Ethical AI in Human Resource and Financial Decision-Making: A Benchmark-Informed Framework for Fairness, Transparency and Accountability in Automated Systems

Rahul Patowary, Dr. Debosree Sarma, Rinti Borah, Dr. Rinku Agarwal, Puspakshi Sarmah

Assistant Professor, Department of Business Administration, NERIM Group of Institutions, Guwahati
E-Mail Id: rr.patowary@gmail.com
Assistant Professor, Department of Business Administration, NERIM Group of Institutions, Guwahati
E-Mail Id: sarma.debosree@gmail.com
Assistant Professor, Department of Computer Science, NERIM Group of Institutions, Guwahati
E-Mail Id: borahrinti91@gmail.com
HOD & Associate Professor, Department of Business Administration, NEF College, Guwahati.
E-Mail Id: rinkuagar@gmail.com
Assistant Professor, Department of Computer Science, NEF College, Guwahati.
E-Mail Id: puspakshi@gmail.com

Abstract

Automated decision systems now shape consequential human-resource and financial outcomes, including recruitment, candidate screening, employee analytics, credit underwriting, fraud detection and portfolio risk assessment. Yet adoption has advanced faster than the evidentiary routines needed to demonstrate fairness, transparency, privacy protection and accountable human control. This article develops and evaluates a benchmark-informed ethical AI framework for HR and financial decision-making using four complementary evidence streams: recent sector surveys and institutional reports, peer-reviewed and public case evidence on algorithmic disparity, reproducible experiments on public credit-risk benchmarks, and structured coding of major governance instruments. The empirical component analyzes the South German/German Credit benchmark and the FICO HELOC Explainable Machine Learning Challenge dataset using transparent and ensemble classifiers, cross-validation, calibration loss, subgroup fairness metrics and permutation importance. The German Credit analysis shows that similar discrimination performance can coexist with materially different fairness profiles: logistic regression achieved mean ROC-AUC of 0.783 but displayed an age-based equal-opportunity difference of -0.122, while random forests achieved ROC-AUC of 0.785 with substantially smaller equal-opportunity difference (-0.004) but lower balanced accuracy. On FICO HELOC, logistic regression achieved holdout ROC-AUC of 0.793 and cross-validated ROC-AUC of 0.796, indicating that interpretable models can be competitive for some credit-risk tasks; however, the absence of protected attributes prevents a complete fairness audit. Source-derived evidence further shows an adoption-governance gap: organizations report extensive AI use in HR and finance, while many lack policy clarity, measurement routines or full understanding of model behavior. The proposed FAIR-HITL framework integrates fairness-aware modelling, explainable AI, privacy-preserving data governance, human-in-the-loop review, regulatory compliance and post-deployment redress into a single auditable lifecycle. The study contributes a reproducible cross-sector methodology and an implementation-ready governance model for high-stakes automated decision systems.

Keywords: ethical AI; algorithmic fairness; human resource analytics; credit scoring; explainable AI; human-in-the-loop governance

How to cite this article: Patowary R, Sarma D, Borah R, Agarwal R, Sarmah P. Ethical AI in Human Resource and Financial Decision-Making: A Benchmark-Informed Framework for Fairness, Transparency and Accountability in Automated Systems. Int J Drug Deliv Technol. 2026;16(63s):1521-1535. DOI: 10.25258/ijddt.16.63s.153

1. Introduction

Artificial intelligence (AI) has moved from a peripheral analytics capability to a decision infrastructure in organizations. In HR, algorithmic systems are used to draft job advertisements, parse resumes, rank candidates, support interview scheduling, infer skills, model employee retention and monitor workforce productivity. In financial services, machine learning assists credit scoring, lending decisions, anti-money-laundering triage, fraud detection, pricing, investment research, customer segmentation and model-risk surveillance. Recent cross-sector indicators [1] show rapid diffusion of AI and generative AI across business functions, while HR surveys [2] and HR governance evidence [3] indicate substantial adoption in the HR domain examined here; finance survey evidence [4] and OECD workplace evidence [5] provide parallel context.

This diffusion is ethically consequential because HR and financial decisions allocate opportunities, income security, credit access and social mobility. Errors are not merely technical defects; they can compound historical disadvantage, obscure responsibility, weaken due-process rights and create new forms of informational asymmetry. The policy environment has therefore shifted toward risk-based governance. NIST emphasizes socio-technical risk management [6] and bias control [7], the EU AI Act classifies many employment and creditworthiness systems as high-risk [8], ISO/IEC 42001 formalizes AI management systems [9], and US regulators have clarified adverse-action duties [10], sample-form expectations [11], employment-discrimination obligations [12] and disability-discrimination duties [13].

The empirical risk is no longer hypothetical. Large-scale deployed hiring evidence [14] reports measurable adverse-impact exposure in algorithmic candidate screening, while HMDA loan-level data [15] and mortgage-denial research [16] continue to identify racial and ethnic denial disparities even after standard credit-risk controls. These findings do not imply that all automated systems are discriminatory, nor that human decision-making is unbiased. They do show that high-stakes AI cannot be responsibly deployed on the assumption that predictive accuracy alone is sufficient evidence of ethical performance.

This article addresses a practical research gap. HR and financial AI scholarship often examines fairness, explainability, privacy and accountability as separate concerns. Industrial governance, by contrast, must manage them jointly across data intake, modelling, deployment, human review, notice, appeal and post-deployment monitoring. The central research question is therefore: how can organizations design and evidence an ethical AI lifecycle that is empirically testable, legally aligned and operationally feasible across both HR and financial decision contexts?

The study makes four contributions. First, it provides a critically appraised evidence base from recent surveys, public benchmarks, case studies, legal instruments and peer-reviewed literature. Second, it conducts reproducible benchmark experiments on public credit-risk data to illustrate performance-fairness-explainability trade-offs. Third, it codes major governance instruments into a transparent control matrix rather than presenting a synthetic compliance score. Fourth, it proposes the FAIR-HITL framework, a novel lifecycle model that combines fairness-aware modelling, explainable AI, privacy-preserving governance, human-in-the-loop accountability and redress mechanisms.

