# AI System Quality as Evidence-Centric Governance: Lifecycle Accountability for High-Risk AI Systems

Fatima Rabia Yapicioglu
University of Bologna, DISI
Automobili Lamborghini S.p.A.
fatima.yapicioglu2@unibo.it

Fabio Vitali
University of Bologna, DISI
fabio.vitali@unibo.it

Rami Mochaourab
Scania CV AB
rami.mochaourab@scania.com

Anna De Carvalho Guimarães
KTH Royal Institute of Technology
andcg@kth.se

**Abstract**

Artificial intelligence (AI) system quality is increasingly central to responsible AI governance, particularly under Article 17 of the EU AI Act, which requires providers of high-risk AI systems to establish a Quality Management System (QMS). Standards, conformity-assessment approaches, and verification methods increasingly operationalise these requirements, but the concrete evidence needed to assess AI system quality remains dependent on the system, intended purpose, risk, and deployment context. A complementary governance problem is therefore how quality claims concerning an individual AI system remain justified as the system and its context evolve. This paper addresses this problem by conceptualising AI system quality as *evidence-centric lifecycle governance*. It distinguishes AI system quality from conventional software and manufacturing quality by highlighting challenges linked to probabilistic behaviour, data dependency, context sensitivity, and temporal degradation. It then organises AI system quality concerns into four governance-oriented clusters: technical robustness and reliability, data and model integrity, human-facing intelligibility and oversight, and institutional accountability. Building on this structure, the paper develops a framework in which quality claims are linked to context-specific and auditable evidence across the AI system lifecycle, together with responsible actors, review and challenge, monitoring, and corrective or remedial action. The contribution is a generalisable governance structure for maintaining the evidential justification of quality claims concerning individual high-risk AI systems, while allowing the concrete evidence, verification methods, metrics, and thresholds to remain system- and context-specific.

## 1 Introduction

Artificial intelligence (AI) systems are increasingly deployed in domains where their outputs can affect rights, access to services, safety, welfare, and institutional decision-making. In such contexts, AI system quality is not only an engineering concern but also an ethical, organisational, and governance concern [1]. A system that performs well on average may still be unreliable for minority subpopulations, unstable under distributional shift, insufficiently transparent, or difficult to audit when harms occur. Yet there is still no widely accepted operational definition of *AI quality* [2], and evaluation often remains centred on aggregate metrics such as accuracy or benchmark performance [3].

This paper starts from the premise that AI system quality cannot be treated as a static property verified once before deployment. Unlike traditional software, AI systems are learned from data, operate probabilistically, and may change in behaviour or relevance as deployment environments evolve [4]. Their quality depends not only on model performance, but also on data provenance, intended purpose and deployment context, uncertainty communication, monitoring, explanation, contestability, and corrective action [5, 6]. Consequently, the evidence required to support a quality claim is necessarily dependent on the system, use case, risk, and operational context. AI system quality is therefore treated here as a relational and evolving property of a sociotechnical system that must be governed throughout the AI system lifecycle.

This challenge has become especially consequential under the European Union AI Act (EU AI Act). For providers of high-risk AI systems, Article 17 requires the establishment and maintenance of a quality management system (QMS) covering organisational arrangements, design and development, data governance, testing, risk management, monitoring, documentation, corrective actions, and post-market processes [7]. This obligation can be situated within the broader concept of an AI management system, as reflected in ISO/IEC 42001:2023 [8], which provides an organisational framework for AI governance through policies, roles, objectives, competence, risk management, monitoring, audit, corrective action, and continual improvement. European standardisation further operationalises AI Act-oriented quality management through EN 18286:2026, which specifies

requirements and guidance for establishing, implementing, maintaining, and continually improving a QMS for providers of AI systems [9].

Research has likewise developed increasingly concrete approaches to AI Act implementation. Conformity-assessment approaches translate regulatory and ethical requirements into assessable criteria and governance procedures [10, 11, 12]. Mustroph and Rinderle-Ma [13] propose and implement a QMS architecture that integrates services addressing AI Act requirements. More recently, Buscemi et al. [14] systematically translate legal requirements into implementable technical and organisational verification activities across the AI lifecycle. Such work substantially advances the operationalisation of regulatory requirements and clarifies what may need to be assessed or verified in practice.

The remaining problem addressed in this paper is therefore not the absence of QMS requirements, conformity-assessment approaches, or technical verification activities. Concrete verification necessarily depends on the AI system, intended purpose, risk, domain, and deployment context: the appropriate datasets, metrics, tests, thresholds, and acceptance criteria cannot be specified uniformly for all high-risk AI systems. A more generalisable governance problem nevertheless remains. Once context-specific evidence is produced, it must be possible to determine which quality claim that evidence supports, under which conditions the claim remains justified, who is responsible for maintaining and reviewing it, how new or contrary evidence can challenge it, and what action follows when its justification no longer holds. This shifts the analytical focus from prescribing universal verification procedures to governing the relationship between quality claims and context-specific evidence over time.

The central argument of this paper is that AI system quality can be conceptualised as *evidence-centric lifecycle governance*. Under this view, quality is not merely the outcome of testing, verification, or certification. It is expressed through claims about an individual AI system whose continued justification depends on structured, traceable, and reviewable evidence. Such evidence may include validation results, dataset documentation, risk records, monitoring logs, incident records, change histories, approval decisions, and corrective actions. Lifecycle governance maintains the relationship between these claims and their supporting evidence as the system moves from problem formulation and data preparation through development, deployment, monitoring, modification, retraining, and retirement. Importantly, the framework does not equate the existence of evidence with accountability. Evidence becomes governance-relevant when it is adequate to support a defined quality claim, linked to responsible actors, subject to review and challenge, and capable of informing corrective or remedial action when the claim is no longer adequately supported.

Accordingly, the paper is guided by two research questions:

- **RQ1:** How can AI system quality be conceptualised as an evidence-centric governance property of high-risk AI systems?
- **RQ2:** How can context-specific evidence supporting AI system quality claims be connected to lifecycle activities, organisational responsibilities, review and challenge, and corrective action within an EU AI Act-oriented QMS?

To address these questions, the paper develops a conceptual framework through a structured synthesis of software quality, AI engineering, manufacturing quality management, responsible-AI governance, AI assurance, and regulatory and standards-oriented sources. The framework organises recurring quality concerns into four governance-oriented clusters and reformulates them as quality claims linked to evidence artefacts, lifecycle stages, organisational responsibilities, and review mechanisms. Its contribution is therefore not a new quality metric, verification catalogue, conformity-assessment procedure, QMS implementation architecture, or QMS standard. Rather, it provides a generalisable governance structure for maintaining the evidential justification of quality claims concerning an individual AI system, while allowing the concrete evidence and verification activities required to support those claims to remain system- and context-specific.

The remainder of the paper is organised as follows. The next section reviews related work on AI quality, responsible AI, auditing, assurance, conformity assessment, AI Act-oriented QMS design, and regulatory verification. The paper then presents the framework development methodology and develops the conceptual foundation for AI quality as a sociotechnical and lifecycle-dependent property. It introduces four governance-oriented quality clusters, develops the evidence-centric assurance framework, maps evidence artefacts to lifecycle stages and organisational responsibilities, and illustrates the approach through an automotive residual-value agent. The paper concludes with implications, limitations, and directions for empirical evaluation.

# 2 Related Work

## 2.1 AI System Quality and Software Engineering

Classical software quality frameworks, such as ISO/IEC 25010, define quality through attributes including reliability, usability, security, maintainability, and performance efficiency [15]. These attributes remain relevant for AI-enabled systems, but they do not fully capture learned behaviour, probabilistic outputs, data dependency, drift, uncertainty calibration, context-sensitive degradation, or the social consequences of AI-mediated decisions [1, 4, 16].

Software engineering research for machine learning similarly shows that AI development differs from conventional software development. AI systems depend on training data, validation processes, model selection, deployment pipelines, monitoring, and

feedback loops [5]. Testing is also different because Machine Learning (ML) systems often lack complete oracles and require statistical evaluation [17, 18]. Moreover, the appropriate verification method is necessarily dependent on the system, intended purpose, risk, data, and deployment context. Technical validation alone therefore cannot determine whether a system remains fair, contestable, explainable, or accountable once embedded in an organisational and social context. Recent work consequently connects AI quality models to regulatory and governance requirements, including transparency, human oversight, and ethical integrity under the EU AI Act [19].

