FROM ALGORITHMIC OPACITY TO GENDER-RESPONSIVE ACCOUNTABILITY: AN EXPERT-VALIDATED G-XAI AUDIT FRAMEWORK FOR HIRING AND CREDIT DECISIONS

Erik Omar Escalona Aguilar[1], Natalia Conejero Riquelme [2]*

[1]*Email: erik.escalona@ubo.cl Universidad UNIACC, Santiago, Santiago Metropolitan, Chile.
ORCID: https://orcid.org/0000-0002-7735-6027
[2]Email: natalia.conejero@ubo.cl Universidad Bernardo O'Higgins, Santiago, Chile.
ORCID: https://orcid.org/0000-0003-2373-4907

Abstract
Artificial intelligence increasingly mediates access to employment and credit, while gendered patterns may remain embedded in historical labels, proxy variables, optimization objectives, and organizational routines. This study develops an expert-validated, stakeholder-oriented Gender-Responsive Explainable Artificial Intelligence (G-XAI) audit framework for hiring and credit decisions. A pilot PRISMA 2020 evidence synthesis identified 86 records, removed 12 duplicates, screened 74 records, assessed 36 full-text reports, and included 20 DOI-verified peer-reviewed studies. The resulting framework was assessed through structured expert validation informed by modified Delphi principles, using quantitative ratings, qualitative comments, and an independent methodological review. The evidence synthesis showed that fairness metrics alone are insufficient, while expert validation supported the framework's relevance and emphasized gender and intersectional analysis, explainability linked to accountability, human oversight, documentation, contestability, and remediation. Experts also identified implementation challenges and instrument-design issues requiring further refinement. Because validation records differed in analytical level, findings were synthesized descriptively rather than treated as formal Delphi consensus. Grounded in stakeholder theory and SDG 5, the framework provides an eight-stage lifecycle architecture for integrating fairness, explainability, governance, and gender-responsive accountability.

Keywords: Algorithmic auditing, explainable artificial intelligence, gender bias, gender-responsive accountability, stakeholder theory.

• Introduction

Artificial intelligence (AI) has moved from an analytical support function to an institutional gatekeeper. In recruitment, automated systems can source candidates, parse résumés, rank profiles, score video interviews, generate job advertisements, and recommend hiring decisions. In financial services, machine-learning models can determine creditworthiness, loan pricing, limits, and access to products. These applications are attractive because they promise speed, consistency, scale, and predictive accuracy. However, when past organizational decisions reflect occupational segregation, unequal care burdens, gendered income trajectories, or discriminatory lending practices, an algorithm may learn and normalize those patterns rather than correct them (Bartlett et al., 2022; Fabris et al., 2025; Fuster et al., 2022; Köchling & Wehner, 2020).

Gender discrimination is not generated only by an explicit variable labelled "sex" or "gender." It can enter through target definitions, selective labels, sampling frames, missing data, measurement practices, and proxies such as employment continuity, salary history, household composition, geography, device type, linguistic style, or occupational title. It can also arise after model deployment, when decision-makers over-rely on algorithmic recommendations or when women cannot understand, contest, or change the factors behind an adverse decision. Suresh and Guttag (2021) therefore describe machine-learning harm as a lifecycle phenomenon, while Mökander et al. (2021) conceptualize automated decision-making as a sociotechnical system rather than a model in isolation.

The central practical difficulty is that fairness is context-dependent and plural. Demographic parity, equal opportunity, equalized odds, predictive parity, calibration, individual fairness, and counterfactual fairness encode different moral and statistical commitments. They may produce incompatible conclusions when base rates differ, and none alone establishes substantive gender justice (Mehrabi et al., 2021; Kozodoi et al., 2022; Pagano et al., 2023). Explainable AI (XAI) creates a second difficulty. A feature-attribution chart may help a data scientist diagnose a model, but it may not give a rejected applicant a faithful, comprehensible, actionable, and legally appropriate explanation. The explanation needs of affected women, model developers, business managers, auditors, and regulators are not identical.

This article uses structured expert judgment informed by modified Delphi principles to address G-XAI decisions where technological change outpaces representative evidence and stable benchmarks. This approach

complements empirical evidence by identifying evidence gaps and supporting provisional safeguards that can be revised as stronger evidence becomes available.

Problem statement

End-to-end algorithmic auditing has been proposed as a mechanism for connecting model assessment with broader organizational accountability (Raji et al., 2020). The reviewed evidence does not provide an integrated, end-to-end audit architecture combining gender-disaggregated testing, meaningful explanations, stakeholder participation, mitigation, human oversight, contestability, and post-deployment monitoring across employment and credit. Existing research addresses these dimensions unevenly and rarely integrates them through gender-sensitive decision criteria. Consequently, narrow technical compliance may coexist with unequal opportunity, opaque procedures, or unrealistic corrective actions for women.

Research question and objectives

Primary research question. How can an expert-validated, stakeholder-oriented G-XAI audit framework define adequate evidence, identify and explain gender bias, and guide mitigation in AI systems used for hiring and credit decisions while preserving predictive validity, operational feasibility, and organizational accountability?

General objective. To design and expert-validate a lifecycle G-XAI audit framework for gender-responsive governance of algorithmic hiring and lending decisions, using a structured expert-validation process informed by modified Delphi principles and including a method for prioritizing data gaps as technologies and use cases evolve.

- Systematically synthesize recent evidence on gender bias, fairness, explainability, auditing, and mitigation in hiring and credit scoring.
- Operationalize stakeholder theory by mapping the rights, interests, knowledge, power, and responsibilities of women applicants and borrowers and of organizational and societal actors.
- Conduct a structured expert validation of the G-XAI framework, informed by modified Delphi principles, and transparently document the validation evidence, analytical approach, methodological limitations, and areas of expert agreement and disagreement.
- Define auditable indicators and mitigation options spanning data, performance, fairness, intersectionality, explainability, human oversight, recourse, and monitoring.
- Develop an adaptive mechanism for prioritizing evidence gaps and revising audit criteria, aligned with SDG 5 commitments to participation, financial access, enabling technology, and accountability.