2. Literature Review2.1 AI in human-resource decision-making Recent HRM research shows that algorithmic systems are increasingly embedded in recruitment and personnel analytics, but the evidence base remains uneven. Reviews of algorithmic decision-making in HR identify recurring risks in data representativeness, construct validity, proxy discrimination, automation bias and opacity in vendor-controlled systems [22]. Ethical HR analytics scholarship further argues that organizational adoption often precedes the development of role clarity, employee voice and accountability structures [23]. A multidisciplinary survey of algorithmic hiring underscores that technical fairness tools, employment-law standards and organizational practice still operate with incomplete translation between them [24].

The HR domain is distinctive because many predictive targets are socially constructed. “Fit”, “potential”, “culture add”, “engagement risk” and “leadership promise” are not direct measurements of immutable traits; they are managerial categories shaped by institutional history, job design and measurement choices. Bias can enter through historical labels, selective observation, performance ratings, referral networks, occupational segregation, language proxies and disability-related accommodations. Evidence from algorithmic hiring also shows a risk of monoculture, where multiple employers or vendors use similar screening models and thereby amplify correlated exclusion across applications and positions [14].

Empirical and sociotechnical studies of hiring technology caution against treating algorithmic scores as neutral evidence. Vendor claims can be difficult to validate externally [25], recruiters may use scores selectively rather than as intended [26], and candidate-facing explanations are often too vague to support meaningful contestation. Recent HRM work on AI assimilation emphasizes that adoption success depends not only on algorithmic performance but also on institutional readiness, user trust, training, governance participation and alignment with employment-law obligations [27]. These findings suggest that ethical HR AI requires job-related validation, adverse-impact testing, candidate notice, disability accommodation, human review authority and post-deployment validity monitoring.

2.2 AI in financial decision-making

Financial institutions have long used statistical models, but contemporary AI expands the scale and complexity of credit, fraud, pricing and risk decisions. In credit scoring, machine-learning methods can capture nonlinear relationships and improve discrimination in some portfolios, but they also create new fairness and explainability challenges. The fairness literature in credit scoring shows that interventions such as pre-processing, in-processing and post-processing

can reduce measured disparities, yet they may alter profitability, calibration or error allocation across groups [31].

Explainability is especially important in finance because adverse-action regimes require specific and accurate reasons, and model-risk management expects traceable validation evidence. Studies of explainable AI in credit risk [32] and SHAP/LIME evaluations [33] show that local and global explanations can help identify dominant credit-risk drivers, but explanations are not automatically faithful, stable or legally sufficient. A model can produce a plausible explanation while relying on unstable proxies or historically biased labels. Conversely, an interpretable model may perform competitively when the decision problem is structured and the feature space has strong domain meaning, as observed in several credit-risk benchmarks.

Algorithmic financial decision-making also intersects with privacy and data minimization. Credit, fraud and investment models may use transaction histories, device fingerprints, geolocation, behavioral signals and third-party data. These features can improve predictive coverage but increase surveillance risk and proxy discrimination. Fair lending therefore requires not only subgroup metric assessment but also reason-code validity, data provenance checks, feature governance, reject-inference caution, model monitoring and appeal mechanisms.

2.3 Ethical AI frameworks and their limitations

Ethical AI frameworks converge around fairness, transparency, human oversight, privacy, robustness and accountability, but they differ in enforceability and operational specificity. The OECD principles [5] articulate values and policy recommendations; the NIST AI Risk Management Framework [6] provides a flexible Map-Measure-Manage-Govern structure; NIST bias guidance [7] complements this structure; the EU AI Act [8] introduces risk-based legal obligations; ISO/IEC 42001 [9] specifies a management-system approach; and US sectoral regulators translate consumer-credit [10], sample-form [11], employment-discrimination [12] and disability-discrimination duties [13] to AI-assisted decisions.

Technical toolkits have made fairness assessment more accessible. Fairlearn [29] and AI Fairness 360 [30], for example, provide metrics and mitigation algorithms that can be integrated into machine-learning workflows. Yet fairness metrics encode normative choices. Statistical parity, equal opportunity, equalized odds and calibration cannot generally be optimized simultaneously when base rates differ. Metric choice must therefore be tied to the decision context, harm model and legal standard rather than selected for numerical convenience. Surveys of bias and fairness in machine learning emphasize that data, measurement, modelling and deployment stages each produce different failure modes [28].

Explainability research provides complementary tools but similar caveats. LIME [34] and SHAP [35] are widely used for local and global feature attribution, but explanation quality depends on model behavior, data distribution, perturbation assumptions and user interpretation. Transparency is therefore not equivalent to publishing feature importances. For HR and finance, transparency should include documentation of model purpose, data provenance, validation results, subgroup performance, reason codes, human review points, appeal routes and monitoring thresholds. Model cards [36], datasheets [37] and external audit proposals [38] offer useful documentation foundations, but adoption remains inconsistent outside mature model-risk functions.

2.4 Research gaps

Three gaps motivate the present study. First, HR and financial AI are often studied in separate literatures, even though both involve high-stakes allocation, protected-class concerns, explanations and contestability. Second, empirical research frequently reports accuracy without documenting fairness, calibration, model interpretability and governance evidence together. Third, many conceptual ethical AI frameworks lack a reproducible bridge to benchmark analysis, source-derived sector evidence and concrete implementation controls. This article addresses these gaps by integrating public-data experiments, published disparity evidence, regulatory control mapping and a cross-sector governance framework.

3. Research Methodology

3.1 Research design

The study uses an evidence-integrated research design combining structured literature synthesis, benchmark-based empirical analysis and governance-document coding. The design is appropriate because the user did not provide proprietary organizational data and because real HR deployment data are rarely publicly available at individual level. Rather than fabricate primary data, the study separates three forms of evidence: source-derived descriptive indicators from recent reports, reproducible model results on public benchmarks, and conceptual synthesis supported by coded governance controls.

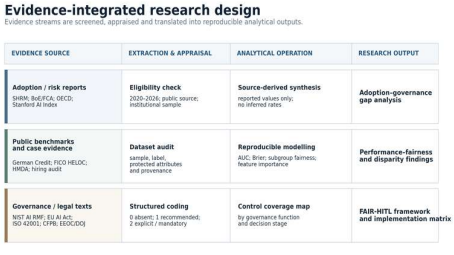

Figure 1. Evidence-integrated research design. The diagram separates public-source descriptive indicators, benchmark-based empirical analysis and governance-document coding to prevent conceptual framework elements from being misrepresented as measured outcomes.