## 2.2 Responsible AI, Auditing, and Assurance

Responsible AI introduces concerns that are not reducible to technical performance, including fairness, transparency, explainability, accountability, privacy, human agency, contestability, and redress [20]. These principles require contextual operationalisation: fairness depends on relevant groups, metrics, data, and thresholds; transparency depends on audience and purpose; and accountability depends on traceability, role allocation, documentation, review, contestation, and remedy mechanisms. Recent Responsible AI framework work similarly emphasises the integration of trustworthy AI design, auditability, accountability, and governance [21].

AI auditing literature highlights the need for independent review, documentation, traceability, and evidence of system behaviour [22]. Yet auditing remains difficult because harms may emerge only after deployment, under particular operating conditions, or for particular subgroups. These harms are not only technical failures; they may include unequal treatment, misleading communication, inappropriate reliance, lack of recourse, or unclear responsibility when AI-supported decisions affect people. The notion of assurance is therefore useful because it shifts attention from isolated tests to justified claims supported by evidence. Picardi et al. [23], for example, develop assurance argument patterns that connect safety claims concerning machine-learning systems to supporting evidence. Such claim–evidence reasoning provides an important foundation for the present work. The focus here is broader sociotechnical AI system quality and the lifecycle governance through which quality claims, context-specific evidence, organisational responsibility, review and challenge, and corrective action remain connected over time.

## 2.3 AI Regulation, Conformity Assessment, and QMS

The EU AI Act gives regulatory importance to lifecycle-oriented quality management for high-risk AI systems, including risk management, data governance, documentation, transparency, human oversight, robustness, post-market monitoring, and incident reporting. Article 17 requires providers of high-risk AI systems to establish and maintain a QMS [7]. Comparable voluntary governance structures exist beyond the EU, including the NIST AI Risk Management Framework [24], while ISO/IEC 42001:2023 defines requirements for establishing, implementing, maintaining, and continually improving an organisation-level AI management system [8].

The operationalisation of AI Act requirements has also been addressed through conformity-assessment, standardisation, verification, and QMS implementation approaches. Floridi et al. [11] propose capAI, a conformity-assessment procedure intended to translate regulatory and ethical requirements into verifiable criteria across the design, development, deployment, and use of AI systems. Thelisson and Verma [25] examine conformity assessment under an earlier legislative stage of the EU AI Act and propose practical tools for its implementation. These contributions provide important foundations for translating regulatory and trustworthy-AI objectives into assessable practices, although they were developed against earlier legislative stages of the AI Act.

At the QMS level, Mustroph and Rinderle-Ma [13] propose a software-based QMS design that connects to an AI system and orchestrates sub-services addressing AI Act requirements, demonstrating the approach through an implemented prototype. European standardisation has subsequently advanced this operational landscape. EN 18286:2026 specifies requirements and guidance for the definition, implementation, and maintenance of a QMS for organisations providing AI systems for EU AI Act regulatory purposes [9]. Accordingly, the problem addressed by the present work should not be understood as an absence of AI Act-oriented QMS requirements, conformity-assessment approaches, or implementation mechanisms.

Recent work also addresses the translation of legal requirements into verification activities more directly. Buscemi et al. [14] systematically decompose AI Act requirements into operational sub-requirements and map them to implementable technical and organisational verification activities across the AI lifecycle. This provides an important bridge between regulatory requirements and practical verification. The concrete instantiation of such verification is necessarily context-specific: appropriate datasets, metrics, tests, thresholds, and acceptance criteria depend on the AI system, intended purpose, risk, domain, affected population, and deployment conditions. The present work addresses a complementary and more generalisable governance layer. Rather than prescribing a universal catalogue of verification activities or technical acceptance criteria, it examines how context-specific evidence, including evidence produced through verification, supports quality claims concerning an individual AI system; under which conditions that support remains adequate; who is responsible for maintaining and reviewing it; and how new or contrary evidence can trigger challenge, revision, and corrective or remedial action as the system and its context evolve.

Recent analysis of European AI standardisation further shows that the availability of standards does not eliminate implementation challenges. Kilian et al. [26] identify organisational, technical, sectoral, and resource-related difficulties associated with implementing the European AI standardisation landscape. Their findings provide additional context for why formal requirements and technical standards alone do not resolve the organisational challenge of interpreting, maintaining, reviewing, and acting upon evidence throughout an AI system's lifecycle.

Table 1: Positioning of the proposed framework relative to selected management-system, QMS, conformity-assessment, verification, and assurance approaches.

| **Approach** | **Primary object** | **Primary focus** | **Relationship to this work** |
|---|---|---|---|
| ISO/IEC 42001 [8] | Organisation | Requirements for establishing, implementing, maintaining, and continually improving an AI management system | Provides the organisation-level management-system context within which this work focuses on quality claims concerning an individual AI system and their connection to lifecycle evidence, responsible actors, review, and corrective action. |
| EN 18286 [9] | AI Act-oriented QMS | Requirements and guidance for establishing, implementing, maintaining, and continually improving a QMS for EU AI Act regulatory purposes | Provides the regulatory QMS structure within which this work examines a complementary assurance problem: how quality claims concerning an individual AI system are justified by evidence, assigned to responsible actors, reviewed and challenged, and revised when the supporting evidence is no longer adequate. |
| Mustroph and Rinderle-Ma [13] | AI Act-oriented QMS design | Design and implementation of a QMS based on EU AI Act requirements | Provides an implementation-oriented QMS architecture, whereas this work develops a conceptual governance framework for maintaining claim–evidence relationships and accountability across the lifecycle of an individual AI system. |
| Floridi et al. (capAI) [11] | AI system conformity assessment | Procedure translating regulatory and ethical requirements into verifiable criteria for assessing trustworthy AI systems | Provides a structured conformity-assessment procedure for evaluating whether an AI system satisfies regulatory and ethical requirements; this work instead focuses on how system-level quality claims remain justified through context-specific evidence and are connected to responsibility, lifecycle review and challenge, and corrective or remedial action over time. |
| Thelisson and Verma [25] | AI system conformity assessment | Governance structure and practical tools for conformity assessment under the developing EU AI Act | Provides an approach for conducting conformity assessment against regulatory requirements; this work addresses the complementary governance problem of maintaining and challenging the evidential justification of quality claims concerning an individual AI system throughout its lifecycle. |
| Buscemi et al. [14] | High-risk AI system assessment | Translation of AI Act requirements into technical and organisational verification activities across the AI lifecycle | Provides a regulatory-to-verification mapping, whereas this work examines how context-specific evidence, including evidence produced through verification, supports and challenges quality claims and is connected to responsibility, review, and corrective action over time. |
| Picardi et al. [23] | Safety-related ML systems | Assurance argument patterns linking safety claims and supporting evidence | Provides claim–evidence reasoning for safety assurance; this work extends that reasoning to broader sociotechnical AI system quality and embeds it within lifecycle responsibilities, monitoring, review, contestability, and corrective action in a QMS context. |
| **This work** | **Individual AI system** | **Evidence-centric lifecycle governance of AI system quality** | **Examines how quality claims remain justified over time by connecting them to context-specific and adequate evidence, responsible actors, lifecycle review and challenge, and corrective or remedial action within an AI Act-oriented QMS.** |

## 2.4 Gap Addressed by This Paper

Prior work provides substantial foundations in ML testing, software engineering for ML systems, responsible AI, auditing, assurance, conformity assessment, regulatory verification, management systems, and AI Act-oriented QMS design [18, 17, 5, 20, 23, 11, 25, 9, 13, 14, 22]. These approaches address complementary aspects of AI quality and governance at different levels. Management-system and QMS standards establish organisational and regulatory structures; conformity-assessment approaches

translate regulatory and ethical objectives into assessable criteria; verification approaches connect legal requirements to concrete technical and organisational assessment activities; and assurance approaches structure claims and supporting evidence.

The gap addressed here is therefore narrower than a general lack of operationalisation. Technical evidence and verification procedures cannot be fully generalised independently of the system and its context: what constitutes adequate evidence for robustness, fairness, transparency, human oversight, or another quality concern depends on the intended purpose, risk, affected population, operating conditions, and applicable requirements. What can be generalised is the governance relationship around such context-specific evidence. At the level of an individual AI system, it must be possible to determine which quality claim the evidence supports, under which conditions that support remains adequate, who is responsible for maintaining and reviewing it, how new or contrary evidence can challenge the claim, and what corrective or remedial action follows when its justification no longer holds.

Table 1 positions the present contribution relative to the closest management-system, QMS, conformity-assessment, verification, and assurance approaches.