Contributions

The article makes five contributions. First, it integrates stakeholder theory with algorithmic auditing and gender-responsive XAI. Second, it incorporates structured expert validation, informed by modified Delphi principles, to assess and refine the framework where fairness metrics, data availability, and stakeholder values are incomplete or contested. Third, it develops an expert-informed data-gap function that identifies evidence requirements for responsible deployment and prioritizes future data collection. Fourth, it specifies an eight-stage audit architecture with differentiated modules for hiring and credit. Fifth, it translates SDG 5 into measurable organizational controls related to equal opportunity, financial access, enabling technology, and enforceable accountability.

Theoretical foundation

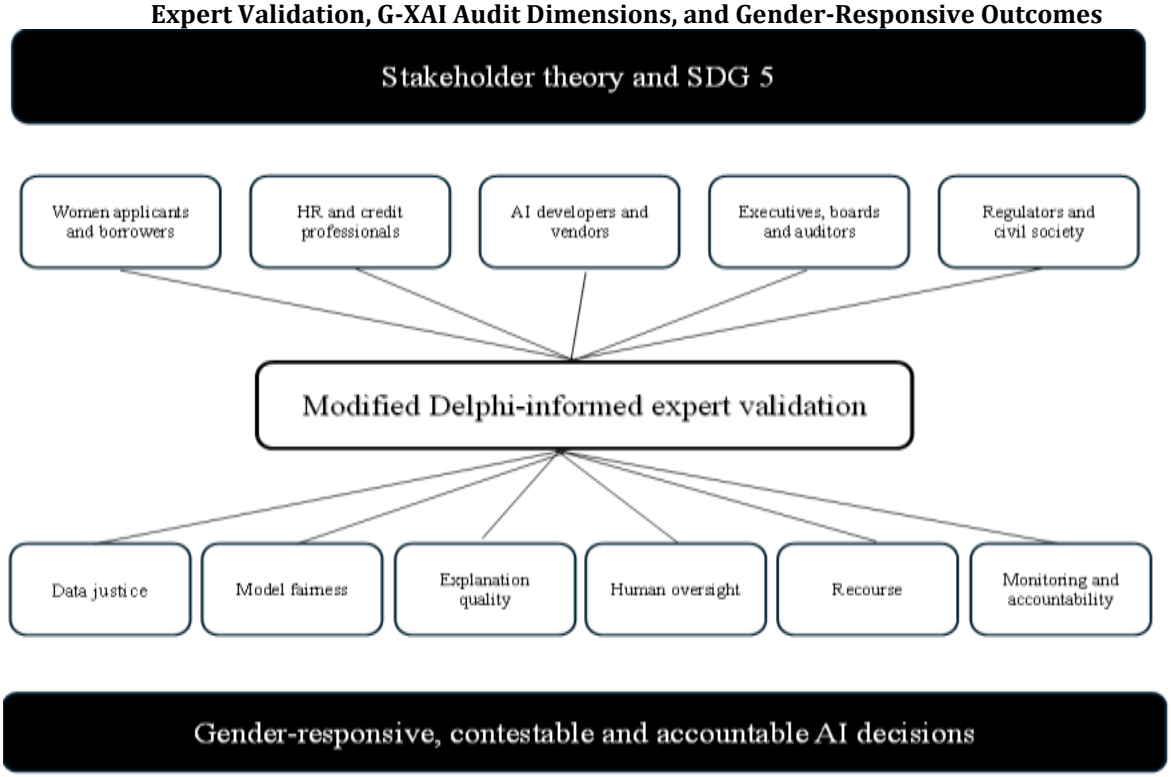
Stakeholder theory and algorithmic decisions

Freeman's (1984) stakeholder approach rejects the view that managerial responsibility is exhausted by shareholder value. Organizations create and distribute value through relationships with groups that can affect or are affected by corporate objectives. Donaldson and Preston (1995) distinguish descriptive, instrumental, and normative dimensions of stakeholder theory: it describes how organizations operate, links stakeholder management to organizational outcomes, and establishes intrinsic obligations toward stakeholders. More recent work emphasizes that stakeholder theory must confront tensions among interests rather than assume frictionless alignment (Freeman et al., 2020).

In algorithmic hiring and credit, women applicants and borrowers are primary stakeholders because decisions affect livelihood, career mobility, income, wealth accumulation, and economic security. Beyond being a protected attribute in fairness calculations, women possess situated knowledge of barriers such as unpaid care, career interruptions, informal income, thin credit files, occupational segregation, and prior exclusion. A legitimate audit must therefore allocate voice, information, and responsibility across affected women and the organizational and societal actors involved.

Stakeholder theory adds normative, instrumental, and descriptive dimensions to technical fairness by addressing stakeholders' rights and dignity, the organizational value of inclusive governance, and the workflows through which algorithmic decisions are implemented. These dimensions support incorporating the perspectives of women affected by hiring and credit decisions.

Figure 1. Conceptual Architecture Linking Stakeholder Theory, SDG 5, Modified Delphi-Informed Expert Validation, G-XAI Audit Dimensions, and Gender-Responsive Outcomes



Note. Authors' elaboration.

Gender-responsive explainable artificial intelligence

G-XAI is defined here as the design, evaluation, explanation, and governance of AI systems so that gendered and intersectional effects are visible, contestable, mitigated, and continuously monitored. The concept extends conventional XAI by asking not only whether a model can be interpreted, but also whether an explanation is faithful, understandable, actionable, non-discriminatory, and connected to an effective review and remedy process.

Local Interpretable Model-agnostic Explanations (LIME)

LIME explains an individual prediction by fitting an interpretable surrogate to perturbed observations around the case of interest (Ribeiro et al., 2016). As a model-agnostic, local method, it can provide intuitive explanations for applicants or borrowers but should not be interpreted as describing the entire model. For auditing, LIME requires documented perturbation settings and tests of local fidelity, stability, and the realism of synthetic cases.

SHapley Additive exPlanations (SHAP)

SHAP assigns feature contributions to the difference between an individual prediction and the expected model output under a reference population, supporting both local explanations and aggregated patterns (Lundberg et al., 2020). In gender audits, SHAP can reveal differences in feature contributions across women and men and identify potential proxy variables or divergent prediction pathways. Because results depend on the background dataset and feature-dependence assumptions, auditors should document these choices and assess sensitivity and subgroup stability. SHAP values indicate model influence rather than causality, legitimacy, or legal justification.

Audit implications

LIME and SHAP are diagnostic methods rather than certificates of fairness. G-XAI therefore triangulates explanation evidence with subgroup performance, counterfactual tests, data provenance, causal reasoning, human decision logs, and affected-person review. Material disagreement among methods requires further investigation.

