Communication

Not peer-reviewed version

Engineering Explainable AI Systems for GDPR-Aligned Decision Transparency: A Modular Framework for Continuous Compliance

Antonio Goncalves * and Anacleto Correia

Posted Date: 21 January 2026

doi: 10.20944/preprints202601.1610.v1

Keywords: Explainable AI; transparency; traceability; GDPR; MLOps; decision provenance; 19 model lineage

Communication

Engineering Explainable AI Systems for GDPR-Aligned Decision Transparency: A Modular Framework for Continuous Compliance

Antonio Goncalves and Anacleto Correia

Military University Institute (IUM) (Portuguese Naval Academy) / Centro de Investigação Naval (CINAV), 2810-001 Almada, Portugal

* Correspondence: agoncalveslx@gmail.com

Abstract

Explainability is increasingly expected to support not only interpretation, but also accountability, human oversight, and auditability in high-risk Artificial Intelligence (AI) systems. However, in many deployments, explanations are generated as isolated technical reports, remaining weakly connected to decision provenance, governance actions, audit logs, and regulatory documentation. This short communication introduces *XAI-Compliance-by-Design*, a modular engineering framework for *explainable artificial intelligence* (XAI) systems that routes explainability outputs and related technical traces into structured, audit-ready evidence throughout the AI lifecycle, designed to align with key obligations under the European Union Artificial Intelligence Act (EU AI Act) and the General Data Protection Regulation (GDPR). The framework specifies (i) a modular architecture that separates technical evidence generation from governance consumption through explicit interface points for emitting, storing, and querying evidence, and (ii) a Technical–Regulatory Correspondence Matrix—a mapping table linking regulatory anchors to concrete evidence artefacts and governance triggers. As this communication does not report measured results, it also introduces an *Evidence-by-Design* evaluation protocol defining measurable indicators, baseline configurations, and required artefacts to enable reproducible empirical validation in future work. Overall, the contribution is a practical blueprint that clarifies what evidence must be produced, where it is generated in the pipeline, and how it supports continuous compliance and auditability efforts without relying on post-hoc explanations.

Keywords: EU AI Act; explainable AI; GDPR; governance; machine learning; regulatory compliance; risk management; transparency

1. Introduction

Artificial intelligence (AI) systems for decision support are facing rapidly increasing regulatory and ethical demands for explainability, traceability, and auditability, driven by the General Data Protection Regulation (GDPR) [1]. This regulation imposes stringent obligations regarding transparency and accountability in the context of automated decision making, while the emerging European Union Artificial Intelligence Act (EU AI Act) [2] and the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) standard ISO/IEC 42001 for AI management systems [3] reinforce the need for human oversight, traceability records, and transparency mechanisms for systems classified as high-risk (for example, under Annex III of the EU AI Act) [4,5].

Scope and standards/practices covered. This communication focuses on *governance integration* rather than proposing a new explanation algorithm. Concretely, we target: (i) lifecycle-oriented risk management and documentation duties in the EU AI Act [2]; (ii) accountability, traceability, and transparency expectations under the GDPR [1]; and (iii) implementation-oriented governance practices found in widely adopted AI management and risk frameworks, including ISO/IEC 42001 (AI management systems) [3], ISO/IEC 23894 (AI risk management) [6], and the National Institute

 doi:10.20944/preprints202601.1610.v1

of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF) [7]. We also align with practical Machine Learning Operations (MLOps) documentation patterns (e.g., model cards and datasheets) as *evidence containers* rather than as standalone reports.

Research questions (RQs) and assumptions. To keep claims testable and non-speculative, we structure the contribution around the following RQs:

- **RQ1:** Which explainability artefacts can be operationalised as *compliance evidence* for high-risk AI governance across the lifecycle?
- **RQ2:** How should such artefacts be produced, versioned, and routed so that governance procedures (oversight, audit, incident handling) can consume them consistently?
- **RQ3:** What evaluation indicators and baselines are minimally required to validate the framework without inflating claims beyond available results?

We assume that (a) the operational context determines acceptable thresholds and that (b) governance indicators must remain configurable to accommodate evolving regulatory guidance.

