information

MDPI

Article

Operationalizing Accountable AI Through Traceable Governance Architecture for Institutional Decision Support

Abdalilah Alhalangy

Department of Computer Engineering, College of Computer, Qassim University, Buraydah 52361, Saudi Arabia; a.alhalangy@qu.edu.sa

Abstract

Institutional artificial intelligence (AI) decision-support systems progressively evaluate cases, determine eligibility, and allocate resources; yet, predicted efficacy alone does not guarantee equity, contestability, or responsible utilization. Current research frequently considers fairness measures, explainability, human oversight, and organizational governance as rather distinct issues. This paper presents a traceable bias-auditing framework that amalgamates prediction, explanation, selective human review, and structured recording into a cohesive operational decision pathway. Through design science research, the artifact was exhibited in a controlled proof-of-concept utilizing 8000 synthetic institutional situations and historically biased data labels. The foundational classifier was a logistic regression model. Selective escalation is initiated by the proximity of boundaries, tension in explanation patterns, and the rules governing review priorities. Three situations were evaluated: baseline prediction, prediction with explanation alone, and comprehensive architecture with review and audit recording. Explanations enhanced reviewability but did not significantly alter fairness outcomes. The proposed architecture improved F1 from 0.781 to 0.795, reduced the demographic parity gap from 0.070 to 0.010, decreased the equal opportunity gap from 0.116 to 0.036, and improved audit completeness from 0.33 to 1.00, while escalating only 4.8% of cases for human review. The results indicate that explanations attain institutional significance solely when linked to procedural regulations and enduring records. The evaluation was simulation-based; thus, the results should be interpreted as proof-of-concept evidence rather than direct field validation.

Keywords: bias auditing; artificial intelligence; explainable AI; institutional decision support; interpretability; accountable AI; traceability

Academic Editors: Marco Guazzone, Manuel Striani and Stefania Montani

Received: 23 June 2026
Revised: 15 July 2026
Accepted: 15 July 2026
Published: 16 July 2026

1. Introduction

Artificial intelligence (AI) now supports decisions that were formerly managed through manual triage or rule-based administrative evaluations. Public agencies, universities, hospitals, financial service providers, and major corporations use machine learning to prioritize cases, assess eligibility, identify abnormalities, and suggest actions. The appeal is evident: AI can analyze extensive case volumes, reveal concealed trends, and enhance reaction times. Recent human-centered artificial intelligence research further emphasizes that AI-based decision-support systems should not be designed as autonomous prediction tools alone, but as socio-technical systems that integrate human judgment, interaction design, data management, and decision logic across the decision pathway [1].

Recent governance research indicates that AI can enhance service personalization, fraud detection, and evidence-based administration when implemented with discipline and contextual awareness [2,3].

The challenges of governance are distinctly evident. Once AI is integrated into an institutional decision-making process, the evaluation of the decision transcends mere predicted accuracy. The assessment is based on the impacted individuals' comprehension, the staff's ability to challenge it, the reviewers' capacity to reconstruct it, and the leadership's demonstration of responsible system usage. The National Institute of Standards and Technology (NIST) presents this as an issue of trustworthiness and risk management, while ISO/IEC 42001 [4] frames artificial intelligence management systems in terms of governance, transparency, accountability, and continual improvement. NIST's generative AI profile further reinforces this rationale by emphasizing documentation, monitoring, and explainability in the management of operational risk [4–6].

The practical issue is that the decision-making process hardly ever incorporates these governance concepts in an operational form. Even while many institutions claim to have a model card, policy document, or fairness dashboard, they nevertheless find it difficult to respond to straightforward audit inquiries at the case level, including the reasons for a specific case's rejection. Which elements were important? Was the decision boundary close to the model? Was there a human review initiated? Where is the rationale for overturning the decision? Research on data governance has previously demonstrated that model architecture and organizational procedures are critical to reliable AI [7]. A grant-eligibility screening or benefit-priority scheme may seem compliant at the policy level, yet remain ambiguous in the context of any contested case.

A key issue is the conceptual fragmentation within the literature. Research on bias and fairness has generated robust classifications of harm sources and mitigation strategies; yet, a significant portion of this work is still centered around models. Recent evaluations indicate that bias may arise from historical data, proxy variables, optimization decisions, interface design, and post-deployment feedback mechanisms. However, the mechanisms that connect these sources of bias to routine institutional review practices remain insufficiently specified [8,9].

Investigations into corporate governance are progressing in this approach. Research on algorithm review boards indicates that institutions are progressively establishing internal frameworks to assess AI risks; nevertheless, these structures often function at the project or portfolio level rather than at individual decision-making events. Similarly, research on public-sector explainability indicates that explanations might enhance perceived fairness or trust in certain contexts; however, the impact is contingent upon the sort of explanation and the institutional environment. Consequently, mere explanations do not address accountability [10].

The missing bridge, this study contends, is attributable to bias auditing at the decision-making level. The point is not just to see whether a model is biased in the aggregate. The intention is to make fairness, explainability and human oversight a controlled sequence of events that can be rebuilt after each decision. An auditing architecture should thus be traceable, and should also answer four questions simultaneously: what the model decided, why this scenario was presented to the model in this way, when a human must intervene, and how the final decision was recorded for future review.

This work meets this need by presenting a bias auditing framework for institutional decision support systems and testing the framework with a simulation-based proof of concept. The design was deliberately functional. Rather than relying on ongoing human supervision, the approach leverages selective escalation, driven by explanation signals and decision uncertainty. It has a systematic audit trail (with the right fields) rather than abstract

transparency. This enables the framework to test a concrete proposition: explanations only become institutionally significant when they are linked to review rules and traceable logs.

Four research questions led this investigation. First, how can we combine fairness, explainability, human review and auditability in one institutional decision support architecture? Second, what changes when an explanation layer is added without modifying the review responsibilities? Third, can a selective human review strategy decrease fairness gaps without appreciably compromising predictive utility? Fourth, which operational indicators are most relevant for evaluating accountable decision support in practice?

The remainder of this paper is organized as follows. Section 2 reviews the literature and identifies the research gap. Section 3 describes the design science process, the proposed artifact, the simulation environment, and the evaluation criteria. Section 4 presents the results. Section 5 discusses the findings in relation to governance and institutional practice. Section 6 presents the conclusions, practical and theoretical implications, limitations, and future work.

This study contributes to the operationalization of trustworthy AI by proposing a case-level governance architecture that transforms explainability from an informational feature into an enforceable institutional control mechanism.