SDG 5 as an impact and accountability framework

SDG 5 provides a substantive orientation for the audit. Targets 5.5 and 5.a connect algorithmic hiring and credit decisions to women's equal participation, leadership opportunities, economic resources, and financial services. Targets 5.b and 5.c further support enabling technologies, documented governance, appeal mechanisms, and corrective action (United Nations, n.d.). SDG 5 therefore functions not as symbolic context but as a basis for impact questions and organizational indicators.

Methodological role of modified Delphi principles

AI systems can change faster than conventional evidence cycles, while relevant evidence may remain limited, unrepresentative, or historically biased, particularly for women and intersectional groups. Governance must therefore consider not only data volume but also relevance, representativeness, timeliness, provenance, and normative adequacy.

Modified Delphi principles provide a structured approach to expert judgment when evidence is incomplete or distributed. Features such as anonymity, iteration, controlled feedback, explicit consensus criteria, and preservation of dissent can reduce status effects, clarify disagreement, and improve traceability. In this study, these principles informed the expert-validation design; the available evidence does not support treating it as a completed multi-round Delphi process.

For G-XAI, structured expert judgment can identify evidence gaps, prioritize data needs, and support provisional decision criteria when empirical evidence is incomplete. These criteria can be reassessed as models, data, legal obligations, affected populations, or empirical evidence change.

Delphi-informed validation complements rather than replaces empirical validation. Because formal consensus and stability criteria were not applied in this study, expert judgments should inform framework refinement without being interpreted as statistical consensus. Formal Delphi studies require predefined validation units, rating dimensions, consensus criteria, and systematic documentation of agreement and disagreement (Beiderbeck et al., 2021; Nasa et al., 2021; Gattrell et al., 2024).

- **Materials and methods**

Design

The study combines a systematic evidence synthesis, structured expert validation informed by modified Delphi principles, and a design-science output: the G-XAI audit framework. The review follows PRISMA 2020 principles (Page et al., 2021a, 2021b).

Review question and eligibility criteria

Table 1. PICOS/SPIDER-Informed Eligibility Framework

Dimension	Operational criterion
Population / context	Women and intersectional groups affected by AI-supported recruitment, hiring, credit scoring, lending, or directly transferable high-stakes audit methods.
Intervention / phenomenon	Algorithmic decision systems; fairness assessment; XAI; algorithmic audit; bias mitigation. Delphi or structured consensus methods were considered through the secondary methodological search.
Comparators	Alternative models, human decisions, hybrid decision processes, fairness interventions, or baseline governance practices.
Outcomes	Gender-disaggregated error and selection rates, fairness metrics, explanation quality, discriminatory effects, governance quality, recourse, stakeholder legitimacy.
Study designs	Peer-reviewed empirical studies, systematic reviews, multidisciplinary surveys, methodological frameworks, and consensus-method guidance.
Time/language	1 January 2021–29 July 2026; English or Spanish. Seminal earlier studies retained only for theoretical or methodological foundations.
Exclusions	Pure accuracy studies without fairness/explainability relevance; clinical-only evidence without transferable audit insight; non-peer-reviewed commentary; duplicate preprints when a final article exists; papers without sufficient methods.

Information sources and search strategy

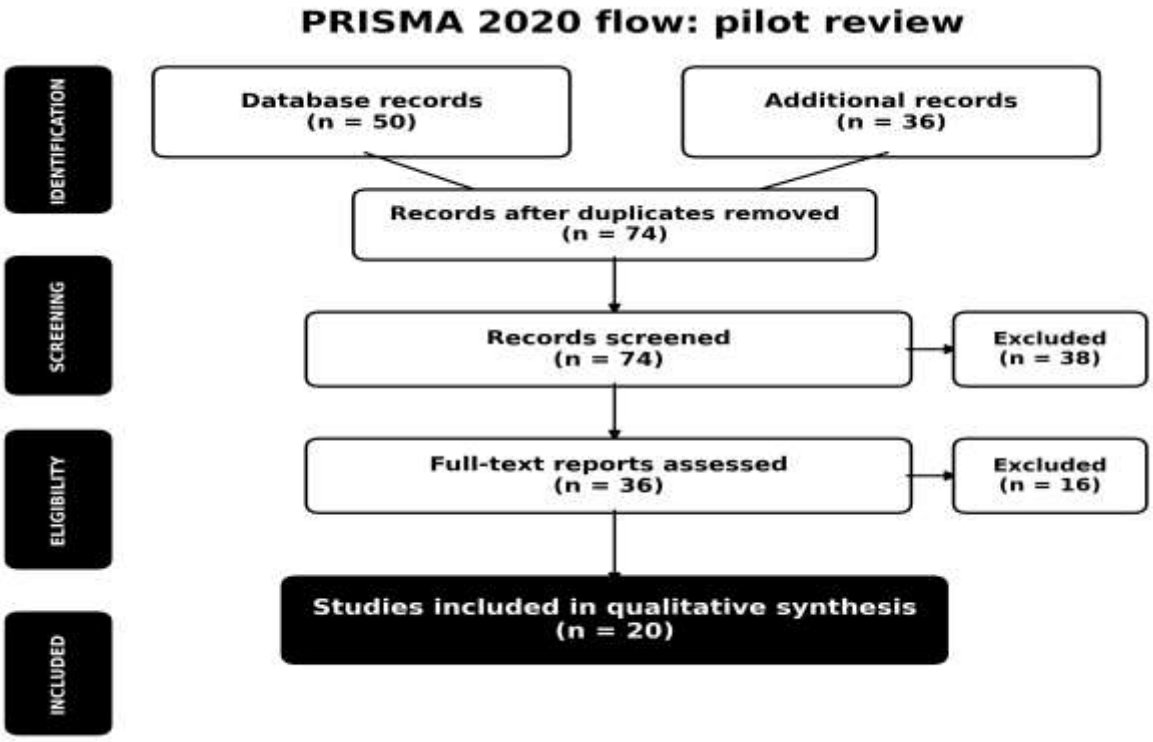
Web of Science Core Collection and Scopus were searched on 29 July 2026 using the reproducible strings in Appendix A, supplemented by citation searching and targeted complementary searches. Of 86 records identified (50 database and 36 additional records), 12 duplicates were removed, 74 records were screened, 36 full texts were assessed, and 20 DOI-verified peer-reviewed studies were included.