Contributions. We contribute: (C1) a modular architecture with explicit data/control flows between technical and governance functions; (C2) a Technical–Regulatory Correspondence Matrix with a transparent mapping methodology; and (C3) an Evidence-by-Design Evaluation Protocol specifying the artefacts and measurements needed for reproducible validation in subsequent empirical work.

These regulatory initiatives underscore the urgency of developing AI systems that embed transparency mechanisms, enable effective human supervision, and demonstrate clear technical accountability, particularly in contexts that involve substantial ethical, legal, and operational implications [8].

Although research in XAI has advanced considerably [9,10], a significant gap remains between explainability as a conceptual/analytical construct and its systematic operationalisation as governance evidence across the lifecycle, particularly regarding traceability, auditability, and human oversight hooks [6,7].

This communication introduces an engineering-oriented approach that interprets GDPR-inspired compliance as a set of design and operational requirements, rather than as a purely legal or normative interpretation. The core contribution lies in the formulation of a modular framework — XAI-Compliance-by-Design — complemented by a Technical–Regulatory Correspondence Matrix that maps regulatory anchors (e.g., GDPR and EU AI Act obligations) to concrete evidence artefacts (including explainability summaries and metrics such as fidelity and stability, provenance records, and audit-log elements) and to governance triggers for oversight and audit. This approach aims to foster AI systems that are explainable, auditable, and compliant by design, and is aligned with the principles reflected in major European legal and governance references for trustworthy AI and continuous accountability, while recognising that empirical validation is required and is specified in Section 3.3.1 [11].

2. Technical Background and Motivation

The implementation of explainability and compliance principles in artificial intelligence (AI) systems remains fragmented between theory and practice. Despite notable progress in defining explainability objectives and quality indicators such as fidelity and stability [12], their systematic incorporation into engineering pipelines remains inconsistent. This gap hinders the translation of regulatory requirements — including transparency, accountability, and human oversight — into verifiable technical specifications throughout the model lifecycle [9,10].

Under the GDPR [1], the traceability of decision-making processes and the maintenance of technical documentation are key obligations, particularly to ensure the auditability of models that operate as "black boxes". However, many real-world deployments provide only partial lineage and logging, without (i) compliance-oriented evidence schemas, (ii) governance triggers, and (iii) audit-ready bundling that links explanations, human oversight actions, and policy updates into a single verifiable decision record.

 doi:10.20944/preprints202601.1610.v1

The heterogeneity of metrics and the absence of standardisation within the MLOps ecosystem further highlight the need for Compliance-by-Design structures that integrate model governance and drift detection. Although significant advances have been made in XAI, standardised engineering practices for mapping explainability outputs to governance-ready, measurable compliance objectives remain limited, including in emerging domains such as Large Language Models (LLMs) [13].

2.1. Limitations of Existing Governance and MLOps Approaches

Current governance guidance and MLOps practice provide valuable building blocks, but they often fail to close the *evidence gap* between technical explainability outputs and audit-ready compliance artefacts. MLOps platforms typically emphasise deployment automation, experiment tracking, and model versioning, while governance requirements demand traceable decision provenance, human oversight hooks, incident evidence, and documentation that can be verified ex post. Similarly, documentation instruments such as model cards and datasheets improve transparency, yet they do not, by themselves, enforce continuous monitoring, trigger governance actions, or provide audit trails that are verifiable and tamper-evident (e.g., via cryptographic integrity controls).

Relation to mainstream MLOps toolchains. In practice, we treat mainstream MLOps stacks as *lifecycle substrates* and add compliance evidence routing as a thin, modular layer: evidence is emitted at explicit interface points (e.g., after training, validation, deployment, and monitoring), stored as versioned artefacts alongside existing registries/artifact stores, and exposed through queryable bundles that governance processes can consume. This preserves interoperability with MLflow/Kubeflow-style pipelines while making evidence schemas, triggers, and audit-ready bundling first-class engineering requirements.

Common MLOps stacks such as MLflow (experiment tracking and model registry) and Kubeflow Pipelines (workflow orchestration) provide important lifecycle primitives, but they typically do not enforce compliance-oriented evidence schemas, governance triggers, or audit-ready bundling that links explainability artefacts, human oversight actions, and policy updates into a single verifiable decision record [14,15].