Accordingly, the contribution of this paper is not a new QMS standard, compliance checklist, conformity-assessment procedure, verification catalogue, or QMS software architecture. Instead, it conceptualises AI system quality as an auditable and revisable claim about a sociotechnical system whose justification must be maintained through adequate lifecycle evidence. The framework generalises the relationship among quality claims, context-specific evidence, lifecycle stages, organisational responsibilities, review and challenge, monitoring, and corrective or remedial action, while leaving the concrete technical evidence, verification methods, metrics, and thresholds to be determined for the particular system and deployment context. In this sense, an AI-specific QMS can function as an accountability infrastructure through which quality claims are not merely documented, but continuously reviewed and acted upon.

# 3 Framework Development Methodology

We develop the framework through a structured conceptual synthesis of five complementary bodies of knowledge: software quality, software engineering for AI systems, manufacturing quality management, responsible-AI governance, and AI assurance and regulation. The objective is not to conduct a systematic literature review or to derive an exhaustive taxonomy of AI quality. Rather, the synthesis identifies transferable concepts relevant to the governance of high-risk AI systems and integrates them into a lifecycle-oriented framework for defining quality claims, identifying supporting evidence, assigning organisational responsibility, and maintaining those claims over time.

The source domains were selected because they address different but complementary dimensions of AI system quality governance. Software quality and AI engineering provide concepts for requirements specification, traceability, verification and validation, non-functional quality attributes, version control, testing, and change management. Manufacturing quality management contributes process control, conformance, auditability, root-cause analysis, corrective and preventive action, and continual improvement. Responsible-AI governance contributes concerns such as fairness, transparency, human oversight, accountability, and contestability. AI assurance contributes the use of explicit claims, supporting evidence, review processes, and assurance arguments. Finally, regulatory and standards-oriented sources provide requirements concerning risk management, data governance, technical documentation, logging, post-market monitoring, incident reporting, and quality management for high-risk AI systems.

These bodies of knowledge were used as conceptual inputs rather than as mutually exclusive literatures. The synthesis focused on concepts that satisfied at least one of three criteria: (i) they represented a recurring concern relevant to the quality or responsible operation of an AI system; (ii) they implied an organisational or lifecycle obligation that could be supported by observable evidence; or (iii) they were directly relevant to governance obligations applicable to high-risk AI systems under the EU AI Act. Concepts concerned only with organisation-wide management-system design, without a clear connection to the quality or assurance of an individual AI system, were treated as contextual rather than as framework elements.

The framework was constructed through four analytical steps. First, recurring quality concerns were identified across the selected bodies of knowledge and expressed at the level of the AI system rather than as isolated technical metrics or organisational policies. Second, conceptually related concerns were grouped according to the principal object of assurance and governance that they address. This resulted in four governance-oriented clusters: technical robustness and reliability; data and model integrity; human-facing intelligibility and oversight; and institutional accountability and governance. These clusters are intended as a parsimonious organisational structure rather than an exhaustive or mutually exclusive taxonomy; individual concerns may therefore interact across cluster boundaries.

Third, the concerns within each cluster were reformulated as assurance-oriented quality claims. A claim expresses a proposition about the AI system that an organisation should be able to justify, such as whether the system performs reliably under defined conditions, whether its data and model changes are controlled, or whether meaningful human oversight is maintained. For each class of claim, we then identified the types of evidence artefact that could make the claim reviewable, including requirements records, test results, dataset documentation, model and configuration histories, monitoring records, approval decisions,

incident records, and corrective-action documentation.

Fourth, the resulting claims and evidence artefacts were mapped to lifecycle stages, organisational responsibilities, review points, and relevant EU AI Act quality-management obligations. This step shifts the framework from a static taxonomy of quality characteristics towards a governance model in which quality claims must remain supported by identifiable evidence as the AI system is developed, deployed, monitored, modified, and, where necessary, corrected or withdrawn. The mapping was iteratively refined to reduce duplication between clusters and to ensure that each framework element could be connected to a lifecycle activity, evidence source, or accountability mechanism.

Figure 1 summarises this construction process.

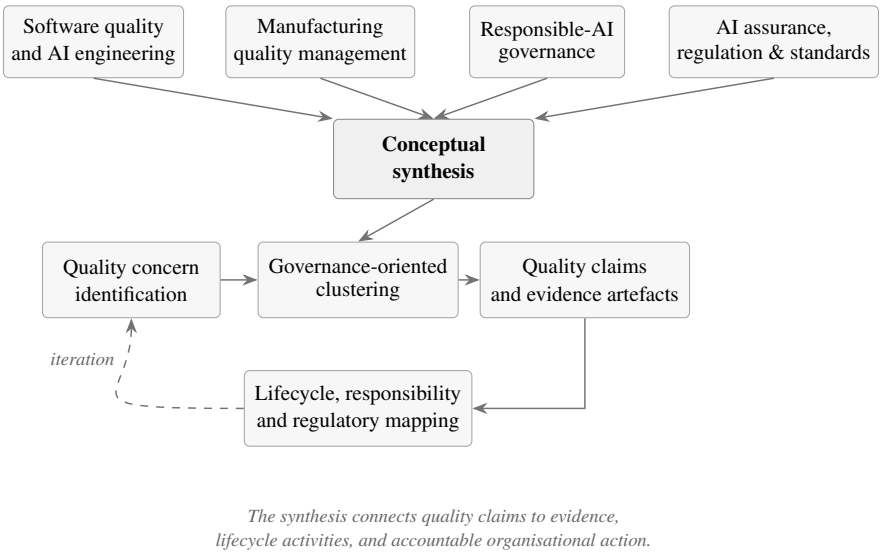


Figure 1: Framework development methodology.

The methodology is design-oriented and interpretive: it does not introduce a new quality metric, empirical measurement protocol, or formally validated taxonomy. The four clusters and their mappings should therefore be understood as an analytically grounded governance structure whose usefulness and completeness require further evaluation through application to concrete AI systems and organisational settings.

# 4 Conceptual Foundation: AI Systems, Quality Claims, and Measurability

## 4.1 AI Systems as Sociotechnical Quality Objects

We treat the object of quality not as an isolated model, but as an AI system operating within an organisational and sociotechnical context. Trustworthy AI depends on data, deployment context, human oversight, organisational processes, and governance mechanisms, not only model behaviour.

An *AI system* is therefore understood as a configuration of six interdependent elements: data ecosystem, learned model, task specification, inference and decision pipeline, deployment context, and governance and monitoring layer. These elements cover data quality and provenance, model assumptions, intended purpose, operational use of outputs, institutional setting, and lifecycle controls such as documentation, monitoring, incident response, retraining, and audit procedures [27, 5].

This system-level view explains why model performance alone is insufficient. An AI system can be low quality if it is deployed outside its intended context, weakly monitored, difficult to contest, unfair for particular groups, or unsupported by governance procedures. High-quality AI therefore requires technical behaviour, human-facing operation, and institutional governance to be jointly supported [16, 28].

## 4.2 Functional and Non-Functional Properties

AI system properties differ in how they are assessed and evidenced. *Functional properties* concern task performance under specified conditions, such as predictive correctness, output validity, or task completion, and are usually assessed through test sets, benchmarks, validation rules, and task-specific metrics.

*Non-functional properties* concern behaviour beyond task correctness, such as robustness, calibration, reliability, fairness, transparency, oversight, accountability, and temporal stability. Since these properties are often probabilistic, contextual, longitudinal, or normative, they require stress testing, monitoring, subgroup analysis, calibration checks, audits, and contextual review rather than one-time testing alone [18, 3]. Table 2 summarises this distinction.

Table 2: Functional and non-functional properties of AI systems.

| **Property type** | **Typical assessment** | **Evidence requirement** |
|---|---|---|
| Functional properties | Tests, benchmarks, rule-based checks, task metrics | Validation reports; benchmark results |
| Non-functional properties | Monitoring, stress tests, subgroup analysis, calibration checks, audits | Monitoring logs; bias assessments; incident, explanation, and governance records |

This distinction is central to the paper's argument. Functional properties may often be evaluated before deployment, although they too may change under distribution shift. Non-functional properties require continuous evidence because they depend on operational context, stakeholder interpretation, system evolution, and organisational response. For example, fairness cannot be established once and assumed indefinitely; it must be reassessed as populations, behaviours, and data distributions change. Similarly, transparency is not only a property of the model, but also of the explanations, interfaces, documentation, and decision procedures surrounding the model [20].

### 4.3 From Properties to Quality Claims

On this basis, we define *AI system quality* as the extent to which an AI system can fulfil its intended purpose reliably, safely, transparently, fairly, and accountably within a specific deployment context over time. This definition has three implications.