Search terms combined explainability/fairness/audit, gender/intersectionality, and hiring/credit concepts; a secondary block addressed structured consensus in AI governance and audit. Exact strings appear in Appendix A.

Screening, extraction, and quality appraisal

Titles and abstracts were screened against predefined eligibility criteria, followed by full-text assessment. Extraction covered study context, data and models, protected and intersectional groups, outcomes, fairness and XAI methods, mitigation, governance, findings, limitations, and DOI. A 0–2 rubric assessed methodological relevance and quality, which informed the interpretive weight of evidence rather than study exclusion. Independent dual screening, formal resolution of reviewer disagreement, and inter-reviewer agreement were not documented for the pilot review.

Figure 2. PRISMA 2020 Flow Diagram for the Pilot Evidence Synthesis



Note. Authors' elaboration following Page et al. (2021a, 2021b).

Modified Delphi-informed expert validation

As summarized in Table 2, the heterogeneous validation records were analyzed at their documented level, with narrative observations treated as complementary qualitative evidence.

Table 2. Modified Delphi-Informed Expert-Validation Design

Element	Implemented validation design
Validation purpose	To assess the relevance, conceptual coherence, applicability, methodological robustness, and areas for refinement of the proposed G-XAI audit framework for hiring and credit decisions.
Validation structure	Structured expert validation informed by modified Delphi principles, combining quantitative ratings, qualitative comments, proposed revisions, and an independent methodological assessment.
Framework evaluated	13 criteria and 39 subcriteria covering gender bias and proxy variables, framework architecture, theoretical and normative foundations, organizational applicability, accountability, fairness, explainability, governance, mitigation, intersectionality, communication, transferability, and methodological validity.
Decision domains	Recruitment/hiring and credit decisions affecting women.
Detailed unit of assessment	The detailed instrument operationalized the 39 subcriteria separately for the two decision domains, yielding 78 case-specific items when completed in full.

Validation records	Five individual spreadsheet-based validation/review records (E1–E5), one independent methodological review (M1), and four additional narrative evidence blocks (N1–N4) were consolidated. Narrative records were not counted as additional unique experts because possible overlap with other validators could not be reliably established.
Quantitative evidence	Available records include item-level ratings, criterion-level ratings, or both, depending on the validation record. Because the records operate at different analytical levels, quantitative findings were synthesized descriptively rather than pooled into a single aggregate statistic.
Qualitative evidence	Expert comments, criterion justifications, methodological observations, and proposed item revisions were examined to identify recurrent strengths, implementation concerns, conceptual ambiguities, and instrument-design issues.
Methodological review	An independent methodological review assessed the unit of validation, evaluation dimensions, methods for assessing agreement and consensus, panel traceability, instrument design, and documentation consistency.
Interpretation of agreement	Favorable expert ratings were interpreted as evidence of expert support for the framework. They were not equated with formal Delphi consensus because the available records did not provide a uniform basis for assessing median, IQR, percentage agreement, or stability across multiple completed rounds.
Use of expert feedback	Quantitative and qualitative evidence was used to identify supported components of the framework and areas requiring clarification, revision, or further empirical validation.

- **Results of the evidence synthesis**

Characteristics of the evidence base

The 20-record DOI-verified corpus provides strongest evidence on statistical fairness metrics, lifecycle sources of harm, and general auditing principles, but comparatively limited evidence on gender-responsive explanation quality, affected-woman participation, intersectional sample adequacy, post-deployment remedy, and integrated hiring-credit governance.

Table 3. DOI-Verified Records Included in the Qualitative Synthesis (2021–2026)

Study	Domain	Design	Audit-relevant finding	DOI
Mehrabi et al. (2021)	General ML	Survey	Taxonomy of bias sources and fairness definitions; no metric is universally sufficient.	10.1145/3457607
Suresh and Guttag (2021)	General ML	Lifecycle framework	Harm can arise during data generation, model building, evaluation, and deployment.	10.1145/3465416.3483305
Mökander and Floridi (2021)	AI governance	Conceptual	Auditing can bridge principles and practice but must be continuous and system-oriented.	10.1007/s11023-021-09557-8
Mökander et al. (2021)	AI governance	Conceptual framework	Defines ethics-based auditing and its technical, organizational, and institutional limits.	10.1007/s11948-021-00319-4
Beiderbeck et al. (2021)	Delphi	Methods guidance	Specifies preparation, execution, feedback, and analysis practices for rigorous Delphi studies.	10.1016/j.mex.2021.101401
Nasa et al. (2021)	Delphi	Methods guidance	Clarifies when Delphi is appropriate and how expert consensus should be designed.	10.5662/wjm.v11.i4.116

Bartlett et al. (2022)	Consumer lending	Econometric audit	FinTech can reduce some pricing disparities while algorithmic discrimination remains detectable.	10.1016/j.jfineco.2021.05.047
Fuster et al. (2022)	Mortgage credit	Empirical/econometric	Machine learning changes approval and pricing patterns and can redistribute benefits unevenly.	10.1111/jofi.13090
Kozodoi et al. (2022)	Credit scoring	Fairness-profit analysis	Fairness interventions have different operational and profit implications; context matters.	10.1016/j.ejor.2021.06.023
Hofeditz et al. (2022)	Hiring	Online experiment	AI recommendations and XAI can alter discrimination, but effects differ across protected groups.	10.1007/s12525-022-00600-9
Pagano et al. (2023)	General ML	Systematic review	Metrics and mitigation methods are diverse and can yield inconsistent assessments.	10.3390/bdcc7010015
Frissen et al. (2023)	Recruitment	ML/NLP study	Detects biased and discriminatory language in job advertisements.	10.1007/s00146-022-01574-0
Garcia et al. (2024)	Credit	Systematic/open forum synthesis	Maps discrimination mechanisms and evidence gaps throughout credit decision systems.	10.1007/s00146-023-01676-3
Armstrong et al. (2024)	Hiring/LLMs	Algorithmic audit	GPT-generated and scored résumés exhibit race- and gender-linked stereotypes.	10.1145/3689904.3694699
Gattrell et al. (2024)	Consensus methods	Reporting guideline	Requires transparent consensus definitions, panel information, rounds, feedback, and attrition.	10.1371/journal.pmed.1004326
Fabris et al. (2025)	Algorithmic hiring	Multidisciplinary survey	Integrates technical, legal, organizational, dataset, and mitigation perspectives.	10.1145/3696457
de Castro Vieira et al. (2025)	Credit decisions	Systematic review	Synthesizes pre-, in-, and post-processing mitigation in credit decisions.	10.3390/jrfm18050228
Mathen (2025)	Credit scoring	Responsible-AI framework	Proposes an evolving governance framework for responsible bank credit scoring.	10.1108/JICES-08-2024-0122
Sivakaminathan and Musi (2026)	HR/LLMs	NLP audit	Finds gendered linguistic patterns and proposes analysis of language roots of bias.	10.1007/s00146-025-02564-8
Bahlool et al. (2026)	AI credit scoring	Systematic review	Finds performance, fairness, and explainability are usually treated in isolation.	10.3390/jrfm19020104

