ISSN: 3049-3277

International Journal of Integrative Studies (IJIS)

Journal homepage: www.ijis.co.in

RESEARCH ARTICLE

Trustworthy Artificial Intelligence Models for Regulated Decision Ecosystems: A Systematic and Empirical Framework

[1]Suresh Babu Narra

[1]Solutions Architect – AI, Machine Learning & Generative AI Cincinnati, Ohio, USA, narra.sureshbabu@gmail.com
ORCID ID: https://orcid.org/0009-0002-5429-0202

ABSTRACT

The proliferation of artificial intelligence (AI) in regulated domains — including consumer credit, clinical diagnosis, and public administration — has intensified regulatory and societal scrutiny of algorithmic decision-making. While predictive performance has improved substantially, persistent deficiencies in fairness, explainability, and accountability impede institutional adoption and regulatory compliance. This paper proposes and empirically evaluates the Regulated Trustworthy AI Lifecycle (RTAIL), a modular framework that integrates fairness-aware learning, post hoc interpretability, and robustness validation into a structured governance lifecycle. Using a synthetic loan approval dataset (N = 10,000; 18 features; sensitive attributes: gender, age group), we compare a logistic regression baseline against an RTAIL-compliant model employing Kamiran and Calders (2012) instance reweighting and SHAP-based explanation consistency. Across five-fold stratified cross-validation, the RTAIL model reduces the Demographic Parity Difference from 0.24 ± 0.03 to 0.08 ± 0.01 (66.7% reduction) and the Equal Opportunity Difference from 0.19 ± 0.02 to 0.06 ± 0.01 (68.4% reduction), at a statistically non-significant accuracy cost of 3.2 percentage points (McNemar p = 0.09). SHAP feature-importance consistency improves from 0.54 to 0.91. Ablation analysis confirms that the fairness constraint is the dominant driver of bias reduction. These results demonstrate that trustworthiness is an achievable, measurable property of AI system design — not merely a normative aspiration — and that the accuracy-fairness trade-off is navigable within operationally acceptable bounds for high-stakes regulated applications.

Keywords: *Trustworthy Artificial Intelligence; Regulated Decision Ecosystems; Explainable AI (XAI); Algorithmic Fairness; AI Governance; Bias Mitigation; Model Interpretability; Robust Machine Learning; Ethical AI Systems; Regulatory Compliance.*

INTRODUCTION

In 2023, the European Union enacted Regulation (EU) 2024/1689 — the AI Act — establishing legally binding requirements for transparency, human oversight, and risk management in high-risk AI systems deployed across employment, credit, healthcare, and law enforcement (European Parliament, 2024). This regulatory milestone reflects a broader recognition that the societal consequences of AI-driven decisions demand governance structures that extend well beyond predictive accuracy. When an algorithm denies a loan, rejects a job application, or stratifies patients into treatment tiers, the legitimacy of that decision depends not only on its statistical properties but on its demonstrable fairness, traceability, and accountability.

Despite substantial progress in machine learning, the dominant paradigm of AI development remains principally oriented toward optimising predictive performance metrics, often at the expense of ethical and regulatory considerations (Li et al., 2022; Liu et al., 2022).

[1]Solutions Architect – AI, Machine Learning & Generative AI Cincinnati, Ohio, USA
Email-id: narra.sureshbabu@gmail.com.

Corresponding Author: Suresh Babu Narra,
Solutions Architect – AI, Machine Learning & Generative AI Cincinnati, Ohio, USA.
Email-id: narra.sureshbabu@gmail.com.

DOI: https://doi.org/10.63856/ijis/v2i5/00041

How to cite this article: Sharma,, V., (2026). Trustworthy Artificial Intelligence Models for Regulated Decision Ecosystems: A Systematic and Empirical Framework. *International Journal of Integrative Studies*, 2(5), 24-31.

Source of support: Nil

Conflict of interest: None.
Received: 18//04/2026 **Revised:** 22/04/2026 **Accepted:** 27/04/2026
Published: 12/05/2026

This misalignment produces what Díaz-Rodríguez et al. (2023) describe as a trust deficit: capable AI systems that cannot be safely deployed in regulated environments because they lack the transparency and fairness guarantees that institutions and regulators require. Studies of credit scoring systems (Barocas et al., 2023), recidivism prediction (Dwork et al., 2012), and medical triage have documented systematic biases that compound pre-existing societal inequities and expose deploying organisations to legal and reputational risk.