2. Literature Review and Research Gap

Institutional AI governance has grown rapidly, yet its application varies. A new synthesis by the OECD suggests that governments are increasingly using AI in critical areas such as service delivery, fraud detection and administrative help, while at the same time grappling with issues of accountability, transparency and lack of competence. Further research in Government Information Quarterly shows that public firms do not implement AI governance as a single, mature framework. Implementation is sometimes incomplete and fragmented, but also strongly shaped by internal capabilities, existing routines and governance culture [2,3].

This distinction is important because institutional AI governance can fail not only at the level of policy ambition, but also at the level of operational execution. A policy may mandate fairness, yet the institution may still be unable to demonstrate how fairness was addressed in a specific case without a case-level trace. Similarly, human oversight may be formally required, but it becomes symbolic rather than effective when escalation criteria are unclear. This argument is well summarized in data governance scholarship that illustrates how organizations define ownership, quality, stewardship, accountability, and reuse at each stage in the data lifecycle to generate trustworthy AI [7].

Research on bias and fairness highlights the need for governance mechanisms that are operational rather than merely declaratory. Algorithmic bias can emerge upstream through sampling imbalance, label construction, and proxy variables, but it can also arise downstream through optimization objectives, threshold settings, interface design, and post-deployment feedback loops. This means that fairness is not a static property of a model alone; rather, it is a sociotechnical condition that may deteriorate even when overall accuracy appears acceptable [8,9].

This is a pivotal moment for institutional decision-making. Administrative systems are frequently refined for uniformity and efficiency. Nevertheless, the same system may subtly perpetuate historical inequities if it assimilates biased historical judgments or factors that indirectly reflect differences among protected groups. The overall performance may remain elevated due to the predominance of the majority group's behavior influencing the training signal. In such instances, a model may seem efficient while generating consistently disparate acceptance rates, mistake rates, or access outcomes.

Research on explainability offers a partial but incomplete response to this problem. Studies of public-sector decision-making show that the type of explanation matters: input-based, group-based, case-based, and counterfactual explanations do not have uniform effects on perceived fairness, trust, or decision acceptance. Related work has also shown that fairness information and explanatory cues shape how individuals evaluate AI-assisted decisions. These findings are important because they show that explanations are not uniform interventions; however, they also indicate that explanation alone cannot substitute for governance [11,12].

A secondary body of literature concentrates on explainable artificial intelligence (XAI) and transparent AI in the context of policymaking and public administration. This study demonstrates that institutions require architectures in which openness is integral to the tools, rather than an afterthought. Consequently, policy-oriented AI systems must have interpretable models, evidence trails, and operational compliance mechanisms. Nonetheless, the majority of this work is confined to architectural or policy frameworks, rather than addressing case-specific institutional determinations that may subsequently be challenged by impacted persons [13].

The literature on human oversight adds a further caution. Governance frameworks frequently recommend human-in-the-loop or human-on-the-loop configurations, but empirical and legal scholarship warns that such requirements can become symbolic if they are not supported by clear procedures. Green [14] argues that human oversight policies may backfire when they assume that human reviewers can automatically correct algorithmic harm. Oversight should therefore be treated not as a nominal safeguard, but as an institutionalized procedure with justified triggers, evidence-based review criteria, and organizational accountability.

Recent work in public-sector governance supports this position. From a procedural justice perspective, institutions often face value conflicts involving fairness, transparency, efficiency, legality, and consistency. Such conflicts cannot always be resolved through a single optimization criterion; they require explicit procedures for deliberation, justification, and resolution. Similarly, research on algorithm review boards suggests that responsible AI governance is more effective when review practices are embedded within existing organizational procedures and supported by leadership, documentation, and clear success criteria [10,15].

From a classical accountability perspective, the framework aligns with Bovens' conception of accountability as a relationship involving information provision, justification, and the possibility of consequences. The proposed architecture operationalizes these elements at the case level: explanation provides information, selective review enables justification and contestation, and structured logging preserves the institutional record necessary for retrospective scrutiny.

Taken together, the literature supports three important observations. First, fairness problems are not merely technical or isolated; they are sociotechnical and continuous. Second, explanations can enhance understanding and perceived legitimacy, but their effectiveness depends on context, design, and institutional use. Third, human oversight is meaningful only when it is institutionalized through explicit review logic and supporting evidence. An implementable mechanism that combines these three insights within a single operational decision pathway remains underdeveloped.

Classical socio-technical and documentation work further sharpens this point. Selbst et al. showed that fairness failures often emerge when technical abstractions detach models from the institutional settings in which they operate. Model cards, datasheets, and algorithmic impact assessments improve the documentation of models, datasets, and antic-

ipated risks, but they do not define how a contested case should be escalated, justified, and reconstructed after deployment [16–19].

Table 1 compares the proposed framework with established governance artifacts and shows why case-level traceability and explicit review logic remain insufficiently specified in existing documentation-oriented literature.

Table 1. Comparison of related governance artifacts and the proposed framework.

Artifact/Stream	Primary Emphasis	Analytical Level	Case-Level Traceability	Explicit Review Logic
Model cards	Model reporting and intended use	Model	No	No
Datasheets for datasets	Dataset provenance and documentation	Data	No	No
Algorithmic impact assessments	Ex ante public-agency accountability and risk review	Organizational	Partial, but not case reconstructability	No
Fairness mitigation studies	Bias detection or correction in the learning pipeline	Model/pipeline	Rarely	Rarely
Proposed framework	Integrated decision-path accountability	Case + workflow	Yes	Yes

Therefore, the research gap does not lie in the absence of fairness metrics, explainability methods, or governance principles. Rather, it lies in the lack of a consistent case-level auditing architecture that integrates these components into a single operational pathway. Existing studies provide limited specification of how a decision-support system should move from raw case input to model output, from model output to explanation, from explanation to justified escalation, and from escalation to a verifiable final record. The present study addresses this gap by designing and evaluating a framework that treats accountability as a prospective decision pathway rather than as a retrospective reporting exercise.

The current study makes three contributions. First, it introduces a design artifact that integrates prediction, explanation, selective human review, and audit recording within a single accountable decision-support architecture. Second, it defines operational evaluation measures that connect predictive utility, fairness gaps, escalation burden, override precision, and audit completeness. Third, it demonstrates through simulation that explanations alone do not substantially improve fairness outcomes; rather, explanations become institutionally meaningful when they are linked to selective review and systematic documentation.