Gender bias in algorithmic hiring

Hiring systems can reproduce gender bias through historical labels, feature construction, gendered language and representation, and automation bias. Models may learn male-dominated occupational patterns or treat employment gaps, career continuity, prior salary, job titles, and institutional prestige as indirect markers of gendered opportunity. Language-based systems may also reproduce gender stereotypes (Armstrong et al.,

2024; Frissen et al., 2023; Sivakaminathan & Musi, 2026), while decision-makers may treat poorly validated scores as objective.

XAI does not necessarily reduce discrimination: explanations can affect outcomes across demographic groups and may make flawed recommendations more persuasive (Hofeditz et al., 2022). Audits should therefore assess model outputs and the effects of explanations on human selection behavior, overrides, confidence, and responsibility.

Gender fairness should also be assessed across the hiring funnel—from exposure and application to screening, interview, offer, salary, and retention—because aggregate rates can conceal stage-specific exclusion.

Gender bias in credit and lending

Credit scoring presents distinct gender risks. Women may have thinner formal credit histories, lower recorded income, interrupted employment, joint accounts, or greater participation in informal entrepreneurship, reflecting structural inequality rather than repayment risk. Alternative data may expand access while introducing opaque proxies from location, social networks, devices, or consumption. Machine learning can alter approvals and prices without eliminating group disparities (Bartlett et al., 2022; Fuster et al., 2022), requiring discrimination to be assessed across data, models, institutional incentives, and regulation (Garcia et al., 2024). Credit fairness must also consider risk calibration and business viability (de Castro Vieira et al., 2025; Kozodoi et al., 2022). A gender-responsive audit therefore examines approval rates, pricing, credit limits, prediction errors, calibration, and repayment outcomes by gender and intersectional group, while assessing whether rejected women receive meaningful reasons and attainable pathways to reassessment.

Cross-cutting gap: fairness, explainability, and governance are evaluated separately

Predictive performance, fairness, and explainability are often evaluated separately, creating a governance blind spot: accurate models may under-approve creditworthy women, parity-adjusted models may be poorly calibrated, and explanations may rely on unjust targets or infeasible changes (Bahlool et al., 2026; Fabris et al., 2025). The G-XAI framework therefore requires a combined evidence package and stakeholder decision gate rather than a single technical pass/fail metric.

Results of the expert validation

The validation records described in Table 2 provided item- and/or criterion-level ratings, complemented by methodological and narrative evidence. Ratings were generally favorable. Item-level means were 3.95 (range 3–5) for E1, 4.24 (3–5) for E2, and 4.71 (4–5) for E3. Criterion-level means were 4.15 for E1, 4.77 for E3, 4.69 for E4, and 4.46 for E5, with all ratings ranging from 4 to 5.

These descriptive findings do not establish formal Delphi consensus because records differed in analytical level and predefined consensus and stability criteria were not applied. The methodological review recommended distinguishing relevance, clarity, and applicability and, in a subsequent formal Delphi study, assessing median, interquartile range, favorable-rating percentages, and stability across rounds.

Qualitative feedback supported the framework's shift from technical explainability toward organizational accountability, particularly traceability, human oversight, contestability, repair, mitigation, gender and intersectional analysis, and organizational learning. Experts also identified implementation challenges involving resources, organizational maturity, differentiated expectations for smaller organizations, progression thresholds, and fairness–performance trade-offs.

Experts cautioned that SHAP and LIME support attribution and diagnosis but do not establish causal discrimination. They supported distinguishing explanation, observed disparity, plausible mechanisms, causal analysis, and normative or legal assessment. Fairness metrics likewise require contextual justification rather than checklist use.

The validation also identified instrument-design issues, including items combining multiple constructs. Recommended refinements included separating data quality from explainability and security, categories of harm, error rates from calibration, and system from incident documentation. One credit-related item was reformulated to remove leading wording. These changes inform subsequent versions of the validation instrument.