The research questions are: RQ1 - What adoption-governance gaps are visible in recent HR and financial AI evidence? RQ2 - How do performance, fairness and explainability trade-offs appear in reproducible public credit-risk benchmarks? RQ3 - Which governance controls are consistently required or recommended across leading AI frameworks and sector regulators? RQ4 - How can these controls be integrated into a lifecycle model that is operationally meaningful for HR and financial institutions?

3.2 Data collection and eligibility criteria

Evidence was collected from four source families: (i) 2020-2026 institutional surveys and industry reports on AI adoption and governance; (ii) peer-reviewed or official public evidence on algorithmic hiring and mortgage disparities; (iii) public benchmark datasets used for credit-risk and fairness research; and (iv) governance instruments with direct relevance to automated HR or financial decisions. Sources were retained only when the reported indicator, dataset provenance or governance requirement was traceable to an official report, public dataset, peer-reviewed study or regulator-maintained document. Blog posts, consultancy graphics without method disclosure and unverifiable numerical claims were excluded.

The empirical datasets were selected on the basis of public accessibility, documented feature definitions, relevance to financial decision-making and reproducibility. German Credit/South German Credit was retained because it remains a standard fairness and credit-risk benchmark, but it is explicitly treated as a diagnostic historical benchmark rather than a contemporary population sample. FICO HELOC was retained because it is a real credit-line dataset designed for explainable machine-learning evaluation. HR analysis relies on peer-reviewed deployment-scale hiring evidence and recent HR survey evidence because individual-level public HR screening datasets with protected attributes and validated outcome labels are generally unavailable for privacy, commercial and legal reasons.

Table 1. Evidence streams used in the study.

Evidence stream	Sources used	Role in analysis	Quality control	Main limitation
Adoption and risk surveys	Stanford AI Index; SHRM AI in HR; Bank of England/ FCA AI survey; OECD algorithmic management study	Establish sector diffusion, governance maturity and reported risk context	Recent, institutionally traceable, explicit samples or methods	Not pooled into a meta-analysis because samples, geographies and denominators differ
Public benchmark datasets	German Credit/South German Credit; FICO HELOC Explainable ML Challenge	Reproducible modelling, fairness diagnostics, calibration and XAI analysis	Public files; documented labels; reusable preprocessing	German Credit is historical; FICO HELOC omits protected attributes
Case and disparity evidence	Algorithmic hiring adverse-impact study; HMDA/ mortgage-denial evidence	Ground empirical motivation in deployed-system and lending disparities	Large-scale or official loan-level evidence	Different designs and legal contexts; not causal proof for every AI deployment
Governance and legal instruments	NIST AI RMF; NIST SP 1270; EU AI Act; ISO/IEC 42001; OECD principles; CFPB and EEOC/DOJ guidance	Build coded control matrix and framework requirements	Official or standards-body provenance	Coverage coding is transparent but not a substitute for legal advice or inter-rater

Evidence stream	Sources used	Role in analysis	Quality control	Main limitation
				consensus

Table 2. Public datasets, audit decisions and preprocessing.

Dataset	Rows	Original columns	Target	Positive rate	Protected attribute used	Preprocessing	SHA-256 prefix
German Credit / South German coding benchmark	1,000	21	credit_risk (1=good, 0=bad)	0.700	age<25 vs age>=25; sex/marital-status field excluded	drop age and personal_status_sex from prediction; one-hot categorical variables; standardize numeric variables	ac003d1158e44018...
FICO HELOC Explainable Challenge	10,459	24	RiskPerformance (Good=1, Bad=0)	0.4778	none released; fairness audit cannot be performed without protected attributes	replace -8 and -9 with missing, retain -7 as condition-not-met; median imputation with missing indicators; standardize numeric variables	e5a914f742ad7f84...

Note. Positive rate denotes the proportion coded as the favorable class. The German benchmark uses age only for fairness diagnostics because the personal-status/sex field is entangled and historically miscoded; FICO HELOC does not release protected attributes, so fairness metrics are not reported for that dataset.

3.3 Preprocessing and model specification

For German Credit, the label was recoded so that favorable credit risk equals 1 and unfavorable risk equals 0. Age was binarized as under 25 versus 25 and older for subgroup diagnostics. Age and the personal-status/sex field were excluded from model features. Categorical variables were one-hot encoded and numeric variables standardized inside cross-validation folds to avoid leakage. The analysis used five stratified folds with a fixed random seed (20260616). Models included logistic regression, a depth-limited decision tree, random forest, gradient boosting, logistic regression with reweighing by age group and a diagnostic equal-opportunity thresholding variant. The equal-opportunity thresholding result is reported as an analytical sensitivity check rather than a deployment recommendation, because group-specific thresholds can raise legal and ethical concerns depending on jurisdiction and use case.

For FICO HELOC, the target was Good repayment performance versus Bad performance. Following the dataset documentation, special values -8 and -9 were treated as missing and imputed using training-

fold medians with missingness indicators; -7 was retained as a substantive "condition not met" code because it denotes inapplicability rather than unavailable information. Models included logistic regression, depth-limited decision tree, random forest and gradient boosting. A 70/30 stratified train-test split was used for holdout ROC curves and Brier scores, and five-fold cross-validation was used for stability checks. Permutation importance was computed as the mean decrease in ROC-AUC on the holdout set.

All experiments used Python with pandas 2.2.3, NumPy 2.3.5 and scikit-learn 1.8.0. The reported confidence-style intervals in result tables are descriptive fold-based intervals, not causal population confidence intervals; they indicate cross-validation variability under the specified splits and preprocessing pipeline.

3.4 Evaluation criteria

Table 3.Evaluation criteria and interpretation.