Hypotheses Development

Building on the theoretical integration of fairness governance and institutional accountability, the study formulates the following hypotheses:

H1. *Adding an explanation layer without modifying review responsibilities will not significantly alter predictive utility or fairness metrics.*

H2. *The full traceable auditing architecture will significantly reduce demographic parity and equal opportunity gaps compared to the baseline model.*

H3. *The full architecture will significantly increase audit completeness relative to explanation-only and baseline conditions.*

H4. *Selective review will not significantly reduce overall predictive accuracy.*

These hypotheses translate the conceptual proposition of the framework into empirically testable claims.

3. Methodology

This study follows a design science research approach because its purpose is to construct and evaluate an artifact for a clearly defined organizational problem rather than to explain a naturally occurring phenomenon. In information systems research, design science is appropriate when the objective is to develop a purposeful artifact and assess its utility in relation to a practical problem environment. The methodological framework is therefore grounded in the design science principles of Hevner et al. and the process model proposed by Peffers et al. [20,21].

The design process followed five stages. First, the problem was identified as an institutional failure to translate monitoring, explainability, and fairness principles into accountable case-level decisions. Second, the design objectives were specified: the artifact should preserve predictive utility, explain decision rationales, selectively trigger human review, and collect sufficient metadata to reconstruct the final judgment. Third, the traceable governance architecture was developed. Fourth, the artifact was demonstrated in a controlled simulation setting. Fifth, the artifact was evaluated using predictive utility, fairness, and auditability indicators.

The proposed artifact is a Traceable Bias Auditing Framework composed of four interrelated layers. The prediction layer generates a probability score and an initial decision recommendation. The explanation layer translates the model output into case-level feature contributions. The human review layer applies selective escalation so that only eligible borderline cases are routed to reviewers. The audit trail layer preserves a structured record of the complete decision pathway. These layers and their minimum recorded outputs are summarized in Table 2.

Table 2. Core layers of the Traceable Bias Auditing Framework.

Component	Primary Function	Minimum Recorded Output
Prediction layer	Generate risk or eligibility score and preliminary decision	Probability score, thresholded recommendation, model version
Explanation layer	Expose case-level rationale for the recommendation	Top positive and negative factors, signed local contributions
Human review layer	Escalate only selected cases for institutional intervention	Review trigger, override/retain action, reviewer rationale
Audit trail layer	Preserve decision trace for later scrutiny and learning	Case ID, timestamp, score, factors, reviewer action, justification

Table 2 summarizes the four layers of the proposed artifact and the minimum information each layer must expose for accountable institutional use.

Figure 1 illustrates the proposed traceable governance architecture. The process starts with institutional case intake, moves through prediction, explanation, selective escalation, and human review, and ends with final decision recording and audit logging. The audit trail is designed as a continuous record connected to both the explanatory and evaluative components of the model. This design choice is important because recording only the final score would not be sufficient for accountability; it would obscure the pathway from model output to institutional action.

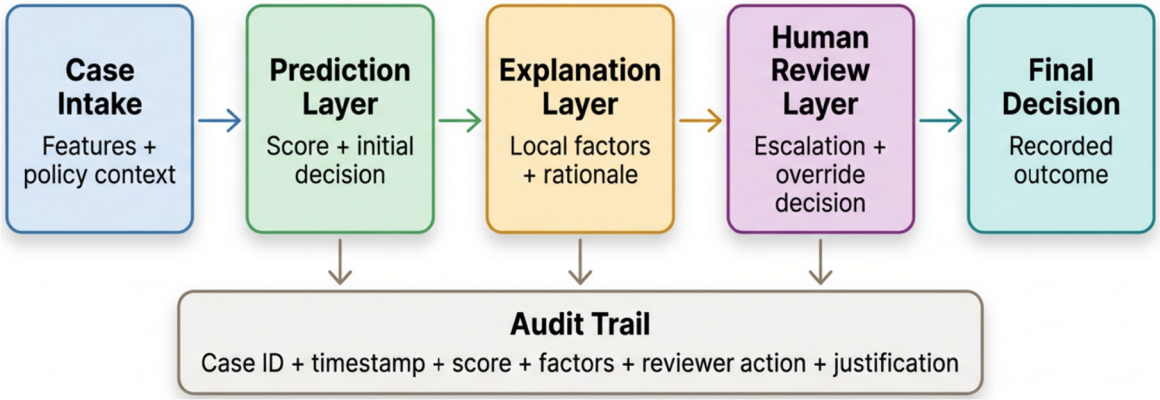

Figure 1. Proposed traceable governance architecture for accountable AI-enabled institutional decision support.

The proposed architecture also emphasizes that model predictions are not, by themselves, responsible for institutional decisions. Accountability becomes possible only when the institution can explain how the recommendation was interpreted, whether it triggered a review, how the final decision was justified, and what evidence was preserved for later reconstruction.

To demonstrate the artifact, we created a synthetic institutional dataset with 8000 samples. The simulated process represents a conventional eligibility or priority-screening setting in which decisions are influenced by six operational attributes: income adequacy, historical compliance, case complexity, document quality, household load, and regional disadvantage. A protected-group indicator was used only for fairness evaluation and policy-review simulation; it was not included as an input to the prediction model. This design choice reflects common institutional practice, where protected variables are often excluded from model inputs even though proxy effects may remain. The audit-log field schema and synthetic variable dictionary are provided in File S2.

Two labeling schemes were defined. The first represented underlying case deservingness and served as the reference evaluation criterion. The second represented historically biased institutional decisions, including a systematic penalty affecting the protected minority and an interaction with regional disadvantage. The baseline model was trained on the historically biased labels. This made the proof of concept more realistic because many institutions learn from prior administrative decisions rather than from normatively ideal outcomes.

A logistic regression classifier was used as the baseline predictive model because the purpose of the study was to demonstrate governance integration rather than to introduce a new modeling technique. This choice also supported interpretability and consistent local decomposition of case-level feature contributions. The dataset was divided into 70% training data and 30% test data, and the decision threshold was set at 0.50. In the explanation layer, feature contributions were estimated for each case using the fitted model coefficients and the observed feature values. These contributions were then ranked to identify the most influential positive and negative factors shaping the case-level recommendation.

The human review policy was intentionally selective. Escalation was considered only for initially negative cases when three conditions were simultaneously satisfied: the model score was close to the decision boundary, the explanation pattern showed policy-relevant tension, and the case met a review-priority rule. Review-priority criteria included protected-group membership and the presence of strong documentation or compliance evidence that conflicted with the initial negative recommendation. The purpose was not to reproduce the judgment of a real administrator perfectly. Rather, the purpose was to evaluate whether

an explanation-based and rule-governed intervention could improve outcomes without requiring manual review of every case.

