidence of reliability, validity, robustness, uncertainty, calibration, explanation quality, and meaningful human intervention. Advanced maturity additionally depends on responsibility mapping, organizational and institutional accountability, lifecycle governance, continuous monitoring, drift review, incident learning, assurance review, and controlled feedback updating. The principal contribution is an integrated maturity structure that connects technical, human, organizational, and lifecycle capabilities while retaining explicit boundaries against overclaiming. Maturity levels may help structure assessment planning, identify capability gaps, and clarify development priorities, but they do not independently prove trustworthiness, safety, ethical sufficiency, legal compliance, certification eligibility, or deployment readiness. The proposed model therefore supports conceptual maturity assessment and future empirical validation rather than operational prescription.

Keywords Trustworthy AI, Transparency, Reliability, Human oversight, Maturity model, Trustworthiness maturity

*Correspondence:

Clara Novak
clara.novak@wu.ac.at

¹ Department of Trustworthy AI and Maturity Models, Vienna University of Economics and Business, Vienna, Austria

² Department of Computer Science and AI Governance, American University in Cairo, Cairo, Egypt

³ Department of Organization and Management, Freie Universität Berlin, Berlin, Germany

Introduction

The question of when an intelligent system becomes trustworthy cannot be answered by identifying a single technical characteristic, ethical principle, organizational policy, or user perception. Trustworthy AI scholarship instead presents

trustworthiness as a multidimensional concern involving transparency, technical robustness, safety, human agency, accountability, fairness, privacy, and institutional responsibility [1–3]. These dimensions interact within socio-technical environments in which system outputs are interpreted, acted upon, contested, and revised. Trustworthiness must therefore be evaluated through the quality and coordination of capabilities surrounding an intelligent system rather than inferred from one favorable property.

Explainability illustrates the limitations of property-based reasoning. An explainable model may reveal influential features, provide local reasons, or support technical inspection, yet those functions do not establish that the underlying data are appropriate, the model is reliable, the explanation is faithful, or the organizational use is accountable [4]. Similar limitations affect transparency, fairness, robustness, and human oversight when each is treated as an isolated achievement. A system can disclose extensive information while remaining difficult to challenge, achieve stable performance while producing unacceptable harms, or include a nominal human reviewer who lacks the information, authority, time, or competence required for meaningful intervention.

A maturity perspective addresses this fragmentation by asking how relevant capabilities are established, maintained, integrated, and updated. Under this perspective, trustworthiness is not reached at a single moment. It develops through progressively stronger evidence concerning system purpose, documentation, data and model provenance, reliability, explanation quality, human control, governance, monitoring, and lifecycle learning. Progression is neither automatically linear nor irreversible. Material changes in data, models, users, organizational structures, or operating environments may weaken previously supported claims and require renewed assessment.

This article proposes an original conceptual trustworthiness maturity model for intelligent systems. Its objective is to organize capability development from unstructured practices through foundational evidence, managed technical and human capabilities, governance integration, continuous assurance, and adaptive learning. The contribution is not a universal scoring instrument, compliance framework, certification scheme, or deployment-readiness checklist. Rather, it provides a structured language for examining what evidence exists, what capabilities remain weak, how different dimensions depend on one another, and why no maturity level can independently prove that an intelligent system is trustworthy.

Trustworthiness as an Evolutionary System Capability

Trust must first be distinguished from trustworthiness. Trust can describe an attitude held by a person or institution, whereas trustworthiness concerns whether available evidence justifies reliance on a system within a defined context. Reliability may support justified reliance, but anthropomorphic language can obscure the fact that an intelligent system does not possess moral intentions or interpersonal commitments in the manner expected of a human trustee [5]. A maturity model should therefore organize evidence-bearing system and organizational capabilities rather than attempt to measure whether users feel confident, reassured, or willing to rely on an AI output.

Trustworthiness maturity also extends beyond statements of responsible or ethical AI principles. High-level commitments can identify desirable values, but principles alone cannot determine how competing values should be interpreted, translated into system requirements, assigned to responsible actors, or evaluated in a specific institutional environment [6]. Capability progression begins when such commitments are converted into maintained practices, evidence, review responsibilities, and decision boundaries. Maturity consequently concerns the institutionalization and continuing effectiveness of trustworthiness-related work, not the number or rhetorical strength of principles an organization endorses.

Operational tools can support this transition by making abstract responsibilities more repeatable. Documentation templates, impact-assessment methods, fairness evaluations, explanation procedures, review protocols, and governance mechanisms may help organizations move from informal intention toward managed practice [7]. Their existence, however, is insufficient evidence of effective use. A mature capability requires clarity about purpose, ownership, inputs,

limitations, review frequency, escalation, and the consequences of identified deficiencies. The relevant question is not whether a tool is available, but whether its use produces credible evidence that informs system and governance decisions.