To clarify this gap in a *capabilities-based* manner (without numerical claims), Table 1 summarises where common approaches stop short and what the proposed framework adds.

Table 1. Capabilities-based comparison (non-quantitative).

Approach	Typical strengths	Key limitation for compliance evidence routing
Model cards / datasheets	Structured transparency summaries; communicability	Weak linkage to continuous monitoring, triggers, and verifiable audit trails
MLOps platforms (generic)	Automation, versioning, deployment pipelines	Explainability often treated as optional output, not as governance input with defined interfaces
Risk/management frameworks (ISO/NIST)	Governance principles; lifecycle requirements	High-level guidance; requires engineering translation into concrete artefacts and routing logic
***XAI-Compliance-by-Design* (this work)**	Evidence routing; governance triggers; audit-ready artefacts	Requires future empirical validation using the protocol in Section 3.3.1

Accordingly, we propose the XAI-Compliance-by-Design framework, which translates legal obligations into verifiable technical parameters, embedding explainability, auditability, and traceability directly within the MLOps pipeline. This approach seeks to bridge the gap between theory and practice, providing a robust foundation for the development of trustworthy, auditable, and regulation-aligned AI systems from the design phase onwards.

 doi:10.20944/preprints202601.1610.v1

3. Proposed Framework: XAI-Compliance-by-Design

The proposed framework directly addresses the gap identified in Section 2, namely the absence of systematic mechanisms that translate regulatory expectations into verifiable engineering artefacts across the AI lifecycle. In contrast to approaches that treat explainability as an isolated post-hoc output or rely on fragmented governance practices, the framework adopts a dual-flow, layered architecture in which explainability, traceability, and compliance are engineered from the outset and operationalised continuously.

The architecture consists of five functional layers—**Data**, **Model**, **Explanation**, **Audit**, and **Interface**—shown in Figure 1 under two complementary flows that traverse the same layers but emphasise distinct roles: (i) an **Upstream Technical Flow** (solid arrows) that propagates evidence/data from *Data* to *Interface* to generate explainable decisions and technical evidence; and (ii) a **Downstream Compliance Flow** (dashed arrows) that propagates governance feedback in the reverse direction, from *Interface* back to *Data*, enabling human oversight, audit outputs, incident handling, and policy/threshold adjustments that reconfigure technical operations. To avoid ambiguity, the layers are displayed in parallel under each flow to highlight their role-specific responsibilities; they do not represent duplicated system components.

Both flows are coordinated through a central core, the **Compliance-by-Design Engine**, which synchronises three categories of governance-critical state: **XAI metrics**, **decision records**, and **compliance parameters**. As depicted in the figure, the Engine receives *Evidence input* (solid link) from upstream technical execution and issues *Governance feedback* (dashed link) to drive downstream controls. Concretely, governance feedback is materialised as *versioned compliance parameters* (e.g., policy-as-code rules, promotion gates, monitoring thresholds, and retention controls) that are persisted by the Engine and then consumed by the upstream technical flow at explicit enforcement points (training/validation, deployment, inference-time checks, and post-deployment monitoring). This closes the loop by updating executable configuration that directly constrains and reconfigures technical execution. This explicit routing makes compliance operational: evidence is produced in standardised forms, consolidated into decision records, and used to trigger or justify governance actions that can be verified *ex post*. The parameters are versioned and configurable, enabling controlled updates when regulatory guidance or organisational thresholds change while preserving traceability.

Operational definitions (used consistently in this paper). We use the following terms in an implementation-oriented sense:

- **Compliance evidence:** structured technical artefacts (e.g., logs, explanation summaries, drift reports, review decisions) that can be inspected, versioned, and verified during audits.
- **Decision records:** structured decision-level documentation linking inputs, model version, explanation artefacts, and relevant human actions for a given outcome.
- **Upstream technical flow:** evidence/data propagation from Data to Interface producing predictions, explanations, and traceable artefacts.
- **Downstream compliance flow:** governance feedback propagation from Interface back to Data, including review outcomes, incident actions, and policy/threshold updates.
- **Compliance-by-Design Engine:** the coordination core that maintains XAI metrics, decision records, and compliance parameters, and mediates evidence input and governance feedback.