Proposed expert-validated G-XAI audit framework

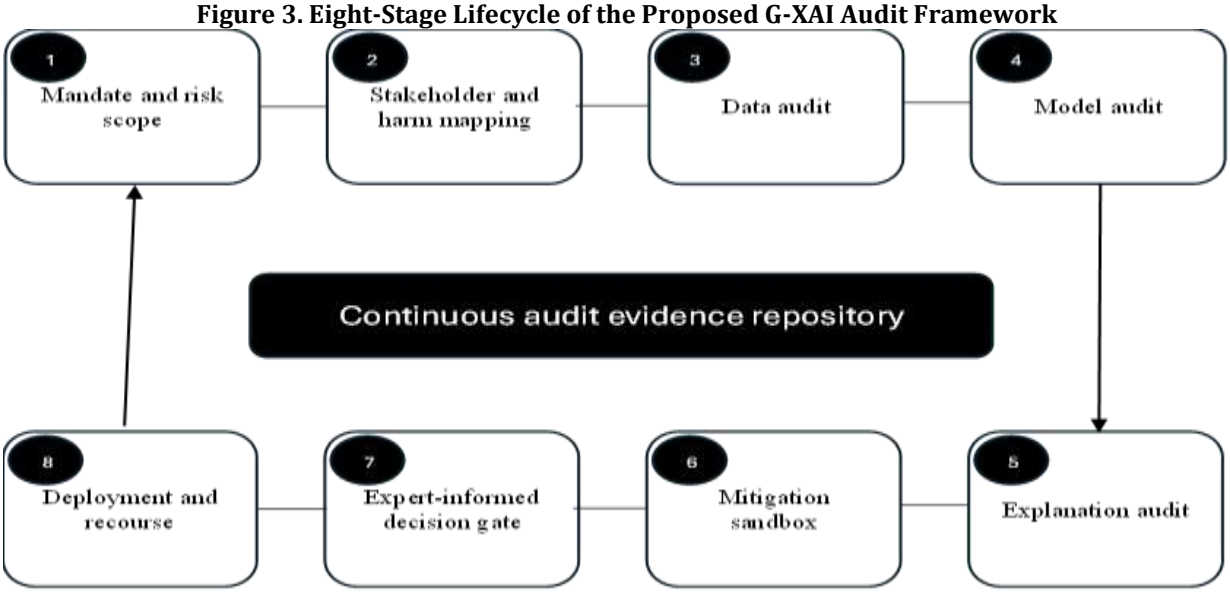
Design principles

Table 4. Normative and Operational Principles of the G-XAI Audit

Principle	Operational meaning
Rights before optimization	Protected groups are stakeholders with legitimate claims to equal treatment, dignity, explanation, and remedy.
Lifecycle scope	Audit the problem definition, data, model, interface, human use, governance, and downstream outcomes.

Intersectionality	Report gender jointly with age, disability, race/ethnicity where lawful, migration, geography, socioeconomic status, and caregiving proxies.
Multiple metrics	Use a justified metric portfolio; document incompatibilities and normative choices.
Explanation as evidence	Evaluate fidelity, stability, comprehensibility, actionability, and feasibility—not visual appeal alone.
Participatory assessment	Include women affected by the system and independent stakeholders in criteria, thresholds, and remedy design.
Contestability	Provide accessible reasons, human review, correction of data, appeal, and documented resolution.
Continuous accountability	Monitor drift, outcomes, complaints, overrides, and remediation after deployment.

Figure 3. Eight-Stage Lifecycle of the Proposed G-XAI Audit Framework



Note. Authors' elaboration.

Stage 1: mandate, purpose, and risk scope

An accountable executive approves the system's purpose, affected population, decision authority, prohibited uses, risk classification, responsibilities, and audit frequency. The audit asks whether AI is necessary and the target legitimate: similarity to prior successful employees may encode historical homogeneity, while poorly supported credit targets may exclude women with thin files. Accountable business, model, data, compliance, and assurance owners are assigned.

Stage 2: stakeholder and gendered-harm mapping

The organization maps stakeholders and gendered-harm scenarios across the decision process. In hiring, harms include reduced exposure, résumé scores, interviews or offers, biased features, and opaque rejection; in credit, lower approval or limits, higher prices, informal-income misclassification, proxy discrimination, and unactionable reasons. Each scenario is linked to severity, likelihood, and a responsible owner.

Stage 3: data audit

The data audit examines provenance, lawful basis, representativeness, missingness, measurement and label validity, temporal coverage, sample selection, and proxies, with distributions assessed by gender and intersectional subgroup. Historical hiring labels may reflect recruiter preferences, while credit labels may reflect prior loan conditions and access. Proxy testing combines statistical association, predictability of gender from non-protected variables, and causal/domain review. Removing the gender variable is insufficient and may prevent detection of discrimination.

Table 5. Minimum Quantitative Metric Portfolio: Metric Selection Must Be Justified Through Expert-Informed Decision Criteria and Domain Context

Metric family	Examples	Audit interpretation

Predictive validity	AUC/ROC or PR-AUC, sensitivity, specificity, precision, recall, Brier score, calibration error	Report overall and by gender/intersection; confidence intervals and sample sizes required.
Selection/approval parity	Selection or approval rate; demographic parity difference; disparate-impact ratio	Useful as an outcome screen; not a universal definition of fairness or a legal safe harbor.
Equal opportunity	True-positive-rate difference	In hiring, qualified women selected; in credit, creditworthy women approved.
Equalized odds	TPR and FPR differences	Captures asymmetric errors but may require trade-offs with calibration.
Predictive parity	Positive predictive value by group	Tests reliability of favorable decisions; can conflict with equalized odds.
Calibration	Observed outcome conditional on predicted risk by group	Critical in credit pricing and risk grades; use reliability curves and ECE/Brier score.
Individual consistency	Similarity-based outcome consistency	Requires defensible similarity metric; can reproduce biased feature spaces.
Counterfactual fairness	Outcome invariance under plausible protected-attribute intervention	Requires causal assumptions and careful treatment of descendants/proxies.
Intersectional worst-group risk	Maximum error, rejection, or price disparity across intersections	Prevents aggregate gender parity from masking harm to specific women.
Operational harm	Appeals, overrides, complaints, time to remedy, adverse price or wage effects	Connects model metrics to lived and organizational outcomes.

Stage 4: model and outcome audit

Metrics are reported with uncertainty, effect sizes, subgroup sample adequacy, and threshold sensitivity. Small intersectional samples require documented uncertainty and appropriate data, estimation, or precautionary responses. A pass requires predictive validity and a defensible fairness profile; aggregate performance cannot compensate for severe worst-group harm.

Stage 5: explanation audit

The explanation audit combines global, local, and counterfactual evidence appropriate to the model, stakeholder, decision, and risk level. Table 6 specifies the required quality tests for affected persons, professionals, and assurance actors. For high-impact decisions, explanation findings are triangulated with counterfactual, subgroup, benchmark, and domain evidence. Explanations that are unstable, misleading, discriminatory, or disconnected from feasible review and remedy are insufficient.

Table 6. Explanation-Quality Audit for Affected Persons, Professionals, and Assurance Actors

Quality	Question	Evidence
Fidelity	Does the explanation accurately represent the model's actual behavior?	For LIME, report local surrogate fit and black-box agreement. For SHAP, verify output reconstruction, explainer type, and background dataset. Compare with exact or transparent benchmarks where possible.
Stability	Do small irrelevant input changes produce similar explanations?	Repeat LIME across seeds and neighborhoods; repeat SHAP across plausible background samples; compare rank, sign, and magnitude consistency.
Comprehensibility	Can the intended stakeholder understand the reason?	User testing with women applicants/borrowers; plain-language comprehension.
Actionability	Does it identify changes that could alter the outcome?	Counterfactual validity and proximity; distinguish mutable from immutable features.
Feasibility	Are suggested changes realistic under social and economic constraints?	Domain and affected-person review; causal and temporal constraints.