First, AI system quality is *systemic*: it depends on the interaction among data, model, task, pipeline, context, and governance. Second, AI system quality is *temporal*: it may degrade as operational data, user behaviour, market conditions, or regulatory expectations change. Third, AI system quality is *evidentiary*: claims about quality must be supported by traceable artefacts that allow system behaviour and governance to be inspected. This claim-based interpretation is aligned with assurance-oriented approaches, where confidence in a system is established through structured arguments supported by evidence rather than by isolated tests alone [29, 30, 31].

## 5 From Software Quality to AI System Quality

### 5.1 Why Traditional Quality Paradigms Are Insufficient

Traditional software and manufacturing quality assurance often assumes explicit requirements, deterministic or repeatable behaviour, process control, and stable specifications, allowing quality to be assessed largely before release. AI systems challenge these assumptions because their behaviour is learned, probabilistic, data-dependent, and context-sensitive [4, 5]. They may also fail silently as conditions change, becoming less reliable, fair, or calibrated while still producing outputs. Table 3 summarises why AI system quality requires lifecycle monitoring, risk management, transparency, fairness, and post-deployment evidence rather than one-time validation alone.

Table 3: Contrasting assumptions across quality paradigms.

| **Aspect** | **Software** | **Manufacturing** | **AI Systems** |
|---|---|---|---|
| Behaviour | Deterministic | Repeatable | Probabilistic |
| Specification | Explicit | Fixed | Partly data-driven |
| Correctness | Binary or rule-based | Conformance-based | Statistical and contextual |
| Testing | Oracle-based | Process validation | Offline and online evaluation |
| Failure | Explicit faults | Defects | Silent degradation |
| Assurance | Pre-release validation | Process control | Lifecycle evidence |
| Governance | Documentation | Compliance | Continuous monitoring and auditability |

The implication is not that traditional quality frameworks should be abandoned. They remain useful for infrastructure, software integration, documentation, process discipline, and risk management. Rather, they must be extended by lifecycle monitoring, data governance, uncertainty-aware evaluation, subgroup analysis, incident reporting, and evidence-based accountability. In this sense, AI system quality can be understood as an extension of software quality rather than a replacement for it: conventional quality attributes remain relevant, but they must be embedded within a broader sociotechnical, regulatory, and assurance-oriented framework [6].

### 5.2 Implications for AI System Quality

The conceptual foundation above explains why AI system quality cannot be reduced to model performance. Because AI systems combine data, models, pipelines, deployment contexts, and governance mechanisms, their quality is systemic rather than model-centric. Because their behaviour may change under new operational conditions, their quality is temporal rather than fixed. Because claims about robustness, fairness, transparency, and accountability require inspection, their quality is evidentiary rather than merely declarative. These characteristics explain why traditional quality paradigms must be extended for AI systems.

## 6 Quality Concerns for Responsible AI Governance

AI system quality involves heterogeneous concerns that arise at different points in the system and its organisational context. Rather than treating these concerns as a flat list of attributes, we organise them according to their *principal object of assurance and governance*. The purpose of the clustering is therefore analytical rather than taxonomic: it distinguishes the principal object to which a quality claim refers, the types of evidence that may support it, and the organisational actors involved in reviewing and acting upon that evidence.

The four clusters reflect four recurring loci at which quality must be established and maintained in an AI system. First, quality concerns may refer primarily to the *behaviour of the deployed technical system*, including whether it performs reliably and remains robust under uncertainty and changing conditions. Second, they may concern the *integrity of the data and model development process* through which that behaviour is produced and modified. Third, they may concern the *interaction between the AI system and human stakeholders*, including whether outputs can be understood, appropriately relied upon, supervised, or contested. Fourth, they may concern the *institutional arrangements* through which responsibility is assigned, operation is monitored, failures are escalated, and corrective or remedial action is taken.

These four loci were selected because together they provide a parsimonious structure spanning the technical, developmental, human-facing, and organisational dimensions of AI system quality. They also correspond to distinct forms of evidence and responsibility across the lifecycle. The classification is not intended to imply that concerns are independent or mutually exclusive. For example, fairness may depend simultaneously on data representativeness, model behaviour, user-facing consequences, and organisational review; similarly, robustness may depend on both technical performance and the quality of monitoring and change-control processes. Where a concern spans multiple clusters, it is placed according to the principal object being assured while its cross-cluster dependencies remain part of the lifecycle analysis.

Accordingly, the framework uses four governance-oriented clusters:

- **Technical robustness and reliability** concerns whether the operational behaviour of the AI system remains sufficiently reliable under intended and reasonably foreseeable conditions. It covers performance, robustness, calibration, availability, latency, security, and resilience to distributional shift. The principal object of assurance is therefore the system's technical behaviour in operation.
- **Data and model integrity** concerns whether the data, model-development, and change processes provide a trustworthy basis for system behaviour. It covers data provenance, representativeness, data quality, bias assessment, validation, reproducibility, versioning, and change control. The principal object of assurance is the integrity and traceability of the data–model production process.
- **Human-facing intelligibility and oversight** concerns whether relevant stakeholders can understand, appropriately rely on, supervise, challenge, or contest the AI system and its outputs. It covers transparency, explainability, user documentation, confidence and uncertainty communication, human oversight, escalation procedures, contestability, and interface design. The principal object of assurance is the relationship between the AI system and the humans who use, supervise, or are affected by it.
- **Institutional accountability and governance** concerns whether the organisation can assign responsibility, maintain oversight, monitor operation, respond to failures, and demonstrate that quality concerns are acted upon. It covers risk

Table 4: Governance-oriented clusters of AI system quality concerns.

| Cluster | Principal object of assurance | Representative concerns |
|---|---|---|
| Technical robustness and reliability | Operational system behaviour | Performance, robustness, calibration, availability, security, drift resilience |
| Data and model integrity | Data–model development and change process | Data provenance, representativeness, bias assessment, validation, reproducibility, versioning, change control |
| Human-facing intelligibility and oversight | Human–system interaction and supervision | Transparency, explanations, confidence communication, user documentation, human oversight, contestability |
| Institutional accountability and governance | Organisational responsibility and control | Risk management, audit trails, monitoring, incident response, corrective actions, decision rights, QMS documentation |

management, documentation, audit trails, incident records, post-market monitoring, corrective and remedial actions, decision rights, and QMS procedures. The principal object of assurance is the organisational capacity to govern the system throughout its lifecycle.

The clustering contributes to RQ1 by providing a parsimonious structure through which heterogeneous quality concerns can be reformulated as reviewable quality claims. Its contribution does not lie in proposing a new taxonomy of AI quality attributes. Rather, the clusters provide an organisational layer for asking, for each type of concern, what proposition about the AI system must be justified, what context-specific evidence can support that proposition, who is responsible for reviewing it, and what action follows when the evidence is no longer adequate.

This classification also supports the subsequent lifecycle framework. The technical, data–model, human-facing, and institutional clusters correspond to different but interacting sources of evidence across development, validation, deployment, monitoring, and change. The four-cluster structure therefore serves as a bridge between heterogeneous AI quality concerns and the paper's central governance relation among quality claims, context-specific evidence, responsibility, review and challenge, and corrective or remedial action.

# 7 From Testing to Evidence-Based Assurance

## 7.1 The Limits of Test-Based Quality

Testing and pre-deployment validation remain necessary components of AI system quality assurance, but they are insufficient as a standalone basis for quality claims. Validation on selected datasets, stress tests, and benchmarks can estimate system performance before release, but cannot fully determine how an AI system will behave across future deployment contexts, changing data distributions, evolving user behaviour, or unforeseen operational conditions. Nor can testing alone demonstrate fairness over time, transparency to affected stakeholders, meaningful human oversight, or accountability after incidents.

This limitation is especially important because many AI system quality concerns are probabilistic, contextual, and temporal. A model may satisfy performance thresholds during validation while later becoming unreliable due to drift, subgroup-specific degradation, data-pipeline changes, or shifts in the operational environment. Similarly, a system may be technically accurate while still being difficult to contest, insufficiently transparent, or weakly governed.

For this reason, AI system quality claims should be treated as assurance claims. Claims such as "the system is robust", "the system is fair", or "the system supports meaningful human oversight" are credible only when supported by appropriate, traceable, and reviewable evidence. Such evidence includes test results, but also monitoring records, data documentation, incident logs, subgroup analyses, explanation reports, oversight procedures, audit trails, and corrective actions.