Figure 1 summarises the dual-flow architecture and the central Compliance-by-Design Engine.

Implementation note (interfaces). The engine can be realised through standard MLOps integration surfaces (artefact registry APIs, evidence-store query interfaces, message buses, and policy-as-code repositories).

Design patterns for overhead, security, and scalability. To keep the real-time decision path lightweight without weakening governance, the architecture supports a set of implementable patterns across the **Data**, **Model**, **Explanation**, **Interface**, and **Audit** layers: (i) *tiered evidence generation*, where a minimal provenance record is emitted on every decision (timestamps, version identifiers, policy version, and input/output fingerprints), while higher-fidelity explanation artefacts are computed in audit

 doi:10.20944/preprints202601.1610.v1

or sampling modes; (ii) *asynchronous evidence emission* via message queues to decouple inference latency from logging throughput; (iii) *sampling and windowed computation* for post-deployment surveillance to control cost while maintaining coverage; (iv) *caching and deduplication* of explanation summaries and baseline artefacts to reduce recomputation; and (v) *secure evidence stores* using append-only logs, integrity protection (e.g., hash chaining and signed manifests), role-based access control, encryption, and retention minimisation. These patterns are treated as explicit configuration choices whose overhead and governance impact are to be reported under the evaluation protocol in Section 3.3.1.