Criterion	Operational definition	Ethical relevance	Limitation
Accuracy	Share of correct classifications	Overall predictive performance	Can be misleading under class imbalance
Balanced accuracy	Mean of sensitivity and specificity	Performance under unequal class distributions	Does not encode group fairness
ROC-AUC	Rank discrimination across thresholds	Model comparison independent of one threshold	Insensitive to calibration and decision costs
Brier score	Mean squared probability error	Calibration and probabilistic quality	Lower is better; affected by prevalence
Statistical parity difference	P(Yhat=1|A=unprivileged) - P(Yhat=1|A=privileged)	Selection-rate disparity	Can conflict with validity when base rates differ
Disparate-impact ratio	Selection rate unprivileged / privileged	Four-fifths-rule style screening diagnostic	Not by itself a full legal test
Equal-opportunity difference	TPR_unprivileged - TPR_privileged	Difference in true-positive access to favorable decisions	Requires reliable outcome labels
Equalized-odds gap	Max absolute TPR/FPR difference	Joint error-rate disparity	May trade off with calibration and accuracy
Permutation importance	Decrease in ROC-AUC after feature shuffling	Global explanation of predictive reliance	Correlated features complicate interpretation
Governance coverage code	0 absent; 1 recommended; 2 explicit or mandatory	Compliance-readiness mapping across controls	Document coding, not proof of implementation quality

3.5 Governance-document coding

Governance instruments were coded across eight control dimensions: risk classification, data governance, fairness testing, XAI/reason-giving, human oversight, privacy/security, monitoring/audit and appeal/remediation. A code of 0 indicates that the control is absent or not meaningfully articulated in the instrument; 1 indicates that the control is recommended or implied; 2 indicates explicit, mandatory or strongly operationalized coverage. Coding was conservative and based on the text and regulatory purpose of each instrument. The matrix is used to identify convergent governance expectations, not to rank frameworks normatively.

4. Proposed Ethical AI Framework

4.1 FAIR-HITL framework

The proposed FAIR-HITL framework is a lifecycle model for high-stakes automated

decision systems in HR and finance. “FAIR” denotes fairness-aware data and modelling, auditable explainability, institutional accountability and rights-preserving governance. “HITL” denotes human-in-the-loop decision authority that is meaningful rather than ceremonial: humans must have competence, time, discretion, escalation pathways and responsibility for overrides. The framework is designed for decisions that materially affect employment, credit access, pricing, monitoring intensity or appeal outcomes.

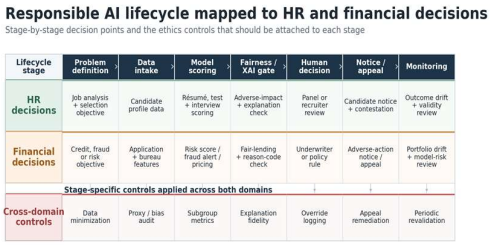

Figure 2.FAIR-HITL framework for ethical AI in HR and financial decision-making. The figure is a conceptual synthesis supported by the benchmark and governance analyses; it is not presented as a measured performance outcome.

The framework begins with use-case risk classification. A low-risk analytics dashboard does not require the same controls as automated candidate rejection or loan denial. Risk classification should consider decision consequence, automation level, scale, affected population, contestability, data sensitivity and likelihood of proxy discrimination. The second stage is data governance, including provenance, lawful basis, data minimization, representativeness, label validity, feature purpose, retention rules and vendor data flows. The third stage is fair model design, where model selection, feature engineering, threshold choice and validation are evaluated against domain-specific harm models. The fourth stage is a deployment gate requiring explainability, human review authority, reason codes, documentation and sign-off. The fifth stage is monitoring and redress, where drift, appeals, incidents, subgroup errors and override patterns are periodically reviewed.

4.2 Ethics lifecycle and sector workflows

The same governance logic must be translated differently across HR and financial domains. HR workflows require job-relatedness, selection validity, adverse-impact analysis, candidate notice, disability accommodation and review of downstream workforce effects. Financial workflows require fair-lending analysis, model-risk validation, affordability checks, adverse-action reasons, fraud-control governance, portfolio monitoring and customer appeals. Figure 3 places these domain workflows on a shared ethical lifecycle.

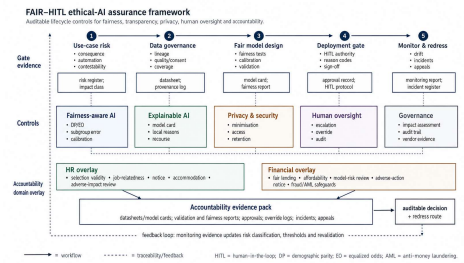

Figure 3.AI ethics lifecycle with HR and financial decision workflows. Controls are activated at need definition, data intake, model scoring, fairness/XAI gate, human decision, notice/appeal and monitoring stages.

4.3 Accountability evidence pack

A central design feature of FAIR-HITL is the accountability evidence pack. Each system should maintain versioned datasheets, model cards, validation reports, subgroup metric logs, reason-code testing, human-override records, vendor assessment, incident registers, appeal outcomes and monitoring reviews. The evidence pack prevents a common governance failure: treating ethical AI as a policy declaration rather than an auditable chain of artifacts. It also supports cross-functional accountability by assigning evidence owners to HR, risk, legal, compliance, model development, procurement, information security and business decision owners.

5. Results and Analysis

5.1 Adoption-governance gap in HR and financial AI

Recent source-derived indicators reveal a consistent pattern: AI adoption is broad, but governance maturity is uneven. Cross-sector adoption indicators are high, HR reports increasing use of AI in HR tasks, and UK financial firms report widespread AI use. However, HR evidence also shows limited measurement of AI implementation success and incomplete AI-use policy clarity, while financial-sector evidence shows that a substantial share of firms report only partial understanding of AI technologies. These values are not pooled statistically because their denominators differ; they are presented as traceable indicators of a governance gap.

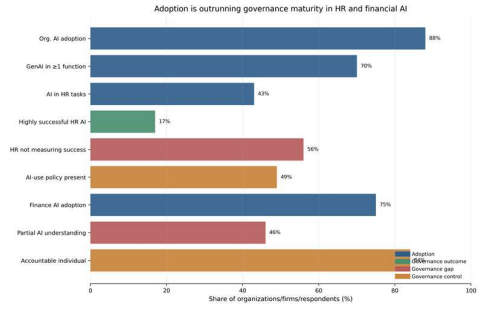

Figure 4. Adoption is outpacing governance maturity in HR and financial AI. Bars report only values stated in cited public sources; they are not estimated from a common sample.