Trustworthiness should ultimately be interpreted as an ecosystem capability involving technical components, human actors, organizational processes, affected stakeholders, and external conditions [8]. Intelligent systems are embedded in decision-support arrangements, autonomous functions, platforms, data supply chains, professional workflows, and institutional mandates. Changes in any of these elements can alter the validity of earlier conclusions. Continuous assurance, lifecycle governance, monitoring, and feedback updating are therefore advanced expressions of maturity, but they remain provisional. No maturity assignment guarantees trustworthiness, and no level independently authorizes system deployment.

Foundational Transparency and Documentation

Transparency is a necessary foundation because assessment is impossible when relevant system properties, assumptions, dependencies, and decision roles remain inaccessible. Nevertheless, transparency should not be equated with accountability. Disclosure can reveal selected technical information while leaving institutional relationships, decision authority, data limitations, or practical consequences obscure [9]. Early maturity therefore requires purposeful transparency: relevant information must be accurate, accessible to its intended audience, connected to the decisions it informs, and maintained as the system changes.

System documentation should explain purpose, intended use, exclusions, architecture, dependencies, interfaces, and known constraints. Model documentation should record the model's origin, design assumptions, training or adaptation process, expected operating conditions, and material limitations. Decision documentation should enable reconstruction of how model outputs, human judgments, and institutional procedures contributed to consequential actions. These records support process-level transparency, traceability, and accountability, but only when they describe actual practices rather than idealized workflows [10].

Data and dataset documentation require particular attention because downstream system behavior depends on choices made during collection, selection, cleaning, labeling, aggregation, and reuse. Documentation should record provenance, development context, transformations, quality controls, known exclusions, representational limitations, and conditions governing later updates. Dataset-development scholarship shows that apparently technical data choices also embed institutional priorities and social assumptions that can shape validity and bias [11]. Data provenance should therefore be assessed as an evidentiary chain rather than a simple source label.

Documentation becomes maturity-relevant when it supports inquiry, auditability, contestability, and lifecycle intervention. Evidence trails should connect requirements, identified risks, system versions, evaluation results, responsible actors, operational incidents, and approved changes. Explanation accessibility may supplement this foundation when users require understandable reasons or routes for challenge. Ethics-based auditing research nevertheless indicates that documents and access mechanisms must be connected to appropriate intervention points and governance capacity if they are to influence practice [12]. Documentation alone does not establish transparency quality, responsible behavior, or trustworthiness. **Table 1** summarises foundational transparency and documentation capabilities required for early trustworthiness maturity in intelligent systems.

Table 1. Foundational Transparency and Documentation Capabilities for Trustworthiness Maturity: System Documentation, Model Documentation, Data Documentation, Decision Documentation, Dataset Documentation, Documentation Quality, Data Provenance, Model Provenance, Decision Provenance, Traceability, Auditability, Evidence Trails, Explanation Accessibility, and Documentation Boundaries

Foundational capability	Core maturity question	Required evidence or practice	Maturity function	Potential output	Failure or overclaiming risk	Assessment need
Transparency	What material system information is available to relevant stakeholders?	Purpose, scope, limitations, dependencies, decision role, and access arrangements	Establishes baseline inspectability	Transparency record	Selective disclosure may create false confidence	Relevance, completeness, and accessibility review
Transparency maturity	Is transparency maintained, repeatable, and audience-appropriate?	Defined ownership, update cycles, access rules, and communication formats	Converts disclosure into a managed capability	Transparency capability profile	More information may be mistaken for better transparency	Currency and stakeholder-usefulness evaluation
System documentation	Is the intelligent system documented beyond the model alone?	Architecture, components, interfaces, intended use, exclusions, and dependencies	Supports system-level understanding	System dossier	Model-centric records may omit important infrastructure and workflow risks	Cross-component consistency review
Model documentation	Are model assumptions, development choices, intended uses, and limitations recorded?	Model origin, design, training or adaptation process, versions, and known failure conditions	Supports technical scrutiny	Model record	Static or promotional documentation may conceal uncertainty	Accuracy and version-control review
Data documentation	Are data sources, collection contexts, transformations, and limitations described?	Source descriptions, collection methods, preprocessing, exclusions, and access conditions	Supports data-quality reasoning	Data record	Technical descriptions may omit social and institutional context	Provenance and suitability assessment
Decision documentation	Can consequential decisions involving the system be reconstructed?	Relevant inputs, model version, output, human intervention, rationale, and outcome pathway	Supports contestability and review	Decision log	Logs may omit informal judgments or imply unsupported causality	Reconstructability and proportionality review