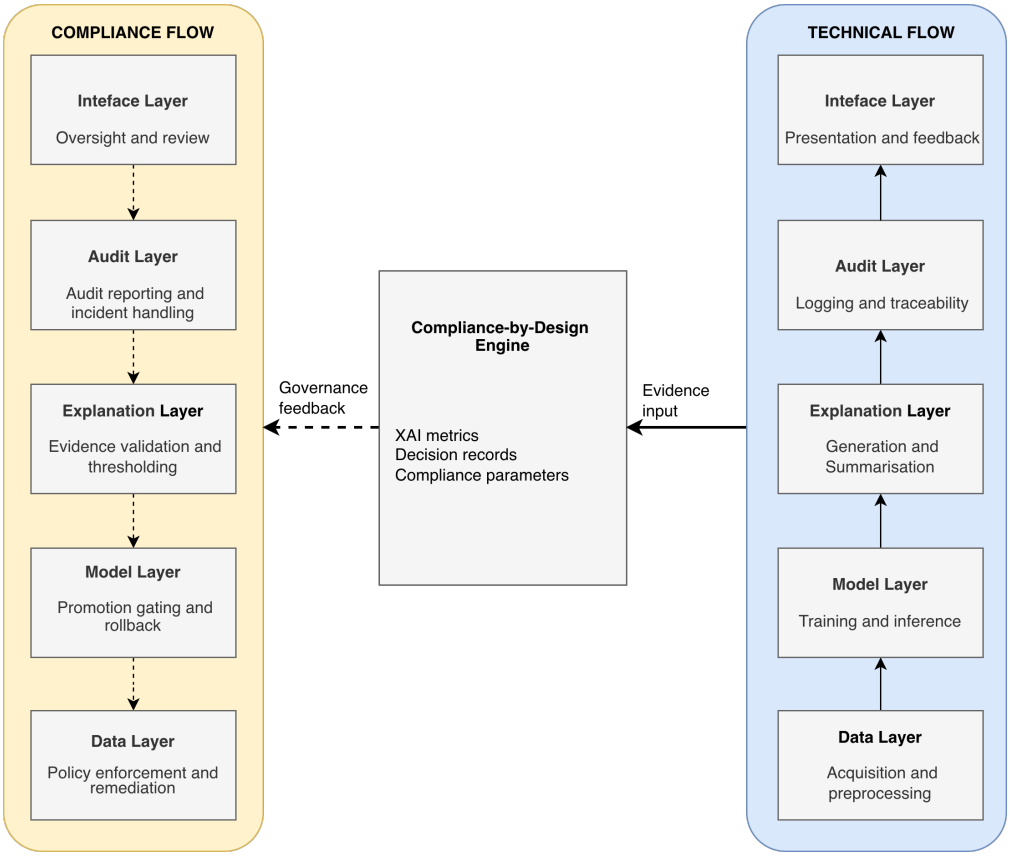

Figure 1. XAI-Compliance-by-Design dual-flow architecture and central Compliance-by-Design Engine. Solid arrows represent the **upstream technical flow** from the Data layer to the Interface layer, producing predictions, explanation artefacts, and traceable evidence. Dashed arrows represent the **downstream compliance flow** from the Interface/Audit layers back to the Data and Model layers, capturing governance feedback such as human review outcomes, incident actions, and updates to policies and thresholds. The Compliance-by-Design Engine coordinates both flows through versioned configuration (e.g., policy-as-code, triggers, and thresholds) and audit-ready evidence bundles, enabling replay and verification while treating audit logs as security-critical assets (integrity protection, access control, encryption, and retention minimisation). EU AI Act = European Union Artificial Intelligence Act; GDPR = General Data Protection Regulation; XAI = explainable artificial intelligence.

3.1. Layers and Responsibilities

To remain consistent with the dual-flow representation in Figure 1, each layer is described by its **technical role** (upstream) and its **compliance role** (downstream).

Data Layer. *Technical role (Acquisition and preprocessing):* acquires, cleans, validates, and prepares data for model operations, maintaining traceable data lineage. *Compliance role (Policy enforcement and remediation):* applies governance constraints such as minimisation, purpose limitation, retention, and remediation actions triggered by audits or incidents.

Model Layer. *Technical role (Training and inference):* trains and executes models, ensuring versioning and reproducibility of model artefacts and configurations. *Compliance role (Promotion gating and rollback):* enforces promotion gates (e.g., staging to production) and supports rollback or deprecation when governance criteria are not met.

 doi:10.20944/preprints202601.1610.v1

Explanation Layer. *Technical role (Generation and summarisation):* generates local and global explanations (e.g., SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME)) and produces concise explanation summaries for downstream consumption. *Compliance role (Evidence validation and thresholding):* validates explanation artefacts against configurable criteria (e.g., stability/fidelity thresholds) [9,10,12] and updates thresholds/policies when warranted.

Audit Layer. *Technical role (Logging and traceability):* records operational events and maintains traceability across inputs, model versions, explanation artefacts, and outcomes (decision-level traceability). *Compliance role (Audit reporting and incident handling):* produces audit-ready outputs, supports incident workflows, and structures evidence for regulatory verification and accountability.

Interface Layer. *Technical role (Presentation and feedback):* delivers user-facing outputs and explanations and captures feedback signals relevant to system operation. *Compliance role (Oversight and review):* supports human oversight, review and override mechanisms, and escalation paths, feeding validated governance outcomes back into lower layers via the compliance flow.

3.2. Technical–Regulatory Correspondence Matrix

The *Technical–Regulatory Correspondence Matrix* establishes the relationship between regulatory anchors, evidence artefacts, and governance triggers, operationalising how engineering mechanisms can be audited against compliance expectations.

Mapping methodology (how Table 2 is constructed). The matrix is built using a rule-based mapping: (1) identify a governance requirement (AI Act/GDPR principle) that implies a *verifiable* duty; (2) derive the minimal *evidence artefact* that would demonstrate satisfaction of that duty during an audit; (3) associate the technical mechanism (XAI metric/technique, logging, monitoring) that produces or supports that artefact; and (4) specify a governance *trigger* (qualitative or quantitative) that would prompt review or mitigation. Where thresholds are domain-dependent, the framework treats them as configurable parameters defined by the organisation's monitoring and governance policies.

Table 2. Technical–Regulatory Correspondence Matrix with explicit evidence artefacts and governance triggers.

Engineering principle	Regulatory anchor (illustrative)	Evidence artefact (audit-ready)	Mechanism + trigger (configurable)
Transparency	EU AI Act (documentation duties); GDPR transparency	Model card + explanation summary; user-facing rationale template	XAI summaries (global/local); Trigger: explanation stability below the configured monitoring criterion
Traceability	EU AI Act (logging/record-keeping); GDPR accountability	Decision provenance record; model/data lineage log	Versioned logs + hash-chaining; Trigger: provenance schema completeness check fails (missing required fields)
Human oversight	EU AI Act (human oversight)	Review/override records; escalation notes	Human-in-the-loop checkpoints; Trigger: decision flagged as high-risk under the organisational risk taxonomy
Technical accountability	EU AI Act (technical documentation); GDPR accountability	Reproducible run bundle; signed artefact manifest	Automated evidence bundling; Trigger: integrity verification failure (hash/manifest mismatch)
Risk minimisation	EU AI Act (risk management); GDPR data protection by design	Drift report; incident report; mitigation ticket	Monitoring + drift metrics; Trigger: drift exceeds the configured threshold in the drift monitoring policy