Existing approaches to trustworthy AI tend to address individual dimensions in isolation — fairness interventions applied post hoc, explainability tools added as an afterthought, or governance frameworks that remain at the level of ethical principles without operational specification (Calegari et al., 2024; Vyhmeister & Castañé, 2024). What is lacking is an integrated lifecycle framework that embeds fairness, explainability, and robustness as first-class design objectives from the earliest stages of model development through ongoing deployment monitoring. The NIST AI Risk Management Framework (NIST, 2023) provides high-level guidance, but does not prescribe the technical mechanisms by which its governance goals are to be achieved.

This paper addresses that gap through two principal contributions. First, we present the Regulated Trustworthy AI Lifecycle (RTAIL), a modular framework that operationalises trustworthy AI across four governance phases: (1) ethical scoping and regulatory alignment; (2) fairness-aware model development; (3) interpretability integration and audit; and (4) continuous monitoring and accountability. Second, we provide the first empirical evaluation of this integrated framework using a simulated loan approval classification task, demonstrating that RTAIL achieves substantial, statistically significant improvements in algorithmic fairness while incurring only a modest and operationally acceptable reduction in predictive performance.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature across the four pillars of trustworthy AI. Section 3 states the research questions. Section 4 describes the RTAIL framework architecture. Section 5 presents the methodology. Section 6 reports experimental results, including ablation analyses. Section 7 discusses findings and regulatory implications. Section 8 acknowledges limitations and directions for future work. Section 9 concludes.

2. Literature Review

2.1 Foundations of Trustworthy AI

The conceptual foundations of trustworthy AI have been progressively formalised over the past decade. Li et al. (2022) provide a landmark survey identifying six core dimensions of trustworthiness — safety, fairness, privacy, robustness, explainability, and accountability — and argue that these must be treated as jointly constitutive properties of a trustworthy system rather than independent objectives. Liu et al. (2022) extend this analysis from a computational perspective, characterising the algorithmic techniques available for each dimension and the trade-offs they introduce. Díaz-Rodríguez et al. (2023) synthesise AI ethics principles, technical requirements, and regulatory standards into a unified reference model, establishing that trustworthy AI is a multidimensional construct requiring both technical and institutional governance.

A significant limitation of these foundational works is their predominantly conceptual character. They describe what trustworthy AI should be rather than demonstrating how an integrated system can be built and measured empirically. This paper contributes precisely that empirical dimension, operationalising the dimensions identified by Li et al. (2022) within a structured experimental evaluation.

2.2 Algorithmic Fairness

Algorithmic fairness has emerged as one of the most actively researched dimensions of trustworthy AI, driven in part by documented instances of discriminatory outcomes in consequential domains. Barocas et al. (2023) provide the definitive treatment of fairness in machine learning, distinguishing between individual fairness (similar individuals should receive similar predictions) and group fairness (aggregate outcomes should not differ systematically across demographic groups). Dwork et al. (2012) formalise fairness through awareness, demonstrating that ignoring sensitive attributes does not prevent discriminatory outcomes.

Two group-fairness criteria are particularly relevant to the regulated credit context of this study. Demographic parity requires that the probability of a positive prediction be equal across demographic groups. Equal opportunity (Hardt et al., 2016) requires that true positive rates be equalised. These criteria are formally incompatible under general conditions (Chouldechova, 2017), a theoretical result that motivates the trade-off analysis in Section 6.

Bias mitigation techniques operate at three points in the machine learning pipeline. Pre-processing methods modify the training data to reduce bias before model training; instance reweighting (Kamiran & Calders, 2012) — the technique employed in this study — assigns sample weights inversely proportional to the product of class label and sensitive attribute probabilities. In-processing methods incorporate fairness constraints into the optimisation objective (Zafar et al., 2017). Post-processing methods adjust decision thresholds on a trained model (Hardt et al., 2016). Each approach introduces different accuracy-fairness trade-offs; pre-processing is preferred here for its compatibility with any downstream classifier.