Non-discrimination	Do explanation patterns differ unjustifiably by gender?	Compare LIME and SHAP patterns, counterfactual costs, reason codes, and recourse availability across gender and intersectional groups.
Privacy and security	Does the explanation reveal sensitive data or enable gaming?	Disclosure minimization, access controls, red-team tests.
Procedural usefulness	Does it support review, appeal, and correction?	Trace reason codes to evidence, human decision, and remediation path.

Table 7. Bias Mitigation Methods: Strengths and Cautions

Layer	Methods	Strengths and cautions
Pre-processing	Reweighing; stratified sampling; targeted data collection; label review; missing-data correction; proxy analysis; representation learning; synthetic augmentation with validation.	Acts before model fitting; can improve coverage but may distort data or conceal structural causes.
In-processing	Fairness constraints; regularization; adversarial debiasing; cost-sensitive learning; interpretable models; monotonic constraints; multi-objective optimization.	Makes fairness part of training; requires explicit metric choice and validation of performance trade-offs.
Post-processing	Threshold optimization; equalized-odds adjustment; reject-option classification; calibrated group-aware review; abstention.	Can change outcomes without retraining; may create legal or operational concerns and must be transparent.
Causal and counterfactual	Causal graphs; path-specific effects; counterfactual fairness; recourse constraints; proxy/descendant analysis.	Addresses mechanisms rather than correlation but depends on contestable causal assumptions.
Explainability-led diagnosis	LIME local-surrogate diagnostics; SHAP attribution analysis by subgroup; concept analysis; counterfactual cost comparison; explanation drift.	Locates potential mechanisms and proxy pathways; post-hoc explanations alone do not prove fairness and must be tested for fidelity and stability.
Organizational controls	Structured rubrics; human-in-command review; diverse teams; escalation; independent audit; vendor clauses; documentation; appeal; training.	Treats bias as sociotechnical; human review must itself be measured for bias and automation dependence.
Monitoring and remedy	Fairness dashboards; drift triggers; complaint analytics; periodic revalidation; rollback; compensation/correction.	Ensures accountability after deployment; requires clear thresholds and owners.

Stage 6: bias mitigation sandbox

Mitigation candidates are evaluated against utility, fairness, explainability, subgroup effects, operational costs, and unintended harms. Selection is based on empirically justified, predefined expert-informed decision criteria rather than the highest parity score alone. Residual risk is documented and assigned to an accountable owner.

Stage 7: expert-informed decision gate

The decision group determines whether evidence supports approval, conditional approval, remediation and retesting, suspension, or prohibition. Minimum evidence is required, and material dissent is documented, particularly when affected women disagree with technical or managerial participants. Hard stops include discriminatory targets, unavailable data provenance, severe unexplained intersectional error, unfaithful explanations, inadequate human review or appeal, insufficient vendor evidence, or unresolved high-severity harm. Conditional approval requires time-bound remediation, accountability, monitoring, and independent verification.

Stage 8: deployment, monitoring, and recourse

Post-deployment monitoring tracks performance, fairness and explanation drift, overrides, appeals, complaints, and relevant downstream outcomes. Material deterioration triggers investigation and potentially rollback. Affected persons receive accessible reasons, data-correction rights, human review, and timely resolution, while audit evidence is retained for independent assurance.

Domain-specific audit modules

Recruitment and hiring module

- Audit the recruitment funnel for gender disparities in exposure, ranking, interviews, offers, compensation, and retention.
- Test gendered language, résumé gaps, salary history, audiovisual features, accessibility, and other potential proxies using matched profiles and subgroup metrics.
- Compare algorithmic, human, and hybrid workflows for automation bias and overrides, and require documented job-relevant criteria, candidate notice, and appeal.

The audit tests whether apparently neutral features reproduce an historically masculine career prototype or structural inequalities in caregiving and occupational segregation, and whether recommendation or generative systems route women toward lower-paid or support-oriented roles.

Credit and lending module

- Audit access, underwriting, pricing, limits, servicing, and renewal by gender and intersection, including calibration and error outcomes.
- Test thin files, informal or part-time income, caregiving interruptions, joint accounts, and alternative-data proxies through subgroup, matched-applicant, and counterfactual analyses.
- Assess recourse costs for women versus men and require specific adverse-action reasons, data correction, human reassessment, and transparent reapplication.

Credit fairness covers both access and contract terms: equal approval rates may coexist with smaller limits or higher prices. The audit also tests whether historical credit policies created training labels that convert past exclusion into apparent future risk.

• Discussion

Theoretical contribution

The framework extends stakeholder theory into algorithmic governance by treating affected women as epistemic stakeholders whose lived experience informs harm and recourse, translating stakeholder legitimacy into auditable rights to information, explanation, participation, and remedy, and assigning differentiated responsibilities across developers, business owners, boards, vendors, auditors, regulators, and civil society.

Structured expert validation links these normative claims to technical audit design and evidence planning. Informed by modified Delphi principles, it makes convergent and divergent judgments and unresolved trade-offs traceable under incomplete evidence without reducing fairness to majority preference.

Methodological contribution

The audit rejects fairness through unawareness, single-metric fairness, and single-visualization explainability. Instead, it integrates performance, group and individual fairness, explainability diagnostics, counterfactual recourse, human behavior, organizational remedy, and an expert-informed data-gap function that treats missing evidence as a governable risk.

Expert validation particularly supported gender and intersectional analysis, human oversight, documentation, contestability, and remediation, while identifying resource, organizational-maturity, threshold, and performance trade-offs. The framework should therefore be treated as an evolving governance architecture requiring further empirical and context-specific testing.

Managerial and policy implications

For executives and boards, the framework translates responsible AI into ownership, evidence, decision gates, monitoring, and procurement requirements for auditability. Internal audit should test models and controls, while risk committees address unresolved dissent and residual gender risk.

For regulators and standard setters, documentation and human oversight must be operational rather than ceremonial. Meaningful oversight requires authority and evidence, while effective contestability requires understandable reasons and accessible mechanisms to challenge and correct decisions.

Contribution to SDG 5

The framework translates the SDG 5 commitments identified above into auditable indicators across hiring, credit access, technology governance, accountability, and remedy, linking AI governance to women's economic participation.