3.3. Conceptual Mini-Case

To illustrate the application of the proposed framework, consider a **decision-support system for network anomaly detection**, designed to identify potentially malicious behaviour in real time while ensuring explainability and regulatory compliance from its inception.

 doi:10.20944/preprints202601.1610.v1

Data Layer. Manages the acquisition and preprocessing of network traffic, applying pseudonymisation to protect sensitive information in line with GDPR principles of minimisation and purpose limitation.

Model Layer. Trains and versions algorithms such as *autoencoders* or *random forests*, storing all versions and parameters in a technical registry to ensure traceability and audit readiness.

Explanation Layer. Employs *SHAP* values to produce local explanations, identifying how variables such as source, destination, protocol, and traffic volume contribute to anomaly classification.

Audit Layer. Records decisions, explanations, and human interventions, monitoring *data drift* and *concept drift*, and triggering alerts upon model degradation or deviation from expected behaviour.

Interface Layer. Adapts explanations to different user profiles and allows controlled *override* of automated decisions, closing the feedback loop and supporting continuous human supervision.

This conceptual case demonstrates how explainability and compliance become structural—not external—components of the AI system, enabling auditability and sustained trust. It illustrates how legal and ethical requirements can be operationalised into auditable engineering mechanisms, and how the same routing logic can be adapted to other transparency-critical AI domains, subject to empirical validation under the protocol in Section 3.3.1.

3.3.1. Evidence-by-Design Evaluation Protocol

Because this communication does not report empirical results, we define a minimal evaluation protocol that enables reproducible validation in future work without inflating claims. The protocol specifies: (i) **baselines** (e.g., MLOps pipeline with documentation-only controls vs. evidence-routing governance); (ii) **indicators** covering performance, explainability, governance coverage, and overhead; and (iii) **artefacts** that must be produced for auditability.

Planned indicators (to be instantiated during empirical evaluation).

- **Model performance:** task-appropriate metrics (e.g., classification/regression measures aligned with the intended purpose).
- **Explainability quality:** stability/consistency of explanations across time windows (assessed via a predefined stability criterion).
- **Governance coverage:** share of decisions with complete provenance and review trace (checked against a required evidence schema).
- **Operational overhead:** latency/compute impact of explanation generation and evidence logging (measured under representative workloads).
- **Monitoring efficacy:** drift detection sensitivity and incident-response trace completeness (evaluated against a documented incident playbook).

Artefacts to be produced (audit-ready bundle).

- Versioned evidence bundle (inputs, model version, explanation summary, provenance fields, reviewer actions).
- Change log of governance parameters (policies/thresholds) with timestamps and rationale.
- Minimal report templates: incident report, audit extract, and compliance checklist.

4. Technical Discussion and Practical Implications

The proposed framework enables the **native and continuous integration of explainability** throughout the AI lifecycle, making transparency and compliance intrinsic to both development and operation. It translates high-level obligations in the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act (EU AI Act) into *verifiable technical artefacts* (evidence schemas, provenance records, explanation summaries, and audit-ready bundles) that can be inspected, replayed, and governed over time.

From an engineering standpoint, implementing *XAI-Compliance-by-Design* within **MLOps** requires balancing **explainability and computational efficiency**. Methods such as *SHAP*, *LIME*, and *counterfactual analysis* can introduce non-trivial overhead (compute, latency, and storage), especially in

 doi:10.20944/preprints202601.1610.v1

time-sensitive settings. The framework therefore treats explanation fidelity, sampling regimes, logging frequency, and evidence retention as explicit *design parameters* that must be configured, justified, and reported under the evaluation protocol in Section 3.3.1.