2.3 Explainable AI (XAI)

Explainability — the capacity of a system to provide human-interpretable accounts of its decisions — is an increasingly explicit regulatory requirement. Article 13 of the EU AI Act mandates transparency for high-risk AI systems, and the CFPB's adverse action notice requirements implicitly demand explanation of individual credit decisions. Lundberg and Lee (2017)

introduce SHAP (SHapley Additive exPlanations), a theoretically principled approach to feature attribution grounded in cooperative game theory, which has become the de facto standard for post hoc explanation of tabular models. SHAP values assign each feature a contribution score for a given prediction, enabling both global (population-level) and local (individual-decision) interpretability.

Despite its widespread adoption, the consistency and stability of SHAP explanations across model runs and data perturbations has received limited empirical attention. Papademas et al. (2025) and Burla and Priya (2025) identify explanation consistency — the degree to which the same features are identified as most important across runs — as a key dimension of explainability that is distinct from mere feature attribution. This study measures SHAP-based consistency quantitatively, contributing evidence on how fairness constraints affect explanation stability.

2.4 AI Governance and Regulatory Frameworks

Governance frameworks provide the institutional architecture within which technical trustworthy-AI mechanisms are embedded. The NIST AI Risk Management Framework (NIST, 2023) organises AI risk management around four functions — Govern, Map, Measure, and Manage — providing a process model for identifying, assessing, and mitigating AI risks across the system lifecycle. The EU AI Act (European Parliament, 2024) establishes a risk-tiered regulatory regime, with the highest obligations applying to AI systems used in credit scoring, employment screening, and healthcare, precisely the domains targeted by RTAIL.

Calegari et al. (2024, 2025) identify a critical gap between individual-dimension trustworthiness assessments and system-level reliability: ensuring that each component satisfies local trustworthiness criteria is necessary but not sufficient for overall system trustworthiness. This motivates the integrated lifecycle approach of RTAIL. Vyhmeister and Castañé (2024) extend this analysis to a project risk management context, showing that governance failures arise predominantly at lifecycle boundaries — transitions between design, development, and deployment — rather than within individual phases. RTAIL is explicitly designed to manage these boundary risks.

2.5 Gap Analysis and Positioning of RTAIL

Despite the richness of the literature across these four pillars, no prior work has proposed and empirically evaluated an integrated lifecycle framework that simultaneously addresses fairness, explainability, robustness, and governance within a single coherent methodology. Existing empirical studies typically evaluate one or two dimensions in isolation, use non-reproducible proprietary datasets, and report point estimates without confidence intervals or statistical significance testing. Nastoska et al. (2025) and Vaghela (2025) propose evaluation metrics and accountability scores but do not provide lifecycle integration or experimental validation. RTAIL fills this gap by providing both a governance architecture and the first cross-validated empirical evidence of its efficacy.

3. Research Questions

This study is structured around three formally stated research questions:

RQ1 (Fairness): Does the RTAIL framework produce a statistically significant reduction in group-level bias — as measured by Demographic Parity Difference (DPD) and Equal Opportunity Difference (EOD) — relative to a performance-optimised baseline classifier, without exceeding a 5 percentage-point reduction in F1-score?

RQ2 (Explainability): Does the RTAIL framework produce more consistent and interpretable feature-importance attributions than the baseline, as measured by SHAP-based top-k feature agreement (Spearman rank correlation ρ) and SHAP consistency score across cross-validation folds?

RQ3 (Trade-off): What is the quantitative shape of the accuracy-fairness trade-off introduced by RTAIL's fairness constraints, and is the accuracy reduction statistically significant under McNemar's test?

4. The RTAIL Framework

The Regulated Trustworthy AI Lifecycle (RTAIL) is a four-phase governance framework designed to embed trustworthiness as a first-class property throughout the AI system development and deployment lifecycle. Each phase produces documented artefacts that support regulatory audit and third-party verification.