Dataset documentation	Are dataset-development and maintenance choices recorded?	Selection, annotation, composition, quality checks, gaps, and update history	Makes limitations examinable	Dataset statement	Checklist completion may be mistaken for dataset validity	Content-quality and context review
Documentation quality	Are records accurate, current, internally consistent, and usable?	Ownership, review dates, versioning, evidence links, and correction procedures	Converts documents into dependable evidence	Documentation-quality finding	Polished records may not correspond to practice	Verification against system behavior
Data provenance	Can material data lineage be traced?	Sources, collection authority, transformations, reuse conditions, and version history	Supports error investigation and responsibility attribution	Data-lineage map	Broken lineage may remain hidden behind summary descriptions	Sampling and lineage verification
Model provenance	Can model origin and modification history be reconstructed?	Base model, components, training stages, fine-tuning, dependencies, and releases	Supports version accountability	Model-lineage record	Proprietary or third-party dependencies may interrupt traceability	Dependency and change review
Decision provenance	Can links among data, model output, human action, and institutional decision be traced?	Time-stamped decision chain and recorded interventions	Supports inquiry and contestability	Decision-provenance trail	Records may oversimplify distributed decision-making	Case-based reconstruction
Traceability	Can requirements, risks, design decisions, tests, and operational evidence be linked?	Maintained requirement-to-evidence relationships	Integrates lifecycle evidence	Traceability matrix	Superficial links may become bureaucratic artifacts	Completeness and claim-fit assessment
Auditability	Can an appropriately authorized reviewer examine relevant	Evidence retention, access rights, review criteria, and independence safeguards	Enables structured scrutiny	Auditable evidence package	Access alone does not guarantee an effective audit	Scope, competence, and independence review

	evidence and processes?					
Evidence trails	Is there a coherent record supporting material trustworthiness claims?	Version-controlled artifacts, tests, decisions, incidents, approvals, and changes	Supports assurance reasoning	Evidence register	Evidence quantity may be confused with evidential sufficiency	Claim–evidence alignment review
Explanation accessibility	Can intended users obtain explanations in usable form when required?	Appropriate timing, language, format, and escalation routes	Connects transparency to human understanding	Explanation-access record	Simplification may reduce fidelity or omit uncertainty	User-centered comprehension evaluation
Transparency boundary	Is it explicit that transparency alone does not prove accountability or trustworthiness?	Cross-domain qualification and limitation statement	Prevents single-factor conclusions	Qualified transparency finding	Disclosure may be marketed as responsible behavior	Overclaiming review
Documentation boundary	Is it explicit that documentation alone does not prove system quality or responsible practice?	Verification against actual behavior and organizational use	Prevents paper-based maturity claims	Qualified documentation finding	Records may diverge from operations	Evidence-to-practice verification

Reliability, Explainability, and Human Oversight

Foundational evidence makes higher-order evaluation possible, but trustworthiness maturity requires more than reconstructability. Reliability, uncertainty, and explanation quality must be evaluated in relation to the system's intended task, operating environment, users, and consequences. Uncertainty-quantification research demonstrates the importance of representing incomplete confidence in deep-learning outputs, while explainability scholarship shows that explanation assessment must address both technical properties and human use [13–15]. Level 2 capability is therefore defined by managed evaluation rather than the mere presence of a high-performing model, an explanation method, or a human reviewer.

Reliability concerns whether a system behaves consistently under stated conditions, whereas validity concerns whether available evidence supports the interpretation and use being made of its outputs. Robustness testing examines performance under plausible perturbations, shifts, missing information, adversarial conditions, or changes in system dependencies. Safety extends further by considering the relationship between system behavior and foreseeable harm. These concepts overlap but should not be collapsed. A reliable model can remain invalid for a particular decision context, and a robust model can remain unsafe when its use, objectives, or institutional controls are inappropriate.

Uncertainty estimation and calibration qualify reliability claims by showing how confidence relates to observed correctness and how that relationship changes across groups, tasks, or time. Mature practice also requires uncertainty communication because technically valid estimates may be misunderstood by operators or end users. Error analysis should examine not only frequency but also type, severity, concentration, and operational consequence. Bias and fairness assessment may be required where system design or use creates differential effects, but selected fairness definitions and metrics must be justified contextually rather than treated as universally interchangeable [13].

Explainability maturity concerns the quality and fitness of explanations rather than their simple availability. Relevant criteria can include fidelity to model behavior, stability under small changes, completeness relative to the question, accessibility, and usefulness for a specified audience. An explanation suitable for a model developer may not assist a decision-support operator, affected individual, governance body, or auditor. Because explainability definitions and evaluation approaches remain heterogeneous, maturity assessment should make the purpose, audience, method, limitations, and evaluation evidence explicit [14]. Explanation quality supports trustworthiness reasoning but does not compensate for unreliable models, inappropriate data, or deficient governance.

Human oversight must likewise be distinguished from nominal human presence. Meaningful control requires that responsible people can understand relevant information, recognize when intervention is needed, possess the authority and practical capacity to intervene, and remain traceable as responsible actors. The tracking and tracing conditions proposed for meaningful human control help clarify why a human-in-the-loop arrangement can remain ineffective when system behavior is poorly aligned with human reasons or responsibility cannot be attributed [16]. Oversight maturity may therefore include competence, workload, timing, escalation, contestability, recourse, and override capacity where relevant. Neither technical reliability nor human supervision independently establishes safety, ethical sufficiency, accountability, or trustworthiness.