Limitations and future research

This study combines a pilot systematic evidence synthesis with structured expert validation informed by modified Delphi principles. Validation records differed in analytical level and may overlap between individual and narrative evidence; findings are therefore descriptive and do not establish formal Delphi consensus. The pilot review also lacked documented dual screening, formal disagreement resolution, and inter-reviewer agreement. Future studies should reproduce and archive the searches, incorporate these safeguards,

prospectively register systematic-review and formal Delphi protocols, and empirically test the proposed thresholds.

Fairness assessment also depends on data availability and legal context; sensitive attributes can support discrimination testing while creating privacy and governance risks. Moreover, the framework has not been evaluated with real-world organizational hiring or credit data or operational deployment. Future research should test its feasibility, audit reliability, and decision effects. Finally, post-hoc explanations do not establish causality and remain method-sensitive; research should compare interpretable and black-box models, examine explanation effects on human behavior, and test whether counterfactual recourse is causally valid and equally attainable.

- **Conclusion**

Gender-responsive AI cannot be achieved by deleting gender, selecting a single fairness metric, or attaching an explanation to an adverse decision. Hiring and credit algorithms operate within organizational histories and social structures that shape data, objectives, decisions, and remedies.

The G-XAI audit addresses these conditions through an eight-stage lifecycle and expert-informed decision process grounded in stakeholder theory. Its central contribution is to integrate technical evaluation with organizational accountability and the perspectives of women affected by hiring and lending systems in defining harms, evidence requirements, mitigation, and recourse. By making disagreement visible, assigning responsibility, and connecting model evaluation to SDG 5, the framework provides a researchable path from algorithmic opacity to gender-responsive accountability.

References

- Armstrong, L., Liu, A., MacNeil, S., & Metaxa, D. (2024). The silicon ceiling: Auditing GPT's race and gender biases in hiring. Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–18. https://doi.org/10.1145/3689904.3694699
- Bahlool, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. Journal of Risk and Financial Management, 19(2), 104. https://doi.org/10.3390/jrfm19020104
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047
- Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021). Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. MethodsX, 8, 101401. https://doi.org/10.1016/j.mex.2021.101401
- de Castro Vieira, J. R., Barboza, F., Cajueiro, D., & Kimura, H. (2025). Towards fair AI: Mitigating bias in credit decisions—A systematic literature review. Journal of Risk and Financial Management, 18(5), 228. https://doi.org/10.3390/jrfm18050228
- Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. Academy of Management Review, 20(1), 65–91. https://doi.org/10.5465/amr.1995.9503271992
- Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F., & Biega, A. J. (2025). Fairness and bias in algorithmic hiring: A multidisciplinary survey. ACM Transactions on Intelligent Systems and Technology, 16(1), 1–54. https://doi.org/10.1145/3696457
- Freeman, R. E. (1984). Strategic management: A stakeholder approach. Pitman.
- Freeman, R. E., Phillips, R., & Sisodia, R. (2020). Tensions in stakeholder theory. Business & Society, 59(2), 213–231. https://doi.org/10.1177/0007650318773750
- Frissen, R., Adebayo, K. J., & Nanda, R. (2023). A machine learning approach to recognize bias and discrimination in job advertisements. AI & Society, 38, 1025–1038. https://doi.org/10.1007/s00146-022-01574-0
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/jofi.13090
- Garcia, A. C. B., Garcia, M. G. P., & Rigobon, R. (2024). Algorithmic discrimination in the credit domain: What do we know about it? AI & Society, 39, 2059–2098. https://doi.org/10.1007/s00146-023-01676-3
- Gattrell, W. T., Logullo, P., van Zuuren, E. J., Price, A., Hughes, E. L., Blazey, P., Winchester, C. C., Tovey, D., Goldman, K., Hungin, A. P., & Harrison, N. (2024). ACCORD (ACcurate COnsensus Reporting Document): A reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLOS Medicine, 21(1), e1004326. https://doi.org/10.1371/journal.pmed.1004326
- Hofeditz, L., Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S. (2022). Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. Electronic Markets, 32, 2207–2233. https://doi.org/10.1007/s12525-022-00600-9

- Köchling, A., & Wehner, M. C. (2020). 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, 13, 795–848. https://doi.org/10.1007/s40685-020-00134-w
- Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083–1094. https://doi.org/10.1016/j.ejor.2021.06.023
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9
- Mathen, M. P. (2025). Toward an evolving framework for responsible AI for credit scoring in the banking industry. Journal of Information, Communication and Ethics in Society, 23(1), 148–163. https://doi.org/10.1108/JICES-08-2024-0122
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607
- Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8
- Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. Science and Engineering Ethics, 27, 44. https://doi.org/10.1007/s11948-021-00319-4
- Nasa, P., Jain, R., & Juneja, D. (2021). Delphi methodology in healthcare research: How to decide its appropriateness. World Journal of Methodology, 11(4), 116–129. https://doi.org/10.5662/wjm.v11.i4.116
- Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Cruz, G. O. R., Peixoto, R. M., Guimarães, G. A. S., dos Santos, L. L., Araujo, M. M., Cruz, M., de Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1), 15. https://doi.org/10.3390/bdcc7010015
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
- Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021b). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. https://doi.org/10.1145/3351095.3372873
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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, 1135–1144. https://doi.org/10.1145/2939672.2939778
- Sivakaminathan, S. S., & Musi, E. (2026). ChatGPT is a gender bias echo-chamber in HR recruitment: An NLP analysis and framework to uncover the language roots of bias. AI & Society, 41, 2841–2861. https://doi.org/10.1007/s00146-025-02564-8
- Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Article 17. https://doi.org/10.1145/3465416.3483305
- United Nations. (n.d.). Goal 5: Achieve gender equality and empower all women and girls. United Nations Department of Economic and Social Affairs. https://sdgs.un.org/goals/goal5

Appendix A

Reproducible Web of Science and Scopus search protocol

Pilot run date: 29 July 2026. Coverage: 1 January 2021 through 29 July 2026. Document the exact indexes and subscription coverage. Export full records and cited references, preserve database search histories, and register any deviations from the protocol.

Web of Science Core Collection - primary domain query

TS=(("explainable artificial intelligence" OR XAI OR "interpretable machine learning" OR fairness OR "algorithmic audit*" OR "AI audit*" OR "algorithmic accountability" OR bias OR