Formally, for each case i, the escalation decision E_i was triggered only when four conditions were jointly satisfied: the initial prediction was negative, the prediction score was close to the decision boundary, the explanation pattern showed policy-relevant tension, and the case satisfied a review-priority condition. This rule is expressed as:

$$E_i = 1 \Leftrightarrow (\hat{y}_i = 0) \wedge (|p_i - 0.50| < \delta) \wedge \left(S_i^+ - \lambda S_i^- > 0\right) \wedge (A_i = 1 \vee u_i > 0.80) \quad (1)$$

The positive and negative explanation contributions were defined as:

$$S_i^+ = \Sigma_j \max(c_{ij}, 0), S_i^- = \Sigma_j \max(-c_{ij}, 0) \quad (2)$$

where p_i denotes the predicted probability for case i, $\hat{y}_i$ denotes the initial binary prediction, c_{ij} denotes the local contribution of feature j to case i, A_i denotes protected-group membership, u_i denotes urgency, δ is the boundary-proximity threshold, and λ is the explanation-tension weight. In the main specification, $\delta = 0.20$ and $\lambda = 0.90$.

Reviewer override followed a second rule. For each escalated case, the reviewer score r_i was computed as:

$$r_i = 1.00d_i + 0.80s_i + 0.65q_i - 0.90m_i + 0.45u_i - 0.20x_i \quad (3)$$

where d_i denotes documentation quality, s_i denotes consistency, q_i denotes prior compliance, m_i denotes missing documents, u_i denotes urgency, and x_i denotes case complexity. A reviewer override was applied when $r_i > 0.10$. Algorithm 1 summarizes the selective escalation and audit logging procedure used in the proposed architecture. The extended implementation note for the escalation logic is provided in File S1.

Algorithm 1. Selective escalation and audit logging procedure

Input: Case features X_i, predicted probability p_i, initial prediction $\hat{y}_i$, local explanation contributions c_{ij}, protected-group indicator A_i, urgency score u_i, and audit-log schema.
Output: Final decision D_i, escalation decision E_i, reviewer action R_i, and completed audit record L_i.

1. Generate the initial prediction score p_i using the trained classifier.
2. Assign the initial binary prediction $\hat{y}_i$ using the decision threshold of 0.50.
3. Compute local explanation contributions c_{ij} for all features j.
4. Compute positive and negative explanation scores S_i^+ and S_i^-.
5. Set $E_i = 0$ by default.
6. If $\hat{y}_i = 0, |p_i - 0.50| < \delta, S_i^+ - \lambda S_i^- > 0,$ and $(A_i = 1$ or $u_i > 0.80)$, then:
6.1 Set $E_i = 1$.
6.2 Route the case to human review.
6.3 Compute the reviewer score r_i.
6.4 If $r_i > 0.10$, override the initial decision; otherwise, retain it.
7. If $E_i = 0$, retain the initial model decision.
8. Record the case identifier, timestamp, model score, top explanatory factors, escalation status, reviewer action, final decision, and justification in the audit trail.
9. Return the final decision and completed audit record.

The use of protected-group membership in the review rule requires careful governance safeguards. In this design, the protected-group indicator was not used to train the classifier, adjust model scores, or assign a lower predicted status to any case. It was

used only as a post-decision safety mechanism within a selective review strategy, with the aim of identifying potentially biased borderline negative recommendations before a final institutional decision was recorded. In practice, such protection-oriented escalation should be subject to legal review, explicit policy authorization, and periodic auditing to ensure that the corrective mechanism does not become a hidden source of unequal treatment.

The components of the selective review rule serve different governance functions and should not be interpreted as interchangeable predictors. Boundary proximity identifies cases in which the model recommendation is uncertain. Explanation-pattern tension identifies cases in which the negative recommendation conflicts with positive case-level evidence. Review-priority conditions identify cases in which the institution has a stronger procedural reason to examine the recommendation before finalization. Protected-group membership was used only within this post-decision safeguard and only under the assumption that such use is legally authorized and institutionally documented. The present proof-of-concept evaluates the combined governance rule rather than estimating the independent causal contribution of each component; therefore, component-level ablation is treated as a sensitivity and future-validation issue rather than as a claim of causal attribution.

To assess the stability of the findings beyond a single threshold specification, the evaluation included two robustness checks. First, a sensitivity analysis varied the boundary-proximity threshold δ across three values: 0.15, 0.20, and 0.25. Second, percentile bootstrap confidence intervals were estimated for the differences between the baseline and full-framework conditions using 2000 resamples of the test set. These checks do not remove the limitations of synthetic data, but they strengthen the internal credibility of the proof-of-concept evaluation. Extended robustness tables are provided in File S1.

These robustness checks should be interpreted as internal stability checks within the proposed simulation setting rather than as evidence of universal parameter robustness. The main analysis varied the boundary-proximity threshold because this parameter directly controls the number of cases eligible for review. Other parameters, including the explanation-tension weight λ and the reviewer override threshold, are institution-specific calibration choices and should be re-estimated or stress-tested before deployment in a real institutional environment. Future validation should therefore examine joint sensitivity to δ, λ, reviewer-score weights, and override thresholds using field data or domain-specific simulation settings.

The audit trail recorded six required fields for each reviewed case: case identifier, timestamp, model score, top explanatory factors, reviewer action, and justification. From these fields, the Audit Completeness Index (ACI) was defined as the number of required fields logged divided by the total number of required fields. For comparison across conditions, the baseline model logged only the case identifier and model score, yielding a nominal ACI of 0.33. The explanation-only condition additionally logged the main contributing factors and reached a score of 0.67. The full framework logged all six fields and reached 1.00.

Escalation rate is used as an operational governance metric rather than as a predictive-performance metric. It measures the proportion of cases routed from automated processing to human review. Formally, if $E_i = 1$ indicates that case i is escalated for human review and $E_i = 0$; otherwise, the escalation rate is defined as:

$$ER = \frac{1}{n}\sum_{i=1}^{n} E_i \tag{4}$$

where n denotes the total number of evaluated cases.

This metric is important because an accountable AI decision-support system must not only improve fairness and traceability, but also remain administratively feasible. Unlike

accuracy, F1 score, or AUC, which evaluate predictive utility, escalation rate estimates the review burden imposed on the institution. A low but effective escalation rate indicates that governance intervention is selective rather than universal, preserving throughput while enabling human review for high-risk borderline cases.