Governance Integration and Continuous Assurance

Advanced trustworthiness maturity requires technical and human capabilities to influence organizational authority, resource allocation, risk acceptance, and system-change decisions. The movement from ethical aspiration toward AI governance requires explicit institutional arrangements; organizational evidence further indicates that governance depends on coordinated practices, management commitment, and the ability to connect principles with operational responsibilities; lifecycle-assurance scholarship similarly shows that evidence must be developed and maintained across system conception, data preparation, model development, evaluation, deployment, operation, and change [17–19]. Governance integration therefore concerns whether trustworthiness evidence can alter consequential decisions, not merely whether an ethics statement, advisory committee, or policy document exists.

Lifecycle governance should assign responsibility for system purpose, data stewardship, model development, operational use, risk review, monitoring, incident response, and controlled updating. Responsibility mapping is especially important where multiple developers, platform providers, operators, professional users, and organizational bodies contribute to outcomes. Risk management should document identified hazards, responsible owners, controls, unresolved uncertainty, and circumstances requiring escalation or reassessment. Monitoring must also account for changing data-generating conditions because concept drift can weaken previously supported performance and validity claims [20]. The absence of detected drift does not establish that all relevant system or social changes have been captured.

Continuous assurance describes the recurring collection, review, and qualification of evidence rather than a permanent declaration that a system is trustworthy. Relevant evidence may include model validation, robustness tests, uncertainty analysis, explanation evaluation, human-oversight findings, governance decisions, operational monitoring, complaints, incidents, and system modifications. An assurance case may help organize claims, arguments, evidence, limitations, and unresolved challenges where appropriate. Ethics-based auditing can similarly provide structured scrutiny, but its effectiveness depends on scope, access, competence, independence, organizational responsiveness, and recognition of its limitations [21]. Auditability is therefore a capability for examination, not proof of ethical sufficiency or certification.

Operational maturity also requires mechanisms for incident reporting, investigation, containment, and learning where the system context supports such practices. Incident response should examine technical and organizational causes rather than assume that failure originates solely in a model component. Feedback updating may incorporate monitoring results, operator experience, stakeholder complaints, audit findings, and changes in the operating environment, but updates should remain controlled, documented, and subject to renewed evaluation. Unrestricted adaptation can weaken traceability, change responsibility relationships, or introduce risks not represented in earlier evidence.

Governance review and stakeholder participation strengthen continuous assurance by testing whether institutional assumptions remain appropriate. Affected stakeholders may reveal harms, accessibility problems, contested values, or barriers to recourse that technical monitoring does not capture. Participation should nevertheless be evaluated for influence, representation, response quality, and protection against tokenism. Governance integration does not automatically create accountability, and continuous assurance cannot eliminate uncertainty or future harm. Both capabilities remain dependent on institutional authority, competent evaluation, context-sensitive judgment, and continuing willingness to act on adverse evidence.

Trustworthiness Levels and Capability Progression

The proposed model contains six conceptual levels: Level 0, unstructured or ad hoc trustworthiness practices; Level 1, foundational transparency and documentation; Level 2, managed reliability, explainability, and oversight; Level 3, defined governance integration and accountability; Level 4, continuous assurance and lifecycle monitoring; and Level 5, adaptive trustworthiness learning and system updating. Responsible-AI maturity research supports integrating technical and organizational capabilities within a socio-technical progression rather than treating maturity as model performance alone [22]. These levels organize the quality, integration, and maintenance of evidence; they do not certify trustworthiness, compliance, safety, ethical sufficiency, or deployment readiness. Figure 1 illustrates the proposed trustworthiness capability progression from unstructured practices to adaptive trustworthiness learning, showing that maturity levels organize evidence and governance capability rather than certifying system trustworthiness.