Table 4.Source-derived AI adoption and governance indicators.

Domain	Indicator	Reported value	Classification
Cross-sector	Organizations using AI (Stanford AI Index 2026)	88%	Adoption
Cross-sector	Organizations using genAI in ≥1 function	70%	Adoption
HR	Organizations leveraging AI in HR tasks (SHRM 2025)	43%	Adoption
HR	HR AI implementations rated highly successful	17%	Governance outcome
HR	AI-in-HR organizations not measuring success (SHRM 2026)	56%	Governance gap
HR	AI-using/piloting organizations with AI-use policies	49%	Governance control
Finance	UK financial firms already using AI (BoE/FCA 2024)	75%	Adoption
Finance	AI-using/planning firms with only partial AI understanding	46%	Governance gap
Finance	AI-using UK firms with accountable individual (FCA)	84%	Governance control

Note. Indicators have different samples and geographies and are therefore interpreted descriptively rather than as a pooled rate.

5.2 German Credit benchmark: performance-fairness trade-offs

The German Credit benchmark illustrates why fairness evaluation cannot be inferred from accuracy. Logistic regression and random forest models achieved very similar mean ROC-AUC values, but their subgroup fairness profiles differed materially. Logistic regression achieved mean ROC-AUC of 0.783 and balanced accuracy of 0.673, but its age-based equal-opportunity difference was -0.122, indicating lower true-positive access for the under-25 group under the chosen favorable-outcome coding. Random forest achieved mean ROC-AUC of 0.785 and a much smaller equal-opportunity difference of -0.004, but balanced accuracy was lower at 0.592. The depth-limited tree was more interpretable but weaker on discrimination and balanced accuracy. The diagnostic equal-opportunity thresholding variant improved balanced accuracy to 0.690 and produced a disparate-impact ratio above 1, but it did not eliminate equalized-odds differences and should not be interpreted as a legally neutral deployment solution.

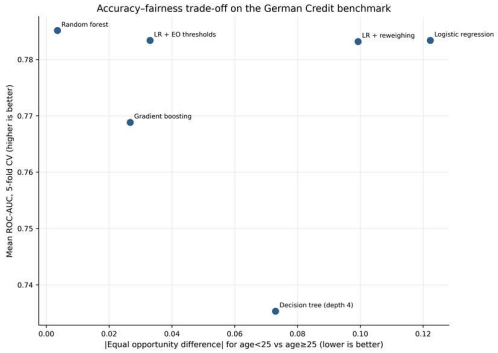

Figure 5.Accuracy-fairness trade-off on the German Credit benchmark. The x-axis reports the absolute equal-opportunity difference for age<25 versus age>=25, where lower is better; the y-axis reports mean ROC-AUC from five-fold cross-validation.

Table 5. German Credit five-fold cross-validation: performance and age-based fairness diagnostics.

Model	ROC-AUC mean +/- SD	Balanced accuracy	Disparate-impact ratio	Equal-opportunity diff.	Equalized-odds gap
Logistic regression	0.783 +/- 0.041	0.673	0.849	-0.122	0.197

Model	ROC-AUC mean +/- SD	Balanced accuracy	Disparate-impact ratio	Equal-opportunity diff.	Equalized-odds gap
LR + reweighing (age)	0.783 +/- 0.042	0.670	0.907	-0.099	0.234
LR + EO thresholds (age; diagnostic)	0.783 +/- 0.041	0.690	1.052	0.033	0.163
Decision tree (depth 4)	0.735 +/- 0.029	0.609	0.944	-0.073	0.141
Random forest	0.785 +/- 0.030	0.592	0.976	-0.004	0.116
Gradient boosting	0.769 +/- 0.030	0.590	1.003	0.027	0.159

Note. Favorable class = good credit risk. Protected attribute = age<25 versus age>=25. Age and personal-status/sex variables were excluded from predictors. Results are benchmark diagnostics and not claims about contemporary lending populations.

The German results support three methodological conclusions. First, fairness intervention is metric-specific: reweighing shifted selection rates and disparate impact but did not uniformly improve equal opportunity or equalized odds. Second, interpretability is not synonymous with fairness: the depth-limited tree is easier to inspect, yet it does not dominate the fairness metrics. Third, threshold selection is ethically loaded. A threshold that improves a group fairness metric can still produce unacceptable calibration, individual fairness, legal or operational consequences. These results justify the framework requirement that thresholding decisions be documented, reviewed and monitored as governance events rather than treated as technical defaults.

5.3 FICO HELOC benchmark: interpretable performance and explanation constraints

The FICO HELOC analysis provides a complementary lesson. On the holdout set, logistic regression achieved ROC-AUC of 0.793, accuracy of 0.721 and Brier score of 0.185. Random forest achieved a similar ROC-AUC of 0.791 and accuracy of 0.720, while gradient boosting and the depth-limited tree performed slightly below logistic regression. Cross-validation results were consistent: logistic regression achieved mean ROC-AUC of 0.796, followed closely by random forest and gradient boosting. The result does not imply that linear models always dominate in credit risk; rather, it shows that a transparent baseline can be highly competitive and should be considered before deploying more opaque models in regulated decision settings.

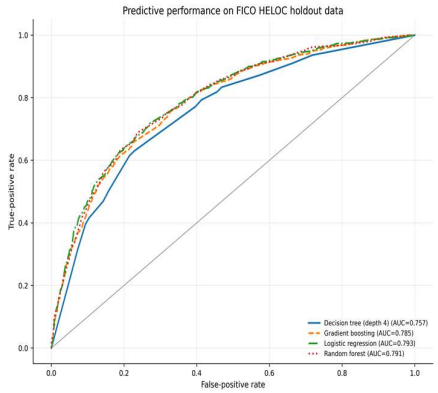

Figure 6.Predictive performance on FICO HELOC holdout data. ROC curves compare logistic regression, depth-limited decision tree, random forest and gradient boosting.