Phase 1: Ethical Scoping and Regulatory Alignment

Prior to any data collection or model development, RTAIL requires the production of an Ethical Scope Document (ESD) that specifies: (i) the regulated domain and applicable legal frameworks (e.g., Equal Credit Opportunity Act, EU AI Act risk classification); (ii) the sensitive attributes and protected groups relevant to the deployment context; (iii) the fairness criteria to be enforced and the acceptable tolerance thresholds for each; and (iv) the explainability obligations arising from the regulatory context (e.g., adverse action notice requirements). The ESD anchors all subsequent technical decisions to a documented governance intent and serves as the primary reference document for regulatory audit.

Phase 2: Fairness-Aware Model Development

Model development under RTAIL incorporates fairness constraints at the pre-processing stage using instance reweighting (Kamiran & Calders, 2012). Sample weights w_i are assigned to each training instance i as:

$$w_i = P(Y) \cdot P(S) / (P(Y, S))$$

where Y is the class label and S is the sensitive attribute. This upweights underrepresented group-label combinations without altering the feature distribution, preserving compatibility with standard classifiers. A regularisation penalty λ is added to the loss function to promote robustness, with λ selected by cross-validated grid search.

Phase 3: Interpretability Integration and Audit
All predictions are accompanied by SHAP explanations generated using the TreeExplainer or LinearExplainer as appropriate to the model family. RTAIL mandates that explanation consistency be measured across cross-validation folds using the Spearman rank correlation of top-k feature importance rankings, with k = 5. A consistency score below 0.80 triggers a model review. Each model version is documented in a Model Card (Mitchell et al., 2019) capturing performance, fairness, and explanation metadata, supporting third-party audit.

Phase 4: Continuous Monitoring and Accountability
Post-deployment, RTAIL prescribes periodic re-evaluation of fairness metrics on production data (at minimum quarterly), with automatic alerts when DPD or EOD exceed the thresholds established in the ESD. An accountability score is computed as a weighted aggregate of fairness, robustness, and explanation consistency metrics (Vaghela, 2025), enabling longitudinal tracking of system trustworthiness. All monitoring results are logged to an immutable audit trail, supporting traceability requirements under the EU AI Act.

5. Methodology

5.1 Research Design

This study employs an experimental comparative design. Two models — a performance-optimised baseline and an RTAIL-compliant trustworthy model — are trained on an identical dataset and evaluated against a pre-specified set of performance, fairness, and explainability metrics. A five-fold stratified cross-validation protocol is applied to all training and evaluation procedures, and all reported metrics are presented as means with 95% confidence intervals. Statistical significance of performance differences is assessed using McNemar's test on the aggregated cross-validation predictions.

5.2 Dataset and Preprocessing

The study uses a synthetic loan approval dataset generated to simulate the statistical properties of real-world consumer credit portfolios while preserving full data governance control. The dataset comprises 10,000 records and 18 features, including demographic attributes (gender, age group [18–35, 36–55, 56+], marital status), financial attributes (annual income, credit score, debt-to-income ratio, loan amount requested, employment tenure), and loan outcome variables (approved/denied). The binary outcome (loan approval) is distributed at 62% positive (approved) to 38% negative (denied), reflecting empirical base rates in consumer credit.

Sensitive attributes are gender (binary: male/female, 51%/49%) and age group (three levels). Intersectional bias is present in the data-generating process, with female applicants in the 18–35 age bracket approved at 48% versus 71% for male applicants in the same bracket — a disparity consistent with documented patterns in consumer credit literature (Barocas et al., 2023).

Preprocessing proceeds in five steps: (1) mean imputation for three features with missing rates below 2%; (2) z-score normalisation of all continuous features; (3) one-hot encoding of categorical variables; (4) computation of instance weights per Kamiran and Calders (2012) for the RTAIL model only; and (5) stratified random splitting into five cross-validation folds, with stratification on both outcome label and sensitive attribute to ensure balanced representation in every fold.

5.3 Model Specifications

5.3.1 Baseline Model

The baseline model is a logistic regression classifier implemented in scikit-learn 1.4 (Pedregosa et al., 2011) with L2 regularisation strength C = 1.0 (default), solver = lbfgs, and maximum iterations = 1,000. No fairness constraints or sample reweighting are applied. This classifier is selected because its linear decision boundary provides a meaningful point of comparison for fairness interventions and because its intrinsic interpretability makes it a natural benchmark in the algorithmic fairness literature.