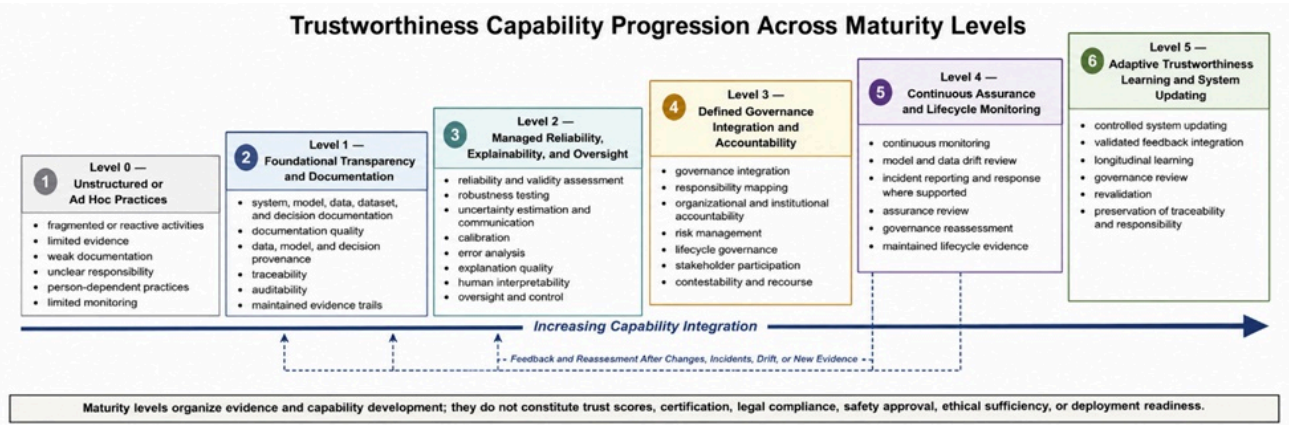


Figure 1. Trustworthiness Capability Progression Across Maturity Levels

Alt-text description:

A maturity progression diagram showing six conceptual levels of trustworthiness capability, moving from ad hoc practices through transparency, reliability, governance, continuous assurance, and adaptive trustworthiness learning.

Level 0 indicates that relevant activities are fragmented, reactive, inconsistently performed, or dependent on individual initiative. Level 1 begins when system purpose, model characteristics, data origins, decision pathways, limitations, and

evidence ownership are documented and maintained. General AI maturity scholarship supports using clearly defined dimensions and stages to structure incremental capability development [23]. In the present model, however, early documentation maturity is only an evidentiary foundation; it does not show that the documented system behaves reliably or is governed responsibly.

Level 2 adds managed technical and human capabilities. Reliability, validity, robustness, uncertainty, explanation quality, and oversight are assessed against intended use and operating conditions. Level 3 integrates these findings into governance, risk ownership, accountability, auditability, stakeholder participation, and escalation. Domain-specific maturity-model development demonstrates the value of explicit capability dimensions, indicators, and progression requirements, although such structures remain context-dependent and should not be generalized as universal proof of readiness [24].

Level 4 maintains evidence through continuous monitoring, drift review, incident learning, governance reassessment, and recurring assurance. Level 5 adds controlled adaptation in which validated technical, operational, organizational, and stakeholder feedback can change the system and its governance while preserving traceability. Readiness frameworks may help organizations characterize conditions for AI-enabled transformation, but readiness should remain distinct from trustworthiness maturity [25]. A system can possess substantial adoption capability while retaining unresolved reliability, accountability, or assurance weaknesses. **Table 2** maps reliability, explainability, human oversight, governance integration, continuous assurance, and trustworthiness capability progression across the proposed maturity levels.

Table 2. Trustworthiness Capability Progression Across Maturity Levels: Reliability, Robustness, Uncertainty, Explainability, Human Oversight, Human Control, Governance Integration, Accountability, Continuous Monitoring, Incident Response, Feedback Updating, and Maturity Boundary Conditions

Maturity level	Capability description	Core capability question	Required evidence or practice	Dominant trustworthiness function	Typical limitation	Progression requirement
Level 0 — Unstructured or ad hoc trustworthiness practices	Activities are fragmented, reactive, undocumented, or person-dependent.	Are repeatable trustworthiness practices and responsible actors identifiable?	Initial system, risk, actor, and evidence inventory	Recognition of capability gaps	Weak documentation, unclear ownership, and limited monitoring	Establish ownership, transparency, documentation and evidence retention
Level 1 — Foundational transparency and documentation	System, model, data, dataset, decision, provenance, and traceability records are maintained.	Can relevant actors understand the system's purpose, limitations, evidence, and decision role?	Accessible documentation, provenance, version control, traceability, and auditability	Inspectability and reconstructability	Records may be inaccurate, inaccessible, static, or disconnected from practice	Verify documentation and add technical, explanation, and oversight evaluation
Level 2 — Managed reliability, explainability, and oversight	Technical behavior, uncertainty, explanations, and human intervention are	Is there credible context-specific evidence supporting	Validity, reliability, robustness, calibration, error, bias, explanation,	Managed technical and human control	Tests may be narrow; explanations may mislead; oversight may be nominal	Integrate evidence with governance, accountability, and risk ownership

	evaluated and managed.	system behavior and human supervision?	and oversight evaluation			
Level 3 — Defined governance integration and accountability	Trustworthiness responsibilities influence institutional decisions and risk management.	Are authority, responsibility, escalation, participation, and review connected to evidence?	Governance mandates, responsibility mapping, risk records, auditability, contestability, and recourse	Institutionalized accountability	Governance may lack authority, resources, independence, or stakeholder influence	Link governance to monitoring, incidents, assurance, and corrective action
Level 4 — Continuous assurance and lifecycle monitoring	Evidence is maintained through monitoring, drift review, incidents, auditing, and governance reassessment.	Can emerging changes and failures be detected, investigated, and acted upon?	Continuous monitoring, drift analysis, incident processes, assurance review, and retained evidence	Ongoing lifecycle assurance	Unknown harms may remain undetected; monitoring may not trigger action	Develop controlled adaptation and validated feedback updating
Level 5 — Adaptive trustworthiness learning and system updating	Validated feedback informs controlled changes to systems and governance.	Can learning occur without losing traceability, responsibility, and review?	Change control, revalidation, stakeholder deliberation, governance review, and evidence updating	Accountable lifecycle learning	Adaptation can introduce new uncertainty, value conflict, and responsibility gaps	Continue domain-specific reassessment after every material change