## 7.2 Evidence Artefacts for AI System Quality

An evidence-centric view of AI system quality requires artefacts that connect quality claims to inspectable records. Rather than forming a universal checklist, these artefacts support the demonstration, review, and audit of claims about reliability, fairness, transparency, oversight, and accountability. Verification activities, including the technical and organisational activities identified by Buscemi et al. [14], can generate important forms of such evidence. The present framework treats these outputs as part of a broader evidential relationship: verification results become governance-relevant when they can be connected to a defined quality claim, its deployment context, responsible actors, review mechanisms, and potential corrective or remedial action.

Table 5 presents this traceability view using the four quality clusters in Table 4. It shows illustrative ways in which evidence artefacts can support assurance claims, review practices, corrective actions, and EU AI Act expectations. The artefacts

Table 5: Clustered evidence artefacts for assurance-based AI system quality management and EU AI Act alignment.

| Quality cluster | Assurance claim | Typical artefacts | Review / action | EU AI Act |
|---|---|---|---|---|
| Technical robustness and reliability | Reliable under uncertainty, drift, and operational stress | Validation report; calibration results; stress tests; runtime logs; monitoring and incident records | Performance review; drift monitoring; recalibration, retraining, rollback, suspension | Arts. 9, 12, 15, 72 |
| Data and model integrity | Data and model processes are traceable, representative, reproducible, and controlled | Data documentation; provenance records; bias assessment; model card; training log; change record | Data-quality review; subgroup analysis; reproducibility checks; revalidation | Arts. 9, 10, 11, 12, 17 |
| Human-facing intelligibility and oversight | Users can understand, contest, and appropriately rely on the system | Explanation report; user documentation; confidence communication; oversight design; escalation workflow | Explanation review; oversight testing; escalation, override, redesign | Arts. 9, 13, 14 |
| Institutional accountability and governance | Responsibilities, risks, incidents, and corrective actions are documented and auditable | Intended-purpose specification; risk record; audit trail; QMS record; incident and corrective-action records | Internal audit; incident investigation; corrective-action tracking; management review | Arts. 9, 17, 72, 73 |

are not intended as a complete or universally applicable evidence catalogue; their relevance and adequacy depend on the quality claim, system, intended purpose, risk, and deployment context. Fairness claims, for example, may require data documentation, subgroup analysis, bias assessment, validation, and monitoring records; oversight claims may require user documentation, escalation procedures, audit logs, and incident records; and robustness claims may require stress tests, drift reports, runtime logs, change records, and corrective-action histories. The table therefore illustrates how context-specific technical and organisational evidence can be incorporated into auditable quality-management practices without prescribing a universal verification procedure.

# 8 Illustrative Case Study: Evidence-Centric Governance for an Automotive Residual-Value Agent

To illustrate the framework, consider an automotive company deploying an AI agent that estimates a customer's vehicle residual value over time. The agent combines vehicle data, customer usage, service history, configuration, market demand, and macroeconomic indicators to detect significant depreciation patterns and notify the customer.

The system is not only a forecasting model, but a sociotechnical AI service involving data pipelines, predictive models, customer communication, monitoring, escalation, and organisational responsibility. Its quality therefore depends not only on predictive accuracy, but also on data appropriateness, robustness under market change, responsible uncertainty communication, fairness, human oversight, and auditability.

Figure 3 presents the case as an evidence-centric system-level QMS view, linking lifecycle stages to responsible roles, governance questions, evidence artefacts, auditor focus, and assessment status.

## 8.1 Lifecycle Interpretation of the Case

Figure 3 frames the residual-value agent as a lifecycle-governed AI service rather than an isolated forecasting model. Each stage connects a quality concern to roles and evidence artefacts:

- **Problem formulation and design:** define the bounded purpose, foreseeable harms, affected persons, accountability arrangements, and exclusion from automated financing, insurance, pricing, or trade-in decisions.
- **Data collection and preparation:** establish reliable, current, and representative data on vehicle history, usage, market trends, macroeconomic factors, and customer/vehicle segments, while assessing whether data gaps or proxies could create unequal impacts across groups.
- **Model development:** build calibrated, version-controlled forecasting and alerting models that communicate uncertainty rather than deterministic value claims, and assess whether model choices could amplify information asymmetries or inappropriate reliance by customers.

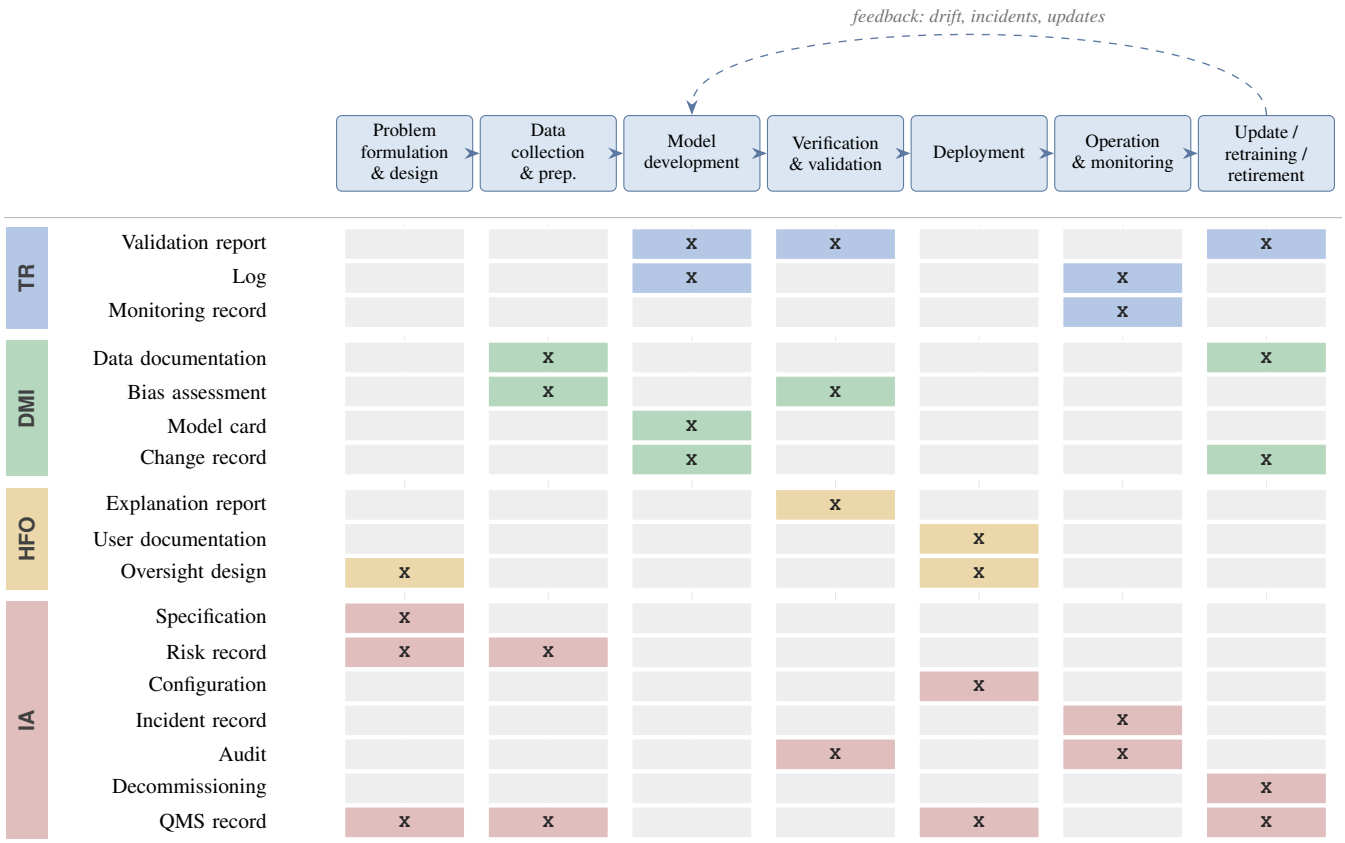


Figure 2: Lifecycle framework for AI-specific QMS. The top band shows seven lifecycle stages, and the heatmap maps evidence artefacts to the stages where they are typically produced or maintained. Artefacts are grouped into four quality clusters: **TR** technical robustness and reliability, **DMI** data and model integrity, **HFO** human-facing intelligibility and oversight, and **IA** institutional accountability and governance. Filled cells indicate stage–artefact relevance; the dashed arrow shows feedback from drift, incidents, or controlled updates.