5.3.2 RTAIL Trustworthy Model

The RTAIL model uses an identical logistic regression specification with the following additions: (i) instance weights computed per Phase 2 of the RTAIL framework; (ii) SHAP LinearExplainer applied post hoc to generate feature attributions for all test instances; and (iii) explanation consistency measured across folds as the mean Spearman rank correlation of the top-5 SHAP features. All random seeds are fixed at 42 across both models to ensure reproducibility.

5.4 Evaluation Metrics

Performance Metrics

- Accuracy, Precision, Recall, and F1-Score (macro-averaged), reported as mean ± 95% CI across five folds.

Fairness Metrics

- Demographic Parity Difference (DPD): $|P(\hat{Y}=1 \mid S=\text{male}) - P(\hat{Y}=1 \mid S=\text{female})|$. Target: DPD ≤ 0.10.
- Equal Opportunity Difference (EOD): $|TPR_{male} - TPR_{female}|$. Target: EOD ≤ 0.10.

Explainability Metrics

- Top-5 SHAP Feature Agreement (Spearman ρ): mean rank correlation of top-5 features across fold pairs.
- SHAP Consistency Score: proportion of cross-validation folds in which the same top-3 features appear.
- Mean Explanation Depth: average number of features required to account for 90% of cumulative SHAP magnitude.

Statistical Testing

- McNemar's test on aggregated cross-validation predictions to assess statistical significance of accuracy differences.

- Paired t-test on fold-level fairness metrics to confirm significance of fairness improvement.

5.5 Ablation Study Design

To isolate the contribution of each RTAIL component, four model variants are evaluated in addition to the baseline and full RTAIL model: (i) fairness constraint only (reweighting, no explainability enforcement, no robustness regularisation); (ii) explainability only (SHAP consistency constraint, no reweighting, standard regularisation); (iii) robustness only (increased regularisation, no reweighting, no SHAP enforcement); and (iv) full RTAIL. This design allows attribution of observed fairness and performance effects to specific framework components.

6. Results

6.1 Predictive Performance (RQ3)

Table 1. Predictive performance metrics across five-fold stratified cross-validation (mean ± 95% CI). pp = percentage points. McNemar's test on accuracy: $\chi^2 = 2.89$, $p = 0.09$ (not significant at $\alpha = 0.05$).

Metric	Baseline Model	RTAIL Model	Δ Change
Accuracy	87.3% ± 1.2%	84.1% ± 1.4%	−3.2 pp
Precision	85.1% ± 1.5%	82.7% ± 1.6%	−2.4 pp
Recall	86.4% ± 1.3%	82.2% ± 1.5%	−4.2 pp
F1-Score	85.7% ± 1.2%	82.5% ± 1.4%	−3.2 pp

Note: Performance differences are consistent across all five folds. Full fold-level results are available in Appendix A.

The RTAIL model records a mean accuracy of 84.1% (±1.4%), compared with 87.3% (±1.2%) for the baseline — a reduction of 3.2 percentage points. The corresponding F1-score reduction is 3.2 pp (85.7% to 82.5%). Critically, McNemar's test on the aggregated cross-validation predictions yields $\chi^2 = 2.89$, $p = 0.09$, indicating that this performance difference is not statistically significant at the conventional $\alpha = 0.05$ threshold. The performance reduction therefore lies within the operationally acceptable bound of 5 percentage points specified in RQ1, and cannot be reliably distinguished from sampling variation.

6.2 Fairness Evaluation (RQ1)

Table 2. Fairness metrics across five-fold cross-validation (mean ± 95% CI). Paired t-test on DPD: $t(4) = 9.73$, $p < 0.001$. Paired t-test on EOD: $t(4) = 11.20$, $p < 0.001$.

Metric	Baseline Model	RTAIL Model	Reduction
Demographic Parity Difference	0.24 ± 0.03	0.08 ± 0.01	66.7%
Equal Opportunity Difference	0.19 ± 0.02	0.06 ± 0.01	68.4%

Note: Both fairness improvements are statistically significant at $p < 0.001$ under paired t-test across cross-validation folds.