Proposed Trustworthiness Maturity Model

The proposed trustworthiness maturity model begins with system-purpose and risk-context definition. Assessment must identify the intended decisions, affected parties, operating environment, system boundaries, material dependencies, and foreseeable consequences before capability can be interpreted. Stakeholder and responsibility mapping then identifies who develops, supplies, governs, operates, relies upon, or is affected by the system. Corporate responsible-AI research demonstrates why declared commitments should be separated from evidence that principles have been implemented in operational practice [26]. The model consequently evaluates observable capabilities and maintained evidence rather than institutional rhetoric.

The second assessment sequence examines foundational transparency, documentation, and provenance. System, model, data, dataset, and decision records are evaluated for accuracy, accessibility, ownership, currency, and consistency with actual practice. Data and model lineage should enable reconstruction of material changes, while decision provenance should clarify how system output, human judgment, and institutional authority contributed to an outcome. Organizational research on operationalizing AI ethics similarly indicates that progression requires institutionalized

practices rather than isolated statements or informal initiatives [27]. Transparency and documentation findings remain qualified by explicit non-sufficiency boundaries.

The third sequence evaluates model reliability, validity, robustness, uncertainty, calibration, error patterns, bias, fairness where relevant, and explanation quality. The fourth examines human oversight, control, intervention authority, competence, workload, contestability, recourse, and meaningful-control conditions. Findings are not converted into universal numerical scores. Instead, they form a qualitative and evidence-linked characterization of capability strength, limitations, unresolved risks, and context dependence. A model may therefore exhibit strong documentation but weak uncertainty communication, or mature reliability testing but inadequate operator authority.

The fifth and sixth sequences assess governance integration, accountability, auditability, monitoring, drift, incidents, assurance review, and feedback updating. Governance assessment asks whether identified evidence can change decisions, trigger escalation, halt use, or require reassessment. Continuous-assurance assessment asks whether evidence remains current as systems and environments change. Stakeholder deliberation can contribute perspectives that are not visible through technical evaluation alone and can help make adaptive learning more responsive and contestable [28]. Such participation must nevertheless be assessed for representation, influence, response, and the risk of symbolic inclusion.

Maturity-level assignment is the model's cautious synthesis stage. An assignment should describe the best-supported capability state, unevenness across domains, evidence limitations, and conditions requiring renewed assessment. It should not be treated as a trust score, certification grade, compliance result, procurement rule, or release authorization. The final integrative step connects technical requirements, human agency, organizational accountability, and lifecycle governance while preserving the principle that no single dimension is sufficient [29]. Capability-progression planning may identify development priorities, but priorities remain domain-specific and subject to institutional judgment. **Table 3** presents the proposed trustworthiness maturity model for intelligent systems, linking assessment stages, maturity functions, evidence requirements, governance needs, feedback mechanisms, and decision boundaries.

Table 3. Proposed Trustworthiness Maturity Model for Intelligent Systems: System Purpose, Risk Context, Stakeholder Mapping, Transparency Assessment, Documentation Assessment, Data Provenance, Model Reliability, Robustness, Uncertainty, Explanation Quality, Human Oversight, Governance Integration, Accountability, Continuous Monitoring, Assurance Evidence, Capability Progression, Feedback Updating, and Deployment Boundaries

Maturity model stage	Core purpose	Required input	Maturity function	Potential output	Assessment or governance need	Failure risk
Purpose and risk-context definition	Establish assessment scope	Intended use, decisions, environment, dependencies, and foreseeable harms	Grounds all later judgments	Context statement	Domain and stakeholder review	Generic evaluation detached from use
Stakeholder and responsibility mapping	Identify affected and responsible actors	Actor, authority, dependency, and impact information	Prevents responsibility gaps	Stakeholder–responsibility map	Governance validation	Hidden actors and diffused responsibility