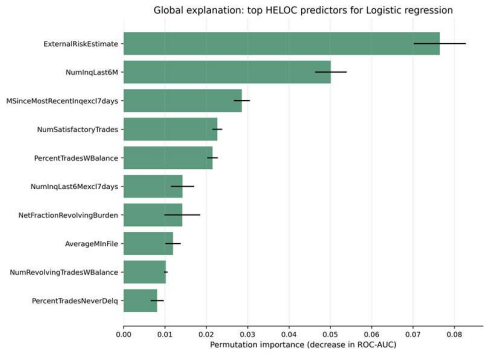

Figure 7. Global explanation: top FICO HELOC predictors for logistic regression. Bars show permutation importance measured as mean decrease in ROC-AUC on the holdout set; error bars show repeated-shuffle variability.

Table 6.FICO HELOC holdout and cross-validation performance.

Model	Hold out accuracy	Hold out balanced accuracy	Hold out ROC-AUC	Hold out Brier	CV ROC-AUC mean +/- SD
Logistic regression	0.721	0.719	0.793	0.185	0.796 +/- 0.009
Decision tree (depth 4)	0.704	0.701	0.757	0.199	0.765 +/- 0.014
Random forest	0.720	0.717	0.791	0.188	0.794 +/- 0.012
Gradient boosting	0.718	0.716	0.785	0.188	0.792 +/- 0.014

Note. Protected attributes are not released in FICO HELOC; therefore, fairness metrics are not reported. Lower Brier score indicates better probabilistic calibration.

Table 7. Top FICO HELOC permutation-importance features for logistic regression.

Feature	Mean decrease in ROC-AUC	SD
ExternalRiskEstimate	0.0765	0.0063
NumInqLast6M	0.0501	0.0038
MSinceMostRecentInqexcl7days	0.0286	0.0020
NumSatisfactoryTrades	0.0227	0.0012
PercentTradesWBalance	0.0215	0.0013
NumInqLast6Mexcl7days	0.0142	0.0028
NetFractionRevolvingBurden	0.0142	0.0043
AverageMInFile	0.0119	0.0019

Note. Feature importance is a global explanation of model reliance, not a causal estimate. Correlated credit variables can distribute importance across related predictors.

The most important HELOC predictors were ExternalRiskEstimate and recent inquiry measures, followed by account-satisfaction, revolving-burden and trade-balance features. This ranking is plausible for credit-risk modelling, but the absence of protected attributes creates a major audit limitation. A model can be accurate and explainable while still producing disparate effects that cannot be measured. This reinforces a central governance point: omitting protected attributes from datasets may reduce direct use but can prevent fairness auditing unless lawful and privacy-preserving evaluation mechanisms are available.

5.4 Evidence of real-world disparity in hiring and lending

The benchmark results should be interpreted alongside deployed-system evidence. In algorithmic hiring, a large-scale study of applications, applicants, positions and employers found that substantial shares of applications by Black and Asian candidates were submitted to positions with adverse-impact exposure, and that a non-trivial share of applicants applying to ten positions could be rejected by all models. This type of correlated exclusion is particularly important because job seekers often apply to multiple openings and may encounter similar models across employers.

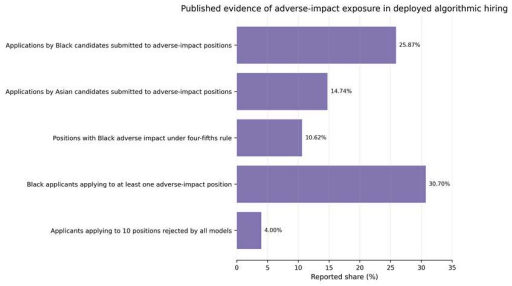

Figure 8.Published evidence of adverse-impact exposure in deployed algorithmic hiring.

Table 8.Source-derived adverse-impact indicators in deployed algorithmic hiring.

Indicator	Reported value
Applications by Black candidates submitted to adverse-impact positions	25.87%
Applications by Asian candidates submitted to adverse-impact positions	14.74%
Positions with Black adverse impact under four-fifths rule	10.62%
Black applicants applying to at least one adverse-impact position	30.70%
Applicants applying to 10 positions rejected by all models	4.00%

Note. The values summarize published deployed-hiring evidence and are included to contextualize governance risk in HR AI, not to infer causality in every hiring platform.

In mortgage lending, denial disparities remain visible after controls such as credit score, loan-to-value ratio and debt-to-income ratio. Published evidence reports excess denial probabilities of 2.9 percentage points for Black applicants, 2.2 percentage points for Asian applicants and 1.5 percentage points for Latinx applicants relative to White applicants. These estimates are not an AI-specific indictment; they are a reminder that credit-risk systems operate in markets where historical, institutional and data-generating processes can produce persistent disparities. AI systems deployed in such settings must therefore be audited against both predictive and distributional criteria.

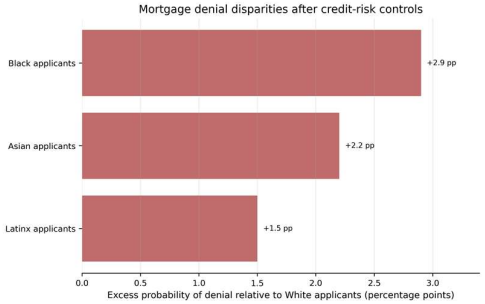

Figure 9.Mortgage-denial disparities after credit-risk controls. Values report excess denial probability relative to White applicants, in percentage points, from the cited mortgage-denial study.

5.5 Governance coverage and compliance-readiness mapping

The governance-document coding shows convergence around core controls but different levels of operational specificity. The EU AI Act and NIST AI RMF provide broad lifecycle coverage; ISO/IEC 42001 emphasizes management-system processes; OECD principles provide value-level orientation; CFPB circulars strongly operationalize adverse-action reasons and fair-lending accountability; and EEOC/DOJ guidance provides domain-specific employment and disability-rights controls. The matrix therefore supports a layered governance strategy rather than a single-framework substitution approach.