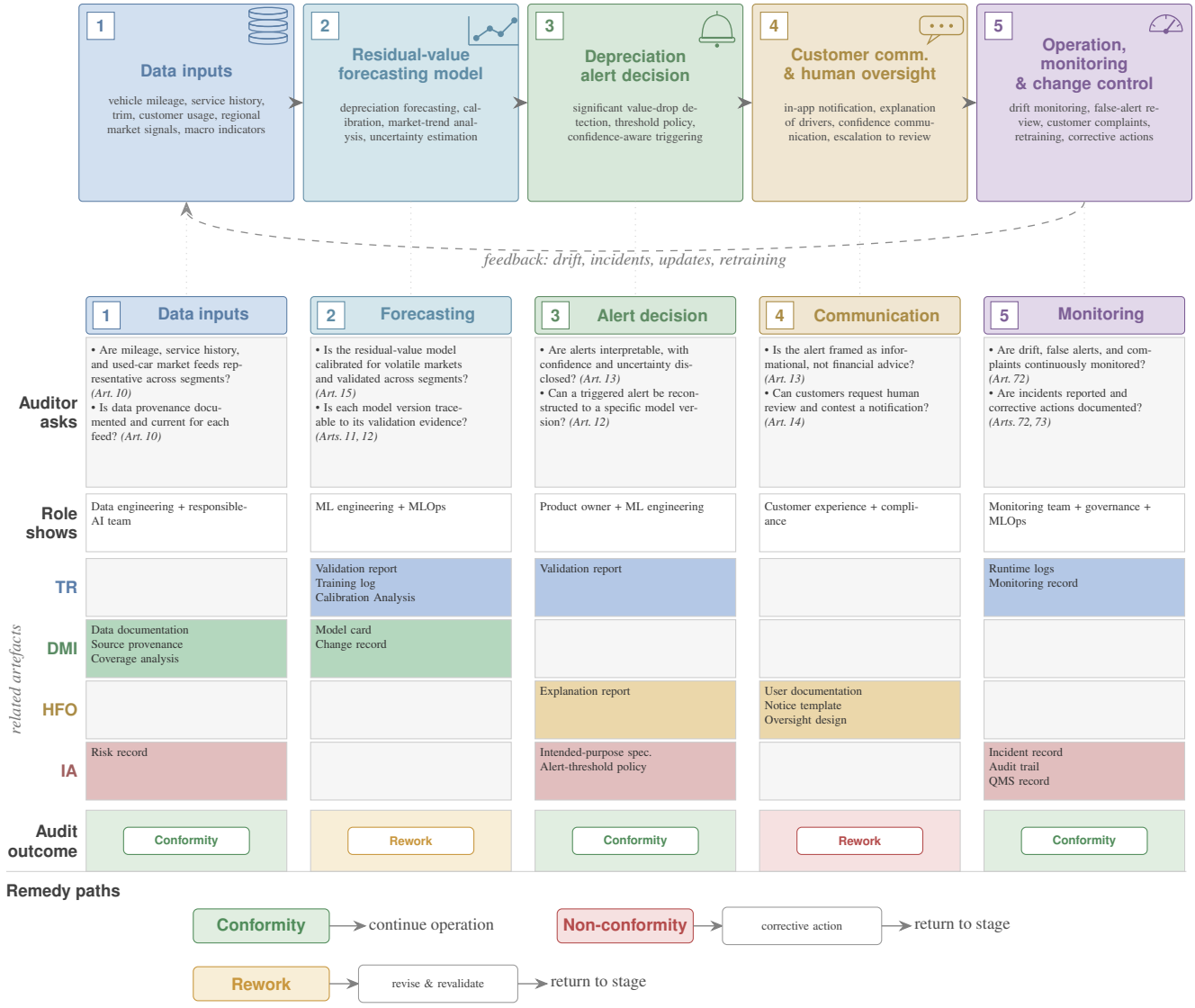


Figure 3: Evidence-centric system-level QMS view for the automotive residual-value agent. For representational clarity and space reasons, the figure condenses the seven-stage lifecycle in Figure 2 into a five-stage operational pipeline. It maps these stages to auditor questions, responsible roles, case-specific evidence artefacts, and illustrative audit outcomes, showing how quality claims, context-specific evidence, organisational responsibility, review, and corrective action can be connected for a specific AI system within a QMS. It uses the same governance clusters as Figure 2: technical robustness and reliability (TR), data and model integrity (DMI), human-facing intelligibility and oversight (HFO), and institutional accountability and governance (IA). The feedback and remedy paths illustrate how drift, incidents, rework, and non-conformity can challenge existing quality claims and trigger review, revision, corrective action, or revalidation. The specific artefacts, evaluation methods, review procedures, and evidence thresholds remain system- and context-dependent and must be determined according to the system's intended purpose, risks, affected stakeholders, and deployment conditions.

- **Verification and validation:** assess accuracy, calibration, subgroup errors, explanation quality, and the risks of misleading alerts, unequal performance, or customers interpreting outputs as financial advice.
- **Deployment:** ensure the active model, alert thresholds, explanations, notices, and escalation pathways match the validated configuration, and that users are informed about system purpose, limits, uncertainty, and available routes for human review.
- **Operation and monitoring:** monitor drift, false alerts, subgroup impacts, customer complaints, market changes, and triggers for recalibration or suspension, with attention to contestability, redress, and recurring harms to affected users.
- **Update, retraining, and retirement:** document changes, reassessment, customer-notification traceability, and decommissioning decisions, ensuring that updates do not silently alter user expectations, accountability, or access to review and remedy.

The case therefore illustrates not only residual-value prediction, but the governance of a customer-facing AI service through bounded purpose, relevant data, calibrated forecasting, validated communication, monitoring, controlled change, and auditable evidence.

## 8.2 Retrospective Diagnostic Application

The preceding case illustrates how evidence-centric lifecycle governance can be designed prospectively for the residual-value AI service. The same framework can also be used retrospectively when the service produces disputed or harmful outcomes after deployment. In the present use case, the purpose of retrospective analysis is to reconstruct how a customer-facing depreciation alert came to be produced, which quality claims supported that outcome, whether the underlying evidence remained adequate, and whether the organisation responded appropriately when contrary evidence emerged.

Consider an illustrative situation in which the residual-value agent begins to generate inaccurate or unstable depreciation estimates for a subset of vehicles after a change in market conditions. For example, changes in regional used-car demand, fuel prices, macroeconomic conditions, or the availability of vehicle history data may cause the relationship between mileage, service history, configuration, and market value to differ from the conditions represented in the original validation data. The forecasting model may therefore remain technically operational while its calibration and residual-value estimates become less reliable for particular vehicle classes, regions, or usage patterns.

The governance problem becomes more consequential when such estimates feed the next stages of the service. An inaccurate forecast may cross the configured depreciation-alert threshold and generate a customer notification indicating a significant expected loss in vehicle value. If the alert is presented without sufficient uncertainty information or explanation of the market assumptions underlying the estimate, a customer may interpret it as a deterministic or authoritative assessment rather than as an uncertain forecast. The resulting failure is therefore not only a forecasting error. It may involve a chain of failures spanning data quality, model calibration, alert-threshold design, customer communication, monitoring, and organisational response.

The evidence-centric framework makes this chain explicit by tracing the retrospective analysis through the five operational stages represented in Figure 3: data inputs, residual-value forecasting, depreciation alert decision, customer communication and human oversight, and operation, monitoring, and change control.

1. **Which quality claim was challenged?**

   At the forecasting stage, the relevant quality claim may be that the residual-value model remains sufficiently accurate and calibrated for the vehicle populations and market conditions in which it is deployed. At the alert-decision stage, a further claim is that the selected depreciation threshold identifies materially significant value changes without producing excessive false alerts. At the communication stage, the service also relies on the claim that customers can understand the informational nature, uncertainty, and limitations of the notification.

   A disputed alert can therefore challenge several connected claims rather than a single global claim that "the system is accurate." Retrospective analysis first identifies which claim or combination of claims no longer appears justified.

2. **What evidence supported those claims, and was it still adequate?**

   The forecasting claim may have been supported by validation reports, calibration analyses, subgroup performance results, model-version records, and assumptions about the market conditions represented in the training and validation data. The data-related claim may have relied on provenance records, coverage analyses, freshness checks, and documentation of vehicle mileage, service history, configuration, and regional market feeds. The alert-decision claim may have been supported by threshold-selection records, false-positive analyses, and validation of the relationship between forecast uncertainty and alert triggering. The communication claim may have relied on explanation reports, user-documentation reviews, notice templates, and usability or oversight testing.

Retrospective analysis asks not only whether these artefacts existed, but whether they still reflected the conditions under which the disputed alert was produced. A validation report may remain available while becoming evidentially weak if the market has shifted materially, a data feed has become stale, or a model update has changed behaviour without corresponding revalidation.