Transparency assessment	Determine whether material information is accessible	System disclosures and audience needs	Establishes inspectability	Transparency finding	Relevance and accessibility evaluation	Disclosure volume mistaken for quality	
Documentation assessment	Evaluate accuracy and maintenance of records	System, model, data, dataset, and decision documentation	Establishes evidence availability	Documentation profile	Verification against practice	Paper compliance	d
Data and model provenance assessment	Reconstruct lineage and modification	Source, transformation, component, and version records	Supports traceability	Provenance map	Steward and dependency review	Broken or proprietary lineage	e
Reliability and validity assessment	Examine dependable behavior and intended-use validity	Performance, failure, and contextual evidence	Supports qualified reliance	Reliability–validity finding	Domain-specific validation	Aggregate metrics conceal critical failure	R
Robustness and uncertainty assessment	Evaluate variation, incomplete knowledge, and confidence	Stress tests, uncertainty estimates, calibration, and error evidence	Qualifies technical claims	Robustness–uncertainty profile	Scenario and subgroup evaluation	Narrow testing or false precision	A fr
Explanation-quality assessment	Evaluate fidelity, stability, completeness, accessibility, and usefulness	Explanations and technical and human evaluation	Supports informed use and review	Explanation profile	Audience-specific testing	Plausible but misleading explanations	
Human oversight and control assessment	Determine whether intervention is meaningful	Roles, competence, information, timing, authority, and override capacity	Evaluates human agency	Oversight-and-control finding	Workflow and authority testing	Oversight theatre and automation bias	C
Governance integration assessment	Determine whether evidence influences	Mandates, risk ownership, decision rights, and escalation	Institutionalizes capability	Governance-integration finding	Effectiveness and authority review	Advisory structures without influence	

	institutional decisions						
Accountability and auditability assessment	Examine attribution, inquiry, and evidence access	Responsibility maps, records, review rights, and contestability routes	Supports scrutiny and recourse	Accountability profile	Independence and access assessment	Responsibility diffusion	c
Continuous monitoring and drift review	Detect material changes in inputs, relationships, performance, and use	Indicators, logs, baselines, and contextual evidence	Maintains operational validity	Monitoring and drift record	Technical and domain interpretation	Alerts without action or missed harms	
Incident-response readiness review	Examine reporting, containment, investigation, and learning where supported	Reporting routes, roles, procedures, and retained evidence	Supports failure response	Incident-readiness finding	Scenario evaluation	Underreporting or narrow root-cause analysis	P
Assurance evidence review	Integrate claims, evidence, limitations, and unresolved challenges	Technical, human, organizational, and lifecycle findings	Supports provisional assurance reasoning	Qualified assurance judgment	Multidisciplinary challenge	Formalism conceals weak evidence	n
Maturity-level assignment with caution	Summarize the evidenced capability state	Domain findings and uncertainty	Organizes progression	Qualified level characterization	Transparent rationale	False precision and level shopping	a
Capability-progression planning	Identify context-sensitive development priorities	Capability gaps, risks, resources, and institutional aims	Guides structured improvement	Provisional progression plan	Local ownership and review	Universalized or mechanically linear planning	o
Feedback updating	Incorporate validated operational and stakeholder learning	Monitoring, incidents, audits, complaints, and deliberation	Enables adaptive maturity	Controlled update decision	Change governance and revalidation	Uncontrolled adaptation	

Compliance and deployment boundaries	Prevent inappropriate use of maturity findings	Explicit qualifications	Controls legal and operational overclaiming	Boundary statement	Separate domain-specific processes	Maturity level used as approval	u
--------------------------------------	--	-------------------------	---	--------------------	------------------------------------	---------------------------------	---

Figure 2 presents the proposed trustworthiness maturity model for intelligent systems, integrating transparency, documentation, reliability, explainability, human oversight, governance, continuous assurance, capability progression, and feedback updating.



Figure 2. Trustworthiness Maturity Model for Intelligent Systems

Alt-text description:

A conceptual maturity model showing assessment stages for intelligent systems, including risk context, stakeholder mapping, transparency, documentation, data provenance, reliability, robustness, uncertainty, explanation quality, human oversight, governance, accountability, continuous monitoring, assurance evidence, capability progression, and feedback updating.

Assessment, Adoption, and Research Implications

For system designers and model developers, the model reframes trustworthiness as an evidence architecture rather than a final-stage review. Purpose, dependencies, documentation, uncertainty, explanation requirements, and intervention pathways should be considered early enough to influence design. Data stewards and platform providers face related questions concerning lineage, versioning, data suitability, third-party dependencies, and notification of material change. These implications are conceptual: they identify assessment domains but do not prescribe particular architectures, models, platforms, thresholds, or development procedures.

Governance bodies and auditors where supported can use the model to identify whether evidence, responsibility, authority, and lifecycle review are connected. Assessment should examine governance effectiveness rather than the existence of policies or committees. Ethics-based auditing may support structured examination, but it remains constrained by evidence access, competence, independence, scope, and institutional response [21]. The model does not provide audit instructions, certification criteria, legal interpretations, or governance structures that are universally appropriate.

Decision-support operators, end users, and affected stakeholders require evaluation of explanation accessibility, uncertainty communication, oversight authority, contestability, recourse, and participation. Human-centered explanation research suggests that explanation usefulness depends on the questions, expectations, and cognitive practices of intended recipients [15]. Future studies should therefore examine whether explanations improve calibrated reliance, whether operators can intervene under realistic conditions, and whether affected people can meaningfully challenge or influence system-supported decisions.