Three families of evaluation metrics were used in this study. Predictive utility was measured using accuracy, F1 score, and AUC. Fairness was measured using the demographic parity gap, equal opportunity gap, and false-positive-rate gap. Governance utility was evaluated using escalation rate, override precision, and the Audit Completeness Index. The operational meanings and formal definitions of these metrics are reported in Table 3.

Table 3. Evaluation metrics and their operational meanings.

Metric	Definition	Operational Meaning
Accuracy	Overall proportion of correct classifications	General predictive utility
F1 score	Harmonic mean of precision and recall	Balance under class imbalance
AUC	Area under the ROC curve	Threshold-independent discrimination
Demographic parity gap	$\left\lvert P\left(\hat{Y}=1 \middle\vert A=1\right)-P\left(\hat{Y}=1 \middle\vert A=0\right)\right\rvert$	Acceptance-rate disparity
Equal opportunity gap	$\left\lvert TPR_{\{A=1\}}-TPR_{\{A=0\}}\right\rvert$	True-positive-rate disparity
False-positive-rate gap	$\left\lvert FPR_{\{A=1\}}-FPR_{\{A=0\}}\right\rvert$	Error-rate disparity
Escalation rate	Share of cases routed to human review	Operational review burden
Override precision	Share of reviewed and changed cases that are correct under the reference labels	Review usefulness
Audit Completeness Index	Logged required fields/total required fields	Traceability and reconstructability

These indicators were selected because they correspond to the three accountability requirements addressed by the proposed framework. Predictive utility metrics assess whether the model remains operationally useful. Fairness metrics assess whether the intervention reduces group-level disparities in access and error patterns. Governance metrics assess whether the framework remains administratively feasible, whether human review is targeted effectively, and whether the final decision pathway can be reconstructed. Broader governance indicators, such as organizational maturity or policy compliance scores, were not used as primary metrics because the present study evaluates a case-level decision pathway rather than an organization-wide AI governance program.

The evaluation compared three experimental conditions. Condition 1 represented the baseline predictive model without explanation, selective review, or extended audit logging. Condition 2 added case-level explanations but did not change the final decisions, allowing the study to isolate the effect of explainability without human review. Condition 3 implemented the full proposed architecture, including explanation-guided selective review and complete audit logging. This design made it possible to test whether explanations become operationally valuable only when they are institutionally connected to review and traceability.

Although the simulation is simplified, it aligns with the core governance logic found in current standards and public-sector scholarship. The NIST and ISO/IEC emphasize measurable controls, documentation, and continual improvement, whereas public-sector studies stress that trustworthy AI requires transparency, evidence, and procedurally justified reviews. Therefore, the simulation serves as a proof-of-concept environment for an artifact intended to be deployed in real institutional settings [4–6,13,15].

The following equations define the predictive, fairness, and governance metrics used in the evaluation. Table 3 summarizes their operational meanings. Mathematically, let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy was defined as (TP + TN)/(TP + TN + FP + FN). Precision was defined as TP/(TP + FP), and recall was defined as TP/(TP + FN). The F1 score was calculated as 2 × Precision × Recall/(Precision + Recall). AUC was calculated as the area under the receiver operating characteristic curve. The demographic parity gap was defined as $|P(\hat{y} = 1 \mid A = 1) - P(\hat{y} = 1 \mid A = 0)|$, where A denotes protected-group membership. The equal opportunity gap was defined as $|TPR\_A = 1 - TPR\_A = 0|$, and the false-positive-rate gap was defined as $|FPR\_A = 1 - FPR\_A = 0|$. Escalation rate was defined as $ER = (1/n)\Sigma E\_i$, where $E\_i = 1$ if case i was escalated to human review and $E\_i = 0$ otherwise. Override precision was defined as the number of correct overrides divided by the total number of overrides. The Audit Completeness Index was defined as the number of logged required fields divided by the total number of required fields.

4. Results

Table 4 presents the quantitative results for the three experimental conditions. The baseline model achieved strong discriminatory performance, with an AUC of 0.913 and an F1 score of 0.781, but it also exhibited notable fairness gaps. The demographic parity gap was 0.070, and the equal opportunity gap was 0.116, indicating lower positive-decision and true-positive rates for the protected group.

Table 4. Comparative results across the three experimental conditions.

Condition	Accuracy	F1	AUC	DP Gap	EO Gap	FPR Gap	Escalation Rate	Override Precision	Audit Completeness
Baseline model	0.834	0.781	0.913	0.070	0.116	0.049	0.000	-	0.330
Model + explanation layer	0.834	0.781	0.913	0.070	0.116	0.049	0.000	-	0.670
Model + explanation + human review + audit trail	0.837	0.795	0.913	0.010	0.036	0.003	0.048	0.540	1.000

The explanation-only condition produced similar predictive and fairness outcomes to the baseline condition. This finding is analytically important because it indicates that explanations alone do not reduce outcome disparities or improve decision utility unless the institution defines when explanations should trigger action.

The full architecture improved both operational utility and fairness. Accuracy increased from 0.834 to 0.837, and the F1 score increased from 0.781 to 0.795. The AUC remained unchanged at 0.913, indicating that the underlying ranking performance was preserved. The observed gain therefore reflects targeted correction among borderline negative cases rather than a broad change in the ranking model. The fairness improvements were more pronounced: the demographic parity gap decreased from 0.070 to 0.010, the equal opportunity gap from 0.116 to 0.036, and the false-positive-rate gap from 0.049 to 0.003.

The escalation rate was 4.8%, indicating a modest operational review burden. This is important because the throughput benefits of algorithmic decision support would be undermined if every case required human review. Override precision was 0.540, indicating that more than half of the reviewed and overridden cases were correct under the reference labels. This value does not imply perfect correction; rather, it shows that the selective review rule concentrated corrective effort on high-risk borderline cases where model uncertainty

and explanation tension were most pronounced. Thus, the governance burden remained selective rather than universal.

The governance indicators showed a clearer contrast across conditions. The Audit Completeness Index increased from 0.33 in the baseline condition to 0.67 in the explanation-only condition and to 1.00 in the full framework. A baseline system may record the model score and final label, but these fields are insufficient to reconstruct the reasoning behind a disputed decision. By logging explanations, review triggers, reviewer actions, and justifications together, the full framework made the decision pathway reconstructable rather than merely stored.

These intervals suggest that the fairness improvements were directionally consistent and substantively meaningful within the simulated context. The reductions in demographic parity and equal opportunity gaps indicate that the review policy primarily affected borderline negative cases with offsetting positive evidence. This pattern is consistent with the intended design of the framework: it does not seek to optimize predictive performance broadly, but to improve fairness, reviewability, and accountability while preserving predictive utility.