3. **Where in the operational pipeline should the problem have been detected?**

   The failure may originate at the data-input stage if market or vehicle history feeds become incomplete, stale, or unrepresentative. It may arise during residual-value forecasting if model calibration deteriorates under volatile or previously unseen market conditions. It may occur at the alert-decision stage if thresholds remain unchanged despite changes in forecast uncertainty or false-alert rates. It may arise at the communication stage if the notification does not make uncertainty, intended purpose, or routes for human review sufficiently clear.

   Alternatively, the original development and deployment may have been reasonable, while the failure emerges later because post-deployment monitoring does not detect drift, recurrent false alerts, subgroup-specific errors, or changing complaint patterns. The lifecycle perspective therefore allows the organisation to distinguish a defect introduced before deployment from degradation that should have been detected during operation.

4. **Who was responsible for reviewing and acting on the evidence?**

   Responsibility can be traced using the role structure already associated with the use case. Data engineering and responsible-AI functions may be responsible for investigating the quality and representativeness of vehicle and market feeds. ML engineering and MLOps may be responsible for model calibration, version traceability, and revalidation. Product ownership and ML engineering may be responsible for depreciation-alert thresholds and triggering logic. Customer-experience and compliance functions may be responsible for notice design, explanation, escalation, and human-review pathways. Monitoring, governance, and MLOps functions may be responsible for detecting drift, false alerts, complaints, and conditions requiring intervention.

   Retrospective analysis therefore asks whether the evidence was connected to actors with clear ownership and decision authority. A technically detectable problem can persist if no actor is responsible for challenging continued operation or if escalation responsibilities are fragmented across stages.

5. **What governance response should have followed?**

   The appropriate response depends on where the evidential breakdown is located. Stale or incomplete market data may require data remediation or temporary restriction of affected inputs. Loss of calibration may require recalibration, retraining, or renewed validation. Excessive false alerts may require revision of depreciation thresholds or confidence-aware triggering. Misleading customer interpretation may require changes to the notification, explanation of value drivers, presentation of uncertainty, or access to human review.

   Where the available evidence no longer supports continued operation, the organisation may need to restrict the service, suspend alerts for affected vehicle populations, roll back a model or configuration, notify affected users, or conduct additional validation before the relevant quality claims can be re-established. The corrective action should itself become part of the evidence chain through incident records, change records, approval decisions, and subsequent revalidation.

Table 6 summarises the diagnostic interpretation for this use case.

This retrospective interpretation shows why a disputed residual-value alert cannot be diagnosed adequately by asking only whether the forecasting model made an inaccurate prediction. The relevant governance chain begins with the vehicle and market data used by the service, continues through forecasting and alert triggering, and extends to the way uncertainty is communicated to the customer and the way operational evidence is monitored and acted upon. A failure at one stage may invalidate evidence relied upon at subsequent stages.

The case also illustrates the distinction between evidence availability, evidence adequacy, and accountability. An organisation may possess validation reports, model cards, monitoring logs, and QMS records while still failing to govern the service effectively. For example, a calibration report produced before deployment may remain available but no longer be adequate after a substantial market shift. Monitoring evidence may reveal increasing false alerts, yet accountability still fails if no responsible actor is authorised or required to challenge the continued use of the existing model or alert threshold.

The diagnostic contribution of the framework is therefore to reconstruct the relationship between an operational outcome and the quality claims that justified it. For the residual-value service, the relevant chain can be read from vehicle and market evidence, through model and threshold decisions, to customer-facing communication, monitoring, and organisational response. Evidence becomes governance-relevant when it can support or challenge a defined claim, is reviewable by responsible actors, and can trigger proportionate corrective or remedial action. In this sense, the retrospective application complements the prospective lifecycle design shown in Figure 3: the prospective view asks what evidence and responsibility should accompany

Table 6: Retrospective diagnostic application to the illustrative residual-value AI service.

| Operational element | Illustrative failure | Relevant evidence | Governance implication |
|---|---|---|---|
| Data inputs | Regional market or vehicle history data become stale, incomplete, or unrepresentative for part of the deployed population | Data provenance; freshness checks; coverage analysis; missingness analysis; source and dataset versions | Determine whether the forecasting claim remains supportable for affected vehicles, regions, or customer segments |
| Residual-value forecasting | Forecast errors or calibration deteriorate under changed market conditions | Validation reports; calibration analyses; subgroup performance; training and model-version records | Reassess model validity and trigger recalibration, retraining, or restricted use where required |
| Depreciation alert decision | Existing alert thresholds produce excessive or unstable alerts as forecast uncertainty changes | Alert-threshold policy; false-alert analysis; validation results; runtime logs | Reassess triggering logic and revalidate thresholds against current operating conditions |
| Customer communication and oversight | Customers interpret an uncertain depreciation forecast as a deterministic statement of future vehicle value or financial advice | Explanation report; user documentation; notice template; uncertainty communication; complaint and escalation records | Revise explanations, uncertainty presentation, and access to human review or contestability mechanisms |
| Operation and monitoring | Drift, false alerts, subgroup-specific errors, or complaint patterns emerge but do not trigger timely intervention | Monitoring records; drift reports; runtime logs; complaint trends; incident records | Strengthen monitoring thresholds, review frequency, and escalation triggers |
| Organisational accountability | Evidence of degradation is available, but authority to challenge continued operation or require revalidation is unclear | Role assignments; review logs; approval history; audit trail; QMS records | Clarify decision rights and ensure that contrary evidence can trigger review and corrective action |
| Corrective action | The service continues issuing alerts although the relevant quality claim is no longer adequately supported | Incident record; change record; rollback or suspension decision; revalidation evidence | Restrict, recalibrate, retrain, modify, or suspend the affected service until the claim can be justified again |

each stage, whereas the retrospective view asks where that evidential and accountability chain broke when the service produced an unacceptable outcome.

# 9 Analytical Grounding by Standards Alignment

Because the framework is conceptual, we ground it analytically by positioning its core elements against established quality, management-system, regulatory, and responsible-AI foundations. The purpose is not to claim equivalence with existing standards, but to show that the framework is grounded in recognised practices while adding an evidence-centric assurance layer. Table 7 summarises this alignment.

Table 7: Analytical grounding of the framework by alignment with existing foundations.

| Foundation | Established focus | Extension in this framework |
|---|---|---|
| Software and AI system quality models [15, 16] | Quality attributes; functional and non-functional properties | Reframes quality attributes as assurance claims requiring lifecycle evidence |
| AI and data management systems [8, 27] | Organisational governance; data-quality management; continual improvement | Links governance processes to artefacts, roles, monitoring, and corrective actions |
| EU AI Act QMS and standards [7, 9, 13] | Risk management; documentation; post-market monitoring; incident reporting | Maps regulatory obligations to evidence-centric QMS and auditability structures |
| Responsible AI and auditing [20, 22, 21] | Fairness; transparency; accountability; oversight; auditability | Operationalises responsible-AI principles through claims, artefacts, and review practices |

This alignment does not constitute empirical validation, but provides a preliminary plausibility check: the framework integrates recognised foundations rather than introducing quality concepts ad hoc.

# 10 Implications

The framework translates AI system quality into an auditable governance process for different stakeholders:

- **Researchers** can specify which quality claims their studies support, under which contexts, with what evidence, and with what limitations.
- **Providers and practitioners** can embed evidence artefacts into AI-specific QMS workflows for development, deployment, monitoring, incident response, and change management.
- **Deployers** can use lifecycle evidence to support appropriate use, user communication, human oversight, escalation, feedback, and contestability in operational settings.
- **Regulators and auditors** can assess whether quality claims are supported by current, context-relevant evidence and corrective-action mechanisms.
- **Affected persons** benefit when quality claims are reviewable, contestable, and connected to explanation, escalation, and redress.

# 11 Discussion, Limitations, and Future Work

The framework connects technical and ethical views of AI system quality by linking quality claims to evidence artefacts, responsible roles, review points, corrective actions, and mechanisms for contestability and redress. In this view, an AI-specific QMS can function as an accountability infrastructure for providers, deployers, auditors, and affected persons rather than merely as an internal compliance mechanism. The framework does not replace regulatory requirements, management-system standards, or domain-specific quality criteria; instead, it provides a governance structure through which claims about an individual AI system can be connected to evidence, responsibility, review, and action across the lifecycle.

## 11.1 From Evidence Availability to Accountable Governance

A central implication of an evidence-centric approach is that the presence of documentation cannot itself be equated with effective governance. Records, logs, validation reports, risk assessments, or audit artefacts may exist while providing weak support for the quality claims they are intended to justify. Evidence may be incomplete, outdated, insufficiently relevant to the claim, difficult to trace to its origin, or maintained primarily to demonstrate procedural compliance.