Research priorities include empirical validation of the six levels, domain-specific adaptation, inter-rater consistency, longitudinal capability change, and the relationship between maturity characterization and observed system or institutional outcomes. Continuous-assurance research should examine how evidence becomes stale, how drift and incidents alter earlier claims, and how assurance arguments should be reopened after material change [19]. Responsible-AI maturity research should also investigate uneven capability profiles rather than assume that organizations progress uniformly across dimensions [22]. These are research directions and conceptual adoption implications, not compliance guidance, procurement advice, model-selection advice, legal advice, certification guidance, or deployment instruction.

Conclusion

This article has proposed a conceptual trustworthiness maturity model in which intelligent systems progress from unstructured practices through foundational transparency and documentation, managed reliability, explainability and human oversight, defined governance integration and accountability, continuous assurance and lifecycle monitoring, and adaptive trustworthiness learning and controlled updating. The model connects technical evidence with human agency, organizational responsibility, stakeholder participation, monitoring, incident learning, and lifecycle feedback. Its levels can help structure capability assessment, expose unevenness, and clarify development priorities, but they do not constitute proof that a system is trustworthy. Trustworthiness maturity must remain context-sensitive, evidence-based, provisional, and open to reassessment after changes in data, models, uses, actors, or environments. It does not independently establish legal compliance, ethical sufficiency, operational safety, certification eligibility, institutional accountability, user trust, or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Jul 2025

Revised: 09 Sep 2025

Accepted: 02 Nov 2025

Published online: 15 March 2026

Rights and permissions

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

References

- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-99.
<https://doi.org/10.1038/s42256-019-0088-2>.
- Floridi L, 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*. 2018;28(4):689-707.
<https://doi.org/10.1007/s11023-018-9482-5>.
- 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>.
- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bannetot 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>.
- 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>.
- 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>.
- Stahl BC. Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems. *Sci Rep*. 2023;13(1):7586.
<https://doi.org/10.1038/s41598-023-34622-w>.
- Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc*. 2018;20(3):973-89.
<https://doi.org/10.1177/1461444816676645>.
- Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. *Philos Technol*. 2018;31(4):611-27.
<https://doi.org/10.1007/s13347-017-0279-x>.

Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: a survey of dataset development and use in machine learning research. *Patterns* (N Y). 2021;2(11):100336.
<https://doi.org/10.1016/j.patter.2021.100336>.

Mökander J, Axente M. Ethics-based auditing of automated decision-making systems: intervention points and policy implications. *AI Soc.* 2023;38(1):153-71.
<https://doi.org/10.1007/s00146-021-01286-x>.

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>.

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>.

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>.

Santoni de Sio F, van den Hoven J. Meaningful human control over autonomous systems: a philosophical account. *Front Robot AI.* 2018;5:15.
<https://doi.org/10.3389/frobt.2018.00015>.

Koniakou V. From the "rush to ethics" to the "race for governance" in artificial intelligence. *Inf Syst Front.* 2023;25(1):71-102.
<https://doi.org/10.1007/s10796-022-10300-6>.

Papagiannidis E, Enholm IM, Dremel C, Mikalef P, Krogstie J. Toward AI governance: identifying best practices and potential barriers and outcomes. *Inf Syst Front.* 2023;25(1):123-41.
<https://doi.org/10.1007/s10796-022-10251-y>.

Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: desiderata, methods, and challenges. *ACM Comput Surv.* 2021;54(5):111:1-111:39.
<https://doi.org/10.1145/3453444>.

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>.

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>.

Akbarighatar P, Pappas I, Vassilakopoulou P. A sociotechnical perspective for responsible AI maturity models: findings from a mixed-method literature review. *Int J Inf Manag Data Insights.* 2023;3(2):100193.
<https://doi.org/10.1016/j.jjimei.2023.100193>.

Sadiq RB, Safie N, Abd Rahman AH, Goudarzi S. Artificial intelligence maturity model: a systematic literature review. *PeerJ Comput Sci.* 2021;7:e661.
<https://doi.org/10.7717/peerj-cs.661>.

Sonntag M, Mehmman S, Mehmman J, Teuteberg F. Development and evaluation of a maturity model for AI deployment capability of manufacturing companies. *Inf Syst Manag.* 2025;42(1):37-67.
<https://doi.org/10.1080/10580530.2024.2319041>.

Holmström J. From AI to digital transformation: the AI readiness framework. *Bus Horiz.* 2022;65(3):329-39.
<https://doi.org/10.1016/j.bushor.2021.03.006>.

de Laat PB. Companies committed to responsible AI: from principles towards implementation and regulation? *Philos Technol.* 2021;34(4):1135-93.
<https://doi.org/10.1007/s13347-021-00474-3>.

Camacho Ibáñez J, Villas Olmeda M. Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. *AI Soc.* 2022;37(4):1663-87.
<https://doi.org/10.1007/s00146-021-01267-0>.

Buhmann A, Fieseler C. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technol Soc.* 2021;64:101475.
<https://doi.org/10.1016/j.techsoc.2020.101475>.

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>.