Bootstrap estimates further supported the stability of these findings. Compared with the baseline condition, the full framework produced a mean F1 improvement of +0.014 (95% CI: 0.003 to 0.024), a mean reduction in the demographic parity gap of −0.052 (95% CI: −0.073 to −0.006), and a mean reduction in the equal opportunity gap of −0.077 (95% CI: −0.107 to −0.044). The change in accuracy was positive but small (+0.003, 95% CI: −0.005 to 0.011; McNemar $p = 0.52$), reinforcing the interpretation that the framework primarily improved fairness and reviewability rather than overall classification accuracy.

The larger reduction in the equal opportunity gap is consistent with the design of the selective review rule. The policy targeted borderline negative cases with strong positive evidence, allowing cases that were likely to be erroneous negative decisions under the reference labels to be restored through review. Demographic parity was also improved, but it is more sensitive to the overall acceptance rate, whereas equal opportunity directly reflects changes in true-positive rates among deserving cases.

A representative repaired case had a baseline score of 0.47, high documentation quality, prior compliance, one missing-document flag, and protected-group membership. The explanation layer identified a tension between the initial negative recommendation and the presence of substantial positive evidence. The case therefore satisfied the escalation criteria, and the reviewer score exceeded the override threshold. As a result, the final decision was changed to positive, and the justification was recorded in the audit trail. This type of borderline correction explains how the framework maintained a low review volume while still improving fairness outcomes.

Figure 2 illustrates the relationship between predictive utility and group disparity across the three experimental conditions. The baseline and explanation-only conditions overlap because explanations did not alter the final decisions. In contrast, the full framework achieved a lower demographic parity gap while preserving, and slightly improving, the F1 score. This pattern is important because it suggests that fairness-oriented governance does not necessarily require a substantial loss of predictive utility when review is targeted to high-risk borderline cases.

Table 5 further shows that the improvements were robust across realistic escalation-boundary values. Tightening the boundary reduced the review rate while preserving a substantial share of the fairness improvement, whereas relaxing the boundary produced only modest changes in review burden, F1 score, and override precision. This pattern indicates that the framework was not highly sensitive to a single tuning choice.

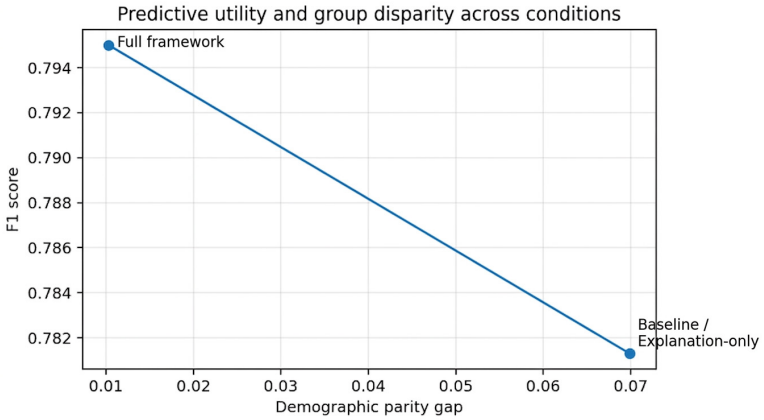

Figure 2. F1 score and demographic parity gap across experimental conditions. Baseline and explanation-only conditions share identical coordinates and are therefore shown with a combined label to avoid overlap.

Table 5. Sensitivity of the full framework to the escalation boundary parameter delta.

Delta	Review Rate	F1	DP Gap	EO Gap	Override Precision
0.15	0.041	0.793	0.019	0.048	0.538
0.20	0.048	0.795	0.010	0.036	0.540
0.25	0.048	0.795	0.009	0.036	0.534

Overall, the results support the central proposition of the study: explanations create visibility, review rules create targeted institutional intervention, and audit trails create accountability. The findings therefore show that explainability becomes institutionally meaningful when it is connected to selective human review and structured logging, rather than when it is treated as a standalone model feature.

5. Discussion

The findings have three important implications for research and practice. First, fairness governance should be understood as a decision-process design problem rather than as a matter of metric evaluation alone. The baseline model achieved a strong AUC, meaning that conventional machine-learning evaluation would likely have treated it as successful. However, group-level disparities remained substantial. The full framework improved outcomes not by replacing the model, but by reshaping the institutional process through which model outputs were interpreted, reviewed, and recorded.

Second, the findings show that explanation and accountability should not be conflated. This distinction is important because many institutional AI initiatives treat explainability as a proxy for responsible use. The results of this study suggest that this assumption is incomplete. Explanation without escalation logic remains informational rather than remedial: it helps an organization describe a decision, but it does not necessarily enable the organization to question, correct, or justify that decision. This finding is consistent with public-sector research showing that the effects of explanations are context-dependent and with critiques warning that human oversight can become symbolic when it is not supported by procedures [11,14].

Third, the study suggests that selective human review may be more defensible than universal review. Current policy debates often frame AI governance as a choice between automation and full human control, whereas institutional practice requires calibrated oversight. The proposed framework operationalizes this middle position by using model

uncertainty and explanation-pattern tension as triggers for review. This approach reduces reviewer burden while preserving the capacity to intervene in the cases where correction is most consequential. The modest escalation rate observed in the simulation indicates that the design is administratively feasible.

These results also enrich the governance discussions in current standards. NIST and ISO/IEC 42001 require organizations to manage AI risks, document controls, and ensure accountability; however, they do not prescribe a single case-level design pattern. Therefore, the traceable bias auditing framework can be considered a candidate for the operationalization of these principles. Prediction and explanation support mapping and measurement are achieved by making the local basis of the recommendation inspectable. This selective review operationalizes managed intervention by defining when human control is mandatory rather than symbolic. The audit trail supports governance and continuous improvement by preserving the evidence required for reconstruction, contest handling, and control review [4–6].

This analysis corresponds with research around organizational governance of responsible AI. Institutions need strong internal processes for responsible AI, and Hadley et al. show how important algorithm review boards are. The current approach enhances these structures by providing them with a characteristic they often miss: consistent case-level control logic. A system approved for deployment by an algorithm review board still needs processes to determine whether its actual use is aligned with fairness and accountability goals [10].

There are also greater consequences for public value. Public AI governance, according to De Fine Licht, contains actual value conflicts, rather than a basic listing of goals that can be held at the same time. The proposed paradigm does not eliminate such arguments, but it does make them manageable. By saving explanations, review actions and arguments, you develop a record that institutions can later discuss to determine whether a given threshold, escalation rule or override pattern is still acceptable. In this respect, traceability extends beyond technical recordkeeping; it becomes institutional memory for contested judgments [15].