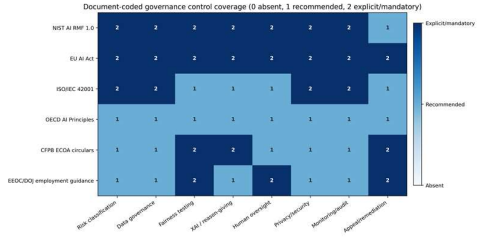

Figure 10. Document-coded governance control coverage. Coding: 0 = absent, 1 = recommended or implied, 2 = explicit, mandatory or strongly operationalized. The heatmap is based on document coding, not measured organizational compliance.

Table 9.Governance-control coverage matrix.

Framework	Risk	Data	Fair	XAI	HITL	Privacy	Monitor	Redress
NIST AI RMF 1.0	2	2	2	2	2	2	2	1
EU AI Act	2	2	2	2	2	2	2	2
ISO/IEC 42001	2	2	1	1	1	2	2	1
OECD AI Principles	1	1	1	1	1	1	1	1
CFPB ECOA circulars	1	1	2	2	1	1	1	2
EEOC/DOJ employment guidance	1	1	2	1	2	1	1	2

Note. Codes: 0 absent; 1 recommended or implied; 2 explicit, mandatory or strongly operationalized. Abbreviations: XAI = explainable AI; HITL = human-in-the-loop.

5.6 Risk and mitigation matrix

Table 10.Cross-sector risk and mitigation matrix for FAIR-HITL implementation.

Risk	HR instantiation	Financial instantiation	Mitigation control	Primary owner / timing
Historical label bias	Past hiring ratings or promotion outcomes encode	Past repayment or denial labels reflect unequal credit access	Label audit, counterfactual review, subgroup	Model owner + domain owner before training

Risk	HR instantiation	Financial instantiation	Mitigation control	Primary owner / timing
	organizational inequality		validation, data-sheet documentation	
Proxy discrimination	Education, gaps, location, referrals or language proxies for protected status	ZIP code, device data, spending patterns or inquiry variables proxy protected class	Feature-purpose review, proxy testing, constrained modelling, lawful protected-attribute audit	Data governance board before feature approval
Opacity and weak reasons	Candidate receives unexplained rejection or generic score rationale	Borrower receives generic adverse-action reason despite complex model	Model cards, local explanations, reason-code testing, human-readable notices	Compliance and model-risk teams before deployment
Automation bias	Recruiters defer to scores despite contradictory evidence	Underwriters or fraud analysts rubber-stamp model flags	HITL training, override authority, sampled review, escalation rules	Business owner during deployment
Privacy overreach	Workforce monitoring	Fraud or credit system	Data minimization,	Privacy/ security teams through

Risk	HR instantiation	Financial instantiation	Mitigation control	Primary owner / timing
	collects excessive behavioral data	s use intrusive third-party signals	retention limits, access controls, privacy impact assessment	out lifecycle
Performance drift	Changing labor markets reduce selection validity	Macroeconomic shocks shift default, fraud or portfolio behavior	Drift metrics, periodic revalidation, challenge models, incident register	Model-risk committee post-deployment
Contestability failure	Candidate cannot challenge inaccurate data or accommodation failure	Applicant cannot understand or appeal automated denial	Notice, appeal workflow, evidence retention, remediation tracking	Legal/compliance and domain owner post-decision

6. Discussion

The empirical analysis indicates that ethical AI cannot be evaluated through predictive performance or procedural compliance alone. The German Credit benchmark shows that models with similar ROC-AUC values may still distribute errors and favorable outcomes unevenly across age groups, while the FICO HELOC benchmark demonstrates that transparent models can remain competitive with more complex ensembles in credit-risk prediction. At the same time, the HELOC case exposes a major audit limitation: when protected attributes are absent, fairness evaluation becomes incomplete. These findings support the need for an ethics lifecycle in which accuracy, calibration, fairness, explainability, governance evidence, and redress mechanisms are assessed jointly before deployment.

A central methodological implication is that fairness depends on data design. Fairness metrics are meaningful only when protected attributes, outcome labels, decision thresholds, and harm assumptions are valid for the target context. In HR, labels such as "high potential" or "successful employee" may reproduce prior managerial bias or unequal opportunity. In finance, repayment outcomes may reflect not only borrower behavior but also access to credit, servicing practices, product design, and macroeconomic exposure. Therefore, ethical AI audits must document label provenance and validity rather than treating benchmark labels as neutral ground truth.

The results also caution against equating interpretability with ethical sufficiency. Transparent models improve inspectability, communication, and contestability, but they do not automatically ensure unbiased labels, lawful feature use, representative training data, stable performance, or meaningful appeal. Conversely, complex models may achieve favorable fairness scores under selected thresholds while failing explanation, monitoring, or accountability requirements. The proposed FAIR-HITL framework therefore separates model performance, explanation quality, decision governance, and redress evidence as distinct but interdependent dimensions of responsible AI deployment.

For HR managers, AI systems should not be adopted solely on efficiency or predictive-validity claims. Deployment should require job-analysis evidence, subgroup adverse-impact testing, candidate notice, disability accommodation procedures, recruiter training, threshold documentation, and override logging. For financial institutions, fair-lending governance must be integrated with model-risk management. Credit, fraud, pricing, and risk models should be evaluated for discrimination, calibration, stability, reason-code accuracy, data provenance, and third-party dependency risk. Transparent baseline models should remain part of challenger-model testing, especially when complex models provide only marginal predictive gains.

For policymakers, the findings suggest that regulation should emphasize auditable evidence rather than broad declarations of trustworthiness. Requirements for subgroup testing, data governance, meaningful explanation, human oversight, monitoring, and appeal are more enforceable when organizations must maintain inspectableartifacts. Regulators should also clarify when protected attributes may be processed for fairness auditing under privacy and anti-discrimination law, since avoiding such data can make discrimination harder to detect. For AI developers and vendors, responsible model development should include datasheets, model cards, validation protocols, subgroup results, explanation-stability evidence, exportable audit logs, and clear limitations. Claims such as "bias-free" or "objective" should be replaced with measurable, context-specific evidence.

Implementation remains challenging because organizational incentives often prioritize efficiency over validation, vendor systems create information asymmetry, protected-attribute auditing may raise legal uncertainty, and human oversight can become symbolic under productivity pressure. Fairness metrics may also be selectively chosen to present favorable results. FAIR-HITL addresses these risks by linking ethical claims to auditable artifacts, decision rights, monitoring duties, and redress pathways, but its effectiveness ultimately depends on sustained institutional commitment.