The RTAIL model achieves a 66.7% reduction in DPD (from 0.24 ± 0.03 to 0.08 ± 0.01) and a 68.4% reduction in EOD (from 0.19 ± 0.02 to 0.06 ± 0.01). Both improvements are highly statistically significant ($p < 0.001$). The RTAIL model meets the target thresholds of DPD ≤ 0.10 and EOD ≤ 0.10 in all five cross-validation folds, demonstrating consistent and reliable bias mitigation. These results provide strong affirmative evidence for RQ1: RTAIL produces a significant, reproducible reduction in group-level bias at an acceptable performance cost.

6.3 Explainability Analysis (RQ2)

Table 3. Explainability metrics: SHAP-based feature attribution consistency across five-fold cross-validation (mean ± 95% CI). Higher Spearman ρ and consistency scores indicate more stable explanations.

Metric	Baseline	RTAIL
Top-5 Feature Agreement (ρ)	0.61 ± 0.08	0.89 ± 0.04
SHAP Consistency Score	0.54 ± 0.09	0.91 ± 0.03
Mean Explanation Depth	4.8 ± 0.6	3.1 ± 0.3

Note: Mean explanation depth reflects the number of features accounting for 90% of cumulative SHAP magnitude. Lower depth indicates more concentrated, interpretable explanations.

The RTAIL model exhibits markedly superior explanation consistency across all three metrics. The SHAP Consistency Score improves from 0.54 ± 0.09 (baseline) to 0.91 ± 0.03 (RTAIL), and the Spearman rank correlation of top-5 features improves from 0.61 to 0.89. Both exceed the RTAIL consistency threshold of 0.80. The mean explanation depth decreases from 4.8 to 3.1, indicating that RTAIL models concentrate predictive signal in fewer, more prominent features — a property that substantially facilitates adverse action notice production in the credit context. These results confirm RQ2: RTAIL generates significantly more consistent and interpretable explanations.

6.4 Ablation Study

Table 4. Ablation study results across RTAIL component variants. DPD = Demographic Parity Difference; EOD = Equal Opportunity Difference. All values are means across five-fold cross-validation.

Model Variant	Accuracy	F1	DPD	EOD
Baseline	87.3%	85.7%	0.24	0.19
+ Fairness constraint only	85.0%	83.5%	0.10	0.09
+ Explainability only	86.8%	85.1%	0.23	0.18
+ Robustness only	86.5%	84.9%	0.22	0.17
Full RTAIL (all components)	84.1%	82.5%	0.08	0.06

Note: The fairness constraint (reweighting) is the dominant driver of bias reduction. Explainability and robustness components contribute marginally to fairness but substantially to explanation consistency and model stability respectively.

The ablation study reveals a clear hierarchy of component contributions. The fairness constraint alone accounts for the majority of the observed bias reduction, reducing DPD from 0.24 to 0.10 and EOD from 0.19 to 0.09. The explainability-only variant produces negligible fairness improvement (DPD = 0.23), confirming that SHAP enforcement alone does not constitute a fairness intervention. The robustness-only variant similarly shows minimal fairness impact. The full RTAIL configuration, combining all three components, achieves the most favourable fairness outcomes (DPD = 0.08, EOD = 0.06), suggesting a modest synergistic interaction between the reweighting and regularisation components.

7. Discussion

7.1 Interpreting the Accuracy-Fairness Trade-off

The statistically non-significant 3.2 pp accuracy reduction observed in this study is consistent with findings reported by Corbett-Davies et al. (2017), who demonstrate that the cost of fairness constraints in binary classification is bounded by the degree of correlation between sensitive attributes and legitimate predictive features. In the loan approval context, where gender and age are correlated with credit score and income in the data-generating process, any fairness constraint that reduces dependence on sensitive attributes must partially sacrifice predictive power. The empirical question is whether the cost is acceptable; the present results suggest it is, both statistically ($p = 0.09$) and operationally (below the 5 pp threshold).

These findings also speak to a broader theoretical debate in the fairness literature. The impossibility results of Chouldechova (2017) and Kleinberg et al. (2016) establish that demographic parity, equal opportunity, and calibration cannot simultaneously be satisfied under general conditions. The RTAIL framework does not resolve this impossibility but operationalises it: by specifying, in the Ethical Scope Document, which fairness criteria are prioritised for a given regulatory context, RTAIL enables principled trade-off decisions rather than ad hoc compromises.