This distinction is also relevant to classic fairness mitigation measures. Reweighting, thresholding and post-processing procedures are applied to change the statistical behavior of the model. The suggested framework does not replace these methods. It offers an overlay of governance on top of them, constraining their institutional usage and providing evidence of where further mitigation is still required upstream. In practice, institutions may need both mitigation to lessen structural disagreement in model outputs and traceable reviews to handle the difficult, high-stakes circumstances remaining.

Another crucial part is learning and improving in a continuous way. Institutions that only track outcomes can assess aggregate success, but learn little about the roots of problematic decisions. If they log the explanation factors and reviewer behavior, they have a basis for periodic audit cycles. Concentrated overrides on a few reasons can be a signal of proxy bias, poor calibration, or missing policy restrictions. This creates a feedback loop from operational audits to model updates, the very sort of lifecycle thinking being championed by modern AI governance [3,7].

The discussion also clarifies a conceptual dimension of bias audits. In the literature on fairness, mitigation measures are often classified into pre-processing, in-processing and post-processing phases. The current paradigm is mainly in the realm of post-decision governance, but is nevertheless linked to upstream remediation. The aim is to detect and track high-risk decisions and to generate evidence on opportunities for improvement in the model or data pipeline. It is therefore a diagnostic and regulatory apparatus in this respect [8,9].

Still, the framework entails operational costs. Institutions need reviewer training, better service capabilities for escalation, and oversight to avoid too much or too little escalation. Too much escalation is a reduction in the throughput benefits of automation, whereas too little escalation is a reduction to a simple ceremonial interface. The design therefore represents a configurable governance trade-off rather than a fixed normative prescription. Institutions with limited review capacity could tighten the escalation boundary or increase the explanation-tension level to save reviewer time. Institutions dealing with cases of legal sensitivity may deliberately widen escalation, at the expense of throughput. The core issue is not the price of a review, but whether the marginal hardship of the review is justified by the reduction in alleged unfairness. The current simulation found that a 4.8% escalation rate offered meaningful gains in equity and holistic audit fulfillment. The results suggest that even with small review budgets, considerable governance benefits can be achieved with a strategic review emphasis. Thus, a good governance protocol is to evaluate the review rate, override correctness, demographic parity discrepancy, equal opportunity disparity, and audit thoroughness on a regular basis and alter the thresholds when drift is noticed.

The framework's independence from a single explanatory technology is one of its main advantages. Although the architecture might support SHAP, counterfactual explanations, monotonic scorecards, or domain-specific reason codes, the proof of concept employed a transparent model and local factor decomposition for simplicity. The integration of the explanation into a controlled procedure that connects it to documentation and escalation is what counts, not the type of explanation.

The study also supports an institutional rather than just individual definition of supervision. Green's critique of oversight policy is potent because humans are imperfect, might defer to automation, or might lack the information necessary to intervene successfully. The design presented here tackles part of this complaint by incorporating human review in the clear rules and obligatory logs. The reviewer does not have to make up an oversight out of whole cloth. It is up to the institution to decide when a review is necessary and what records must be created. This is not a cure-all for every oversight problem, but it does put oversight on a better defensible footing than imprecise policy language [14].

Finally, the results of the simulation showing that fairness increased combined with F1 should be read with caution. This is not to say that fair actions always improve utility. This suggests that a selective review method can restore both fairness and accuracy when there are concentrated errors on borderline cases due to historically biased labels. The bigger lesson is that organizations should not assume that fairness and performance are always aligned. Instead, they should look to where biased histories have distorted current operating thresholds and where reviews can restore a better degree of congruence.

This practical argument is illustrated via a brief scholarship screening scenario. Consider, for example, a university utilizing an AI algorithm to score applicants for financial aid. One applicant from a historically deprived area was given a low recommendation despite having exceptional prior academic performance, complete documentation, and confirmation of consistent compliance. Under a score-only approach, the institution could just report the final suggestion. The case is escalated in the suggested framework because of the proximity of the score to the decision boundary and the contradiction of the explanation pattern with a definite negative decision. The reviewer can uphold or overturn the advice, but the explanation is documented in the institutional record, whether or not the suggestion is upheld. The value of the framework, therefore, is not that it guarantees a morally optimal outcome in all cases. It is useful because it makes an opaque rejection into a contestable and reconstructable decision event.

6. Conclusions

In this study, we propose and demonstrate a Traceable Bias Auditing Framework for institutional decision support systems. The approach was supposed to extend beyond abstract calls for ethical AI to incorporate four practical layers: prediction, explanation, human review and audit trail. The essential thesis was that accountability in institutional AI relied not only on explanation, but on the ability to link the explanation to selected intervention and a thorough reconstruction of the final choice.

This assertion was verified with a simulation-based proof of concept. Transparency improved with the addition of explanations not subject to review, but fairness or usefulness was not altered. Enabling the full framework reduced significant fairness gaps, improved F1, kept ranking quality, and provided a complete audit record with a reasonably minor review burden. The results therefore demonstrate that traceable auditing might be a valuable bridge between fairness evaluation, human oversight and corporate governance.

The practical message for organizations that are using AI for eligibility assessment, prioritization or risk flagging is clear: Do not ask about the model's progress. All critical decisions should be explainable, reviewable as necessary, and reproducible later with sufficient evidence to justify the decision. This difference makes a technically competent system institutionally accountable. This paradigm does not compete with fairness mitigation measures, but governs the operational usage of such strategies in relevant institutional workflows.

Practical and Theoretical Implications and Limitations

This work expands the literature on responsible AI and information systems in theory by transferring the burden of accountability from broad principles to the design of decision paths. This suggests that fairness metrics, explanations and monitoring are best seen as inter-related regulations. The artifact and assessment approach provided in this study can be used by other researchers in various organizational areas. The framework offers institutions a simplified paradigm for responsible implementations. The policy team is able to set the escalation protocols. The technical team can identify the effects of case-level features. The audit team can specify the log fields to be logged. Together, these methods provide a decision-support process that can be managed, instead of an unclear model transfer.

The contributions can be summarized as four concepts of reusable design: First, accountability at the event level is better than accountability at the dashboard level. Second, explanations should be linked to action rules rather than treated as passive signs of transparency. Third, human monitoring should not be general and symbolic, but chosen and restricted by regulations. Fourth, logging should preserve not just the ultimate result but also the re-constructability of the whole decision route.