7. Conclusion

Ethical AI in HR and financial decision-making requires more than responsible rhetoric. The study demonstrates a defensible research pathway: use traceable public evidence, separate conceptual synthesis from empirical claims, evaluate public benchmarks reproducibly, code governance instruments transparently and interpret results within legal and sociotechnical constraints. The German Credit and FICO HELOC analyses show that predictive performance, fairness, explainability and auditability can diverge, while source-derived HR and mortgage evidence shows that algorithmic and institutional disparities are concrete governance risks. The proposed FAIR-HITL framework translates these findings into a lifecycle model that organizations can operationalize through risk classification, data governance, fairness-aware modelling, explainable AI, human review, privacy safeguards, monitoring and redress. For Q1-level research and practice, the key standard is not the appearance of ethical compliance but the ability to produce reliable evidence that consequential AI systems are valid, contestable, monitored and accountable.

References

- [1] Stanford Institute for Human-Centered Artificial Intelligence. The AI Index Report 2026.Stanford University; 2026.
- [2] Society for Human Resource Management. 2025 Talent Trends: AI in HR. SHRM; 2025.
- [3] Society for Human Resource Management. State of AI in HR 2026: AI Use and Governance in Human Resources. SHRM; 2026.
- [4] Bank of England and Financial Conduct Authority. Artificial Intelligence in UK Financial Services: 2024. Bank of England/FCA; 2024.
- [5] OECD. Algorithmic Management in the Workplace.OECD Publishing; 2025.

- [6] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0).NIST AI 100-1; 2023.
- [7] Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. NIST Special Publication 1270; 2022.
- [8] European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union; 2024.
- [9] International Organization for Standardization. ISO/IEC 42001:2023 Information Technology - Artificial Intelligence - Management System.ISO/IEC; 2023.
- [10] Consumer Financial Protection Bureau. Circular 2022-03: Adverse Action Notification Requirements in Connection with Credit Decisions Based on Complex Algorithms. CFPB; 2022.
- [11] Consumer Financial Protection Bureau. Circular 2023-03: Adverse Action Notification Requirements and the Proper Use of the CFPB Sample Forms Provided in Regulation B. CFPB; 2023.
- [12] U.S. Equal Employment Opportunity Commission. Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures under Title VII of the Civil Rights Act of 1964. EEOC; 2023.
- [13] U.S. Equal Employment Opportunity Commission and U.S. Department of Justice. Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring.Technical Assistance Document; 2022.
- [14] Bommasani R, Bana SH, Creel KA, Jurafsky D, Liang P. Algorithmic monocultures in hiring. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency; 2026. doi:10.1145/3805689.3812400.
- [15] Federal Financial Institutions Examination Council. Home Mortgage Disclosure Act 2024 National Loan-Level Dataset.FFIEC; 2025.
- [16] Ky W, Lim K. The Role of Race in Mortgage Application Denials.Federal Reserve Bank of Minneapolis, Community Development Working Paper; 2023.
- [17] Ding F, Hardt M, Miller J, Schmidt L. Retiring Adult: New datasets for fair machine learning. Advances in Neural Information Processing Systems. 2021;34:6478-6490.
- [18] U.S. Census Bureau. American Community Survey Public Use Microdata Sample Documentation.U.S. Department of Commerce; 2024.
- [19] Grömping U. South German Credit Data: Correcting a Widely Used Data Set. Reports in Mathematics, Physics and Chemistry, Beuth University of Applied Sciences Berlin; 2019.
- [20] FICO. Explainable Machine Learning Challenge: Home Equity Line of Credit Dataset. Fair Isaac Corporation; 2018.
- [21] Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T. An interpretable model with globally consistent explanations for credit risk.arXiv:1811.12615; 2018.
- [22] Köchling A, Wehner MC. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research. 2020;13:795-848. doi:10.1007/s40685-020-00134-w.
- [23] Bankins S. The ethical use of artificial intelligence in human resource management: A decision-making framework. Ethics and Information Technology. 2021;23:841-854. doi:10.1007/s10676-021-09619-6.
- [24] Fabris A, et al. Fairness and bias in algorithmic hiring: A multidisciplinary survey. ACM Computing Surveys. 2025. doi:10.1145/3696457.
- [25] Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. 2020:469-481.
- [26] Li L, Lassiter T, Oh J, Lee MK. Algorithmic hiring in practice: Recruiter and HR professional perspectives on AI use in hiring. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2021:166-176.
- [27] Prikshat V, Malik A, Budhwar P. AI-augmented HRM: Antecedents, assimilation and multilevel consequences. Human Resource Management Review. 2023;33(1):100860.
- [28] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning.ACM Computing Surveys. 2021;54(6):1-35. doi:10.1145/3457607.
- [29] Weerts H, Dudik M, Edgar R, Jalali A, Lutz R, Madaio M. Fairlearn: Assessing and improving fairness of AI systems. Journal of Machine Learning Research. 2023;24(257):1-8.
- [30] Bellamy RKE, Dey K, Hind M, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development. 2019;63(4/5):4:1-4:15. doi:10.1147/JRD.2019.2942287.
- [31] Kozodoi N, Jacob J, Lessmann S. Fairness in credit scoring: Assessment, implementation and profit implications.

European Journal of Operational Research. 2022;297(3):1083-1094. doi:10.1016/j.ejor.2021.06.023.

- [32] Misheva BH, Osterrieder J, Hirsa A, Kulkarni O, Lin SF. Explainable AI in credit risk management. Frontiers in Artificial Intelligence. 2021;4:782989.
- [33] Gramegna A, Giudici P. SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence. 2021;4:752558. doi:10.3389/frai.2021.752558.
- [34] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier.Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:1135-1144.
- [35] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions.Advances in Neural Information Processing Systems. 2017;30:4765-4774.
- [36] Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019:220-229.
- [37] Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Communications of the ACM. 2021;64(12):86-92. doi:10.1145/3458723.
- [38] Raji ID, Smart A, White RN, et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. 2020:33-44.