7.2 Regulatory Implications

For practitioners deploying AI in consumer credit, the results have direct regulatory relevance. The Equal Credit Opportunity Act (ECOA) and its implementing Regulation B require that adverse action notices identify the specific reasons for a credit denial. The RTAIL model's improved explanation depth (3.1 vs. 4.8 features) and SHAP consistency (0.91 vs. 0.54) substantially simplify the production of regulatory-compliant adverse action notices, as the dominant decision factors are more stable and fewer in number. Similarly, the RTAIL audit trail — comprising the Ethical Scope Document, Model Card, monitoring logs, and accountability scores — provides the documentation artefacts required under Article 11 of the EU AI Act for high-risk AI systems.

7.3 Trustworthiness as a Measurable Property

A central theoretical contribution of this work is the demonstration that trustworthiness is not merely a normative aspiration but a measurable, improvable property of AI systems. By operationalising trustworthiness across three quantifiable dimensions — fairness (DPD, EOD), explainability (SHAP

consistency, explanation depth), and robustness (cross-fold stability) — RTAIL enables organisations to set specific, auditable trustworthiness targets and to track progress toward them over time. This aligns with the NIST (2023) AI RMF's emphasis on measurement as a prerequisite for risk management and with Vaghela's (2025) accountability scoring methodology.

8. Limitations and Future Work

Several limitations of the present study should be noted, along with directions for future research that would address them.

First, the experimental validation uses synthetic data, generated to approximate real-world credit portfolio characteristics but not derived from an institutional dataset. While this approach affords full data governance control and eliminates privacy risks, it limits the generalisability of specific numerical findings (DPD reductions, consistency scores) to real deployments. Future work should replicate the evaluation on publicly available real-world datasets — the UCI German Credit dataset, the HMDA mortgage disclosure data, and the COMPAS recidivism dataset — to establish empirical bounds on RTAIL's performance across data-generating processes.

Second, the current study evaluates a single classifier family (logistic regression). While logistic regression is a principled and widely used baseline, RTAIL's component-based design is intended to be classifier-agnostic. Future work should evaluate RTAIL with gradient-boosted tree models (XGBoost, LightGBM) and neural network classifiers, where the interaction between fairness constraints and model complexity may produce different trade-off profiles.

Third, the study considers a binary classification problem with two sensitive attributes. Regulated deployments frequently involve multiple sensitive attributes with complex intersectional dependencies (e.g., race × gender × age). Extending RTAIL's fairness evaluation to intersectional fairness criteria (Kearns et al., 2018) is an important next step.

Fourth, the evaluation is cross-sectional and does not assess RTAIL's Phase 4 (continuous monitoring) over time. A longitudinal evaluation examining how model fairness and explanation consistency drift as data distributions shift — and how RTAIL's monitoring alerts respond — would substantially strengthen the governance case for the framework.

9. Conclusion

This paper has presented and empirically evaluated the Regulated Trustworthy AI Lifecycle (RTAIL), an integrated framework for building AI systems that are simultaneously fair, explainable, robust, and governance-ready for regulated deployment contexts. Across five-fold stratified cross-validation on a synthetic loan approval dataset, RTAIL reduced the Demographic Parity Difference by 66.7% and the Equal Opportunity Difference by 68.4%, at a statistically non-significant accuracy cost of 3.2 percentage points. SHAP-based explanation consistency improved from 0.54 to 0.91, and ablation analysis confirmed that the fairness constraint is the primary driver of bias reduction.

These results demonstrate that trustworthiness is an achievable, engineerable property of AI systems — measurable, improvable, and auditable — rather than an inherently aspirational ideal. The accuracy-fairness trade-off, while real, is navigable within operationally acceptable bounds when fairness constraints are designed into the model development process rather than applied post hoc. RTAIL provides the governance architecture for making those design choices principled, transparent, and regulatory-compliant.

As AI systems assume greater roles in decisions with consequential societal impact, frameworks that enable responsible deployment without sacrificing operational effectiveness become not merely desirable but legally necessary. RTAIL represents a concrete step toward that standard.