Accordingly, the robustness findings should not be interpreted as evidence of broad generalizability across all governance policies or institutional environments; they indicate stability only within the assumptions of the present proof-of-concept simulation.

But there are important limitations to this study. The evaluation is simulation-based and hence does not capture the entire complexity of real institutional behavior, such as reviewer unreliability, strategic applicant behavior, or political impacts on thresholds. Second, the protected-group concept is binary and reductive; real-world institutions often face many overlapping protected features and multiple harm effects. The proof of concept uses logistic regression to maintain transparency and includes a clear escalation system. More sophisticated models might require more elaborate explanation strategies, alternative interfaces for the reviewers, and more robustness evaluations. The audit completeness index indexes the reconstructability of records, not the substantive quality of each reason.

But a complete journal can nevertheless record bad decisions, even so. The simulation considers overrides relative to a simple threshold of deservingness. Foundational truths are seldom normative in real institutions; eligibility may be contested, value-laden, and contingent on policy development. Hence, the current reference should be seen as a normative analytical standard, rather than as a declaration that institutions can identify in practice an infallible truth label. Future work should focus on improving the thoroughness of the audit by drawing on more detailed indications, such as the quality of rationale, coherence of review, and consistency of explanation override.

These limitations suggest clear avenues for future exploration. Future work could apply the concept to real organizational datasets, extend fairness analysis to multi-group and intersectional situations, and compare alternative explanation kinds, reviewer methodologies, and combination mitigation and governance mechanisms. Longitudinal evaluation is also required so that recurrent overriding trends can be fed back into data curation and model adjustment. Another possibility is to adapt the architecture to generative and conversational decision support, where explanation, evidence grounding and logging are becoming increasingly important in the current governance frameworks [6].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info17070694/s1, File S1: Online Resource 1. Supplementary methods, escalation pseudocode, extended sensitivity results, and bootstrap confidence intervals are provided. File S2: Online Resource 2. Synthetic variable dictionary, reviewer-rule parameters, and audit-log schema (.xlsx). File S3: Online Resource 3. Reproducibility notes, file manifests, and submission-ready supplementary captions.

Funding: This research received no external research funding. The article processing charge (APC) was funded by the Deanship of Graduate Studies and Scientific Research at Qassim University (www.qu.edu.sa) for financial support (QU-APC-2026).

Institutional Review Board Statement: Not applicable. This study did not involve human participants, identifiable personal data, animal subjects, or live institutional interventions.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article and Supplementary Materials. The evaluation used synthetic data generated for a simulation-based proof of concept; no personal, institutional, human-participant, animal, or confidential data were used. Supplementary Materials, including supplementary methods, escalation pseudocode, extended results, synthetic variable dictionary, reviewer-rule parameters, audit-log schema, reproducibility notes, and file manifests, are provided in Files S1–S3 (Online Resources 1–3).

Acknowledgments: The researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University (www.qu.edu.sa) for financial support (QU-APC-2026).

Conflicts of Interest: The author declares no competing interests.

References

1. Fanfani, M.; Ipsaro Palesi, L.A.; Nesi, P. Human-Centered AI for Decision Support Systems: A Systematic Review of Application Domains, Architecture Designs, Current Trends and Future Directions. *Big Data Cogn. Comput.* **2026**, *10*, 186. [CrossRef]
2. Organisation for Economic Co-Operation and Development. *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*; OECD Publishing: Paris, France, 2025. [CrossRef]
3. de Almeida, P.G.R.; dos Santos Júnior, C.D. Artificial intelligence governance: Understanding how public organizations implement it. *Gov. Inf. Q.* **2025**, *42*, 102003. [CrossRef]
4. *ISO/IEC 42001:2023*; Information Technology-Artificial Intelligence-Management System. ISO/IEC (International Organization for Standardization and International Electrotechnical Commission): Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/42001 (accessed on 13 April 2026).

5. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1)*; U.S. Department of Commerce: Washington, DC, USA, 2023. [CrossRef]
6. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*; U.S. Department of Commerce: Washington, DC, USA, 2024. [CrossRef]
7. Janssen, M.; Brous, P.; Estevez, E.; Barbosa, L.S.; Janowski, T. Data governance: Organizing data for trustworthy artificial intelligence. *Gov. Inf. Q.* **2020**, *37*, 101493. [CrossRef]
8. González-Sendino, R.; Serrano, E.; Bajo, J.; Novais, P. A review of bias and fairness in artificial intelligence. *Int. J. Interact. Multimed. Artif. Intell.* **2024**, *9*, 5–17. [CrossRef]
9. Ahmad, A.; Vallès, Y.; Idaghdour, Y. Bias in AI systems: Integrating formal and socio-technical approaches. *Front. Big Data* **2026**, *8*, 1686452. [CrossRef] [PubMed]
10. Hadley, E.; Blatecky, A.; Comfort, M. Investigating algorithm review boards for organizational responsible artificial intelligence governance. *AI Ethics* **2025**, *5*, 2485–2495. [CrossRef]
11. Aoki, N.; Tatsumi, T.; Naruse, G.; Maeda, K. Explainable AI for government: Does the type of explanation matter to the accuracy, fairness, and trustworthiness of an algorithmic decision as perceived by those who are affected? *Gov. Inf. Q.* **2024**, *41*, 101965. [CrossRef]
12. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and explanation in AI-informed decision making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. [CrossRef]
13. Papadakis, T.; Christou, I.T.; Ipektsidis, C.; Soldatos, J.; Amicone, A. Explainable and transparent artificial intelligence for public policymaking. *Data Policy* **2024**, *6*, e10. [CrossRef]
14. Green, B. The flaws of policies requiring human oversight of government algorithms. *Comput. Law Secur. Rev.* **2022**, *45*, 105681. [CrossRef]
15. de Fine Licht, K. Resolving value conflicts in public AI governance: A procedural justice framework. *Gov. Inf. Q.* **2025**, *42*, 102033. [CrossRef]
16. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 30 August–2 September 2019; pp. 59–68. [CrossRef]
17. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 220–229. [CrossRef]
18. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé, H., III; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92. [CrossRef]
19. Reisman, D.; Schultz, J.; Crawford, K.; Whittaker, M. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*; AI Now Institute: New York, NY, USA, 2018; Available online: https://ainowinstitute.org/wp-content/uploads/2023/04/aiareport2018.pdf (accessed on 15 April 2026).
20. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design science in information systems research. *MIS Q.* **2004**, *28*, 75–105. [CrossRef]
21. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A design science research methodology for information systems research. *J. Manag. Inf. Syst.* **2007**, *24*, 45–77. [CrossRef]