4.1. Overhead, Log Security, and Regulatory Evolution

Operational overhead. Explainability computation and evidence logging can impose measurable runtime and storage costs. Rather than weakening governance, the framework supports concrete, framework-compatible mitigations: (i) *tiered explanations* (full-fidelity explanations in audit/review contexts; summarised or sampled explanations in production), (ii) *windowed and sampled computation* for post-deployment surveillance to control cost while maintaining coverage, (iii) *caching* of explanation summaries and reusable baselines, (iv) *monitoring-only replicas* using simplified models for explanation regression testing, and (v) *asynchronous evidence emission* via message queues to decouple inference latency from logging throughput. These mitigations are treated as configurable engineering choices whose impact (overhead, coverage, and governance adequacy) must be documented and evaluated via Section 3.3.1.

Audit log security and privacy. Auditability requires retained evidence, but evidence must remain protected and proportionate. We therefore treat evidence stores as security-critical assets: append-only logging, integrity protection (e.g., hash chaining and signed manifests), strict role-based access control, encryption in transit and at rest, and retention minimisation aligned with the GDPR [1]. Where explanation artefacts may increase leakage risk (e.g., revealing sensitive feature contributions or enabling model inversion or membership inference), the framework supports redaction, aggregation, and *tiered access* (technical vs. managerial views), together with explicit access logging and periodic review of evidence necessity. This positions evidence preservation as an accountability control that remains bounded by proportionality and purpose limitation.

Audit scalability and reproducibility. As systems grow in complexity and volume, evidence collection and verification can become resource-intensive. To reduce manual burden without claiming measured improvements, the framework promotes automation through metadata tracking, version control of artefacts and policies, and standardised evidence schemas. These mechanisms strengthen reproducibility by enabling consistent reconstruction of decision provenance and the corresponding oversight context (model/data versions, explanation summaries, governance checks, and human review actions).

Continuous compliance under drift. Operational maintenance depends on monitoring *data drift* and *concept drift*, as well as *explainability regression* (stability of explanations and shifts in feature attributions over time). By treating drift signals and explanation stability as governance-relevant indicators, the framework supports automated triggers for review and recalibration, ensuring that transparency properties do not silently degrade while conventional performance metrics remain acceptable.

Regulatory change as a lifecycle condition. To remain operational under evolving guidance, governance parameters (policies, triggers, thresholds, documentation requirements) are treated as *versioned configuration items*. This enables controlled updates and traceability when legal interpretations or obligations change, without rewriting technical modules. Recent initiatives such as the European Commission's Digital Omnibus package illustrate that obligations and timelines may be revised, reinforcing the need for versioned, auditable governance parameters rather than hard-coded compliance assumptions [16].

Bias and fairness as core risk signals. Bias and fairness are increasingly recognised as central governance and legal concerns. We integrate bias/fairness checks as first-class governance indicators (not optional diagnostics), producing evidence artefacts that support audit inspection, human oversight, and traceable escalation. Where thresholds are domain- and risk-context dependent, they are treated as configurable governance parameters to be instantiated and reported under Section 3.3.1 [17].

From a practical perspective, adopting the framework *may* reduce manual audit effort by structuring evidence generation, retrieval, and replay into auditable bundles; however, any such benefit is an

 doi:10.20944/preprints202601.1610.v1

evaluation target rather than a demonstrated outcome, and must be empirically assessed under the protocol in Section 3.3.1.

5. Conclusions and Future Perspectives

The *XAI-Compliance-by-Design* framework is presented as an engineering-oriented blueprint that links explainability artefacts to governance processes through explicit evidence routing, traceable decision provenance, and audit-ready documentation. Rather than claiming validated improvements, this communication clarifies what evidence must be produced, how it should be versioned and protected, and how governance parameters can be updated across the lifecycle under the EU AI Act and the GDPR. A key limitation is that empirical results are not reported here; accordingly, we provide an Evidence-by-Design Evaluation Protocol that defines indicators, baselines, and artefacts required for reproducible validation. Future work will instantiate the protocol in domain case studies, quantify overhead and governance coverage, and assess how explanation stability and monitoring signals can operationalise risk management in practice.

Its main contribution lies in translating legal concepts such as accountability, transparency and human oversight into verifiable technical artefacts and interfaces that can support continuous governance workflows. Rather than asserting validated benefits, the framework provides an engineering blueprint and an Evidence-by-Design protocol (Section 3.3.1) that specifies what must be measured and reported to substantiate compliance-relevant claims in future empirical studies.