The framework therefore distinguishes between *evidence availability*, *evidence adequacy*, and *accountability*. Evidence availability concerns whether relevant artefacts exist and can be retrieved. Evidence adequacy concerns whether those artefacts are sufficiently relevant, traceable, timely, complete, and reviewable to justify a particular quality claim. Accountability introduces a further organisational requirement: responsible actors must be able to review the evidence, challenge the corresponding claim, and initiate corrective or remedial action when the claim is no longer supported.

Under this interpretation, evidence becomes governance-relevant when it is connected to an explicit quality claim and embedded within processes of review, challenge, decision, and action. Relevant properties of governance evidence therefore include provenance, relevance, timeliness, completeness, reviewability, explicit linkage to a quality claim, assigned ownership, and the capacity to support escalation, corrective action, or redress. These properties do not constitute universal evidence thresholds; their required form and strength remain dependent on the system, risk, deployment context, and applicable regulatory or domain-specific requirements.

This distinction is particularly important for high-risk AI systems because documentation requirements can otherwise encourage performative compliance: organisations may accumulate compliance artefacts without demonstrating that those artefacts influence operational decisions or responses to emerging risks and harms. An evidence-centric QMS should therefore not be evaluated only by whether required artefacts exist, but also by whether evidence remains adequate to support the associated claims and whether contrary evidence can trigger review and organisational action.

In this sense, evidence availability is necessary but insufficient for accountability. Evidence becomes governance only when it can be used to challenge a quality claim and trigger accountable action.

## 11.2 Limitations and Future Work

The proposed framework is conceptual and interpretive rather than prescriptive or empirically validated. It does not define universal thresholds for fairness, robustness, explainability, evidence adequacy, or acceptable risk, since these depend on domain,

jurisdiction, deployment context, system purpose, and stakeholder expectations. Its purpose is instead to bridge responsible-AI principles, AI Act-oriented QMS obligations, and auditable lifecycle practices at the level of an individual AI system.

The framework is further limited by its focus on high-risk AI systems and its interpretive mapping to the EU AI Act. Although the underlying relationships between claims, evidence, responsibilities, and lifecycle review may be applicable beyond this regulatory context, this broader applicability is not established in the present work. Similarly, the framework uses a deliberately broad evidence-artefact vocabulary. Concrete artefacts, formats, review procedures, and evidential thresholds require domain- and organisation-specific refinement.

A further limitation concerns the organisational cost of evidence-centric governance. Producing, maintaining, reviewing, and connecting lifecycle evidence can create substantial documentation and coordination overhead. More evidence does not necessarily produce better governance, particularly where artefact production becomes detached from decision-making, challenge, and corrective action. Future evaluation should therefore consider not only whether the proposed evidence structures can be implemented, but also their proportionality, organisational burden, and resistance to performative compliance.

Future work should evaluate the framework through domain-specific and organisational case studies, including retrospective application to documented AI failures and prospective application during system development and operation. Such studies could examine whether the framework helps identify assurance gaps, clarify responsibility, improve lifecycle traceability, or support corrective action compared with alternative QMS or assurance approaches. Further work should also develop criteria for evidence adequacy, investigate integration with structured assurance cases and QMS templates, and examine whether evidence-centric governance produces measurable improvements in AI system quality and accountability over time.

# 12 Acknowledgments

The work of Fatima Rabia Yapicioglu was supported by the Piano Nazionale di Ripresa e Resilienza (PNRR) and Automobili Lamborghini S.p.A. under Grant CUP J33C23001490009.

# References

[1] Michael Felderer and Rudolf Ramler. Quality assurance for ai-based systems: Overview and challenges (introduction to interactive session). In *International Conference on Software Quality*, pages 33–42. Springer, 2021.

[2] Markus Borg. The aiq meta-testbed: Pragmatically bridging academic ai testing and industrial q needs. In *International Conference on Software Quality*, pages 66–77. Springer, 2021.

[3] Edwin Farley. Ai auditing: First steps towards the effective regulation of artificial intelligence systems. *Available at SSRN 4676184*, 2023.

[4] Valentina Lenarduzzi, Francesco Lomio, Sergio Moreschini, Davide Taibi, and Damian Andrew Tamburri. Software quality for ai: Where we are now? In *International conference on software quality*, pages 43–53. Springer, 2021.

[5] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300. IEEE, 2019.

[6] Seung-Hee Kim. Suggestion for an iso 25010 quality model encompassing ai-based software. *Journal of Internet Computing & Services*, 25(5):67, 2024.

[7] European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, 2024.

[8] ISO/IEC. Iso/iec 42001: 2023 information technology—artificial intelligence—management system, 2023.

[9] Artificial intelligence — quality management system for eu ai act regulatory purposes, 2026.

[10] Luciano Floridi, Josh Cowls, Monica Beltrametti, et al. AI4People — an ethical framework for a good AI society. *Minds and Machines*, 28(4):689–707, 2018.

[11] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. capAI — a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act, 2022. SSRN Scholarly Paper.

[12] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. A procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act. 2025.

[13] Henryk Mustroph and Stefanie Rinderle-Ma. Design of a quality management system based on the eu ai act. In *Legal Knowledge and Information Systems*, pages 314–320. IOS Press, 2024.

[14] Alessio Buscemi, Tom Deckenbrunnen, Fahria Kabir, Kateryna Mishchenko, and Nishat Mowla. Assessing high-risk AI systems under the EU AI Act: From legal requirements to technical verification, 2026. arXiv preprint arXiv:2512.13907.

[15] ISO/IEC 25002:2024. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model overview and usage. Standard, International Organization for Standardization, Geneva, CH, 03 2024.

[16] ISO/IEC 25059:2023. Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems. Standard, International Organization for Standardization, Geneva, CH, 06 2023.

[17] Christian Murphy, Gail E Kaiser, and Marta Arias. A framework for quality assurance of machine learning applications. 2006.

[18] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1):1–36, 2020.

[19] Jessica Kelly, Shanza Ali Zafar, Lena Heidemann, Joao-Vitor Zacchi, Delfina Espinoza, and Núria Mata. Navigating the eu ai act: A methodological approach to compliance for safety-critical products. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 979–984. IEEE, 2024.

[20] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering. *ACM Computing Surveys*, 56(7):1–35, 2024.

[21] Andrés Herrera-Poyatos, Javier Del Ser, Marcos López de Prado, Fei-Yue Wang, Enrique Herrera-Viedma, and Francisco Herrera. A framework for responsible ai systems: Building societal trust through domain definition, trustworthy ai design, auditability, accountability, and governance. *arXiv preprint arXiv:2503.04739*, 2025.

[22] Jakob Mökander. Auditing of ai: Legal, ethical and technical approaches. *Digital society*, 2(3):49, 2023.

[23] Chiara Picardi, Colin Paterson, Richard David Hawkins, Radu Calinescu, and Ibrahim Habli. Assurance argument patterns and processes for machine learning in safety-related systems. In *Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020)*, pages 23–30. CEUR Workshop Proceedings, 2020.

[24] NIST AI. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, 1(0):0–1, 2023.

[25] Eva Thelisson and Himanshu Verma. Conformity assessment under the eu ai act general approach. *AI and Ethics*, 4(1): 113–121, 2024.

[26] Robert Kilian, Linda Jäck, and Dominik Ebel. European ai standards–technical standardisation and implementation challenges under the eu ai act. *European Journal of Risk Regulation*, 16(3):1038–1062, 2025.

[27] ISO/IEC 5259-3:2024. Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines. Standard, International Organization for Standardization, Geneva, CH, 07 2024.

[28] Jesús Oviedo Lama, Jared David Tadeo Guerrero-Sosa, Moisés Rodríguez Monje, Francisco Pascual Romero Chicharro, and Mario Piattini Velthuis. Towards quality assessment of ai systems: A case study. In *ICSOFT*, pages 254–261, 2025.

[29] Robin Bloomfield and John Rushby. Assurance of ai systems from a dependability perspective. *arXiv preprint arXiv:2407.13948*, 2024.

[30] Alpay Sabuncuoglu, Christopher Burr, and Carsten Maple. Justified evidence collection for argument-based ai fairness assurance. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 18–28, 2025.

[31] Odd Ivar Haugen, Dag McGeorge, Tore Myhrvold, Christian Agrell, and Andreas Hafver. Assurance of ai-enabled systems. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, pages 100–106, 2024.