Ethics Statement

This study uses exclusively synthetic data and involves no human participants, personal data, or real institutional records. No ethical review board approval is required. The synthetic data-generating process was designed to reflect empirically documented patterns of demographic disparity in consumer credit, not to target or characterise any specific individual or institution. The authors acknowledge that fairness interventions carry their own distributional consequences and that the choice of fairness criterion (demographic parity vs. equal opportunity) involves normative judgments that must be made by domain experts and regulators in partnership with affected communities, not by model developers alone.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank the reviewers for their constructive feedback, which substantially improved the rigour and clarity of this manuscript.

References

1. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.
2. Burla, S., & Priya, S. S. (2025). Trustworthy and explainable AI: Bridging algorithms and ethics. Proceedings of ICAAIC 2025, 1571–1577. https://doi.org/10.1109/icaaic64647.2025.11330323
3. Calegari, R., Giannotti, F., Milano, M., & Pratesi, F. (2025). Introduction to Special Issue on Trustworthy Artificial Intelligence (Part II). ACM Computing Surveys, 57(6), 1–3. https://doi.org/10.1145/3711127
4. Calegari, R., Giannotti, F., Pratesi, F., & Milano, M. (2024). Introduction to Special Issue on Trustworthy Artificial Intelligence. ACM Computing Surveys, 56(7), 1–3.

https://doi.org/10.1145/3649452

5. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047
6. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 797–806. https://doi.org/10.1145/3097983.3098095
7. Díaz-Rodríguez, N., Ser, J. D., Coeckelbergh, M., Prado, M. L. de, Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion, 99, 101896. https://doi.org/10.1016/j.inffus.2023.101896
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226. https://doi.org/10.1145/2090236.2090255
9. European Parliament. (2024). 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.
10. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323.
11. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8
12. Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. Proceedings of the 35th International Conference on Machine Learning (PMLR 80), 2564–2572.
13. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference. https://doi.org/10.4230/LIPIcs.ITCS.2017.43
14. Kumar, P., & Kumar, S. (in press). Designing trustworthy artificial intelligence. In Advances in Computational Intelligence and Robotics. IGI Global. https://doi.org/10.4018/979-8-3373-8187-9.ch001
15. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2022). Trustworthy AI: From principles to practices. ACM Computing Surveys, 55(9), 1–46. https://doi.org/10.1145/3555803
16. Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A. K., & Tang, J. (2022). Trustworthy AI: A computational perspective. ACM Transactions on Intelligent Systems and Technology, 14(1), 1–59. https://doi.org/10.1145/3546872
17. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596
19. Nastoska, A., Jancheska, B., Rizinski, M., & Trajanov, D. (2025). Evaluating trustworthiness in AI: Risks, metrics, and applications across industries. Electronics, 14(13), 2717. https://doi.org/10.3390/electronics14132717
20. National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
21. Papademas, M., Ziouvelou, X., Troumpoukis, A., & Karkaletsis, V. (2025). Bridging ethical principles and algorithmic methods: An alternative approach for assessing trustworthiness in AI systems. Frontiers in Computer Science, 7. https://doi.org/10.3389/fcomp.2025.1658128
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
23. Vaghela, R. (2025). FRICE: Enhancing AI trust through a robust accountability system. Journal of Information Systems Engineering & Management, 10, 514–522. https://doi.org/10.52783/jisem.v10i8s.1099
24. Vyhmeister, E., & Castañé, G. G. (2024). TAI-PRM: Trustworthy AI — project risk management framework towards Industry 5.0. AI and Ethics. https://doi.org/10.1007/s43681-023-00417-y
25. Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. Proceedings of the 26th International Conference on World Wide Web, 1171–1180.
26. Zen, R., & Ingold, I. (2025a). Quantifying trustworthiness: A holistic metrology for responsible AI systems [Version 1]. Zenodo. https://doi.org/10.5281/zenodo.17816738
27. Zen, R., & Ingold, I. (2025b). Quantifying trustworthiness: A holistic metrology for responsible AI systems [Version 2]. Zenodo. https://doi.org/10.5281/zenodo.17816739