Future work should focus on the following priorities:

1. **Empirical validation** of explainability metrics (*fidelity*, *stability*, *comprehensibility*) across real-world domains;
2. **Development of continuous compliance dashboards** integrating both technical and regulatory indicators in real time;
3. **Extension to distributed environments** through *federated learning* and *differential privacy*, enabling transparency without compromising confidentiality.

In summary, this framework constitutes a concrete step towards aligning explainable AI engineering with regulatory compliance, providing a robust foundation for trustworthy, auditable, and regulation-ready sustainable intelligent systems.

Author Contributions: Conceptualization, António Gonçalves and Anacleto Correia; methodology, António Gonçalves; software, António Gonçalves; validation, António Gonçalves and Anacleto Correia; formal analysis, Anacleto Correia; investigation, António Gonçalves; resources, Anacleto Correia; writing original draft preparation, António Gonçalves; writing—review and editing, António Gonçalves and Anacleto Correia; visualization, António Gonçalves; supervision, Anacleto Correia; project administration, António Gonçalves and Anacleto Correia; funding acquisition, Anacleto Correia. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The case study uses synthetically generated data produced by the described workflow; no personal data were used. Data can be regenerated from the configuration and generation procedures reported in this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. Accessed: 2025-12-12.

 doi:10.20944/preprints202601.1610.v1

2. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 — Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, 2024. Accessed: 2025-12-12.
3. International Organization for Standardization. ISO/IEC 42001:2023 — Artificial intelligence management system. Standard. https://www.iso.org/standard/81230.html, 2023. Accessed: 2025-12-12.
4. Thomaidou, A.; Limniotis, K. Navigating Through Human Rights in AI: Exploring the Interplay Between GDPR and Fundamental Rights Impact Assessment. *J. Cybersecur. Priv.* **2025**, *5*, 7. https://doi.org/10.3390/jcp5010007.
5. Sovrano, F. Metrics, Explainability and the European AI Act Proposal. *J* **2022**, *5*, 10. https://doi.org/10.3390/j5010010.
6. International Organization for Standardization. ISO/IEC 23894:2023 — Artificial intelligence — Risk management. Standard, 2023. Accessed: 2025-12-12.
7. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework, 2023. Accessed: 2025-12-12.
8. Ahangar, M.N.; Jalali, S.M.J.; Dastjerdi, A.V. AI Trustworthiness in Manufacturing: Challenges, Toolkits and Best Practices. *Sensors* **2025**, *25*, 4357. https://doi.org/10.3390/s25144357.
9. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. https://doi.org/10.3390/e23010018.
10. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. https://doi.org/10.3390/app11115088.
11. Feretzakis, G.; Vagena, E.; Kalodanis, K.; Peristera, P.; Kalles, D.; Anastasiou, A. GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet* **2025**, *17*, 151. https://doi.org/10.3390/fi17040151.
12. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Pedreschi, D.; Giannotti, F. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* **2018**, *51*, 93:1–93:42. https://doi.org/10.1145/3236009.
13. Ranaldi, L. Survey on the Role of Mechanistic Interpretability in Generative AI. *Big Data and Cognitive Computing* **2025**, *9*, 193. https://doi.org/10.3390/bdcc9080193.
14. MLflow. MLflow Documentation: Model Registry Tutorial. https://mlflow.org/docs/3.6.0/ml/model-registry/tutorial/, 2025. Accessed: 2025-12-12.
15. Kubeflow. Kubeflow Pipelines: Getting Started. https://www.kubeflow.org/docs/components/pipelines/getting-started/, 2025. Accessed: 2025-12-12.
16. European Commission. Simpler EU digital rules and new digital wallets to save time and money for businesses and citizens. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2718, 2025. Accessed: 2025-12-12.
17. Lendvai, G.F.; Gosztonyi, G. Algorithmic Bias as a Core Legal Dilemma in the Age of Artificial Intelligence: Conceptual Basis and the Current State of Regulation. *Laws* **2025**, *14*, 41. https://doi.org/10.3390/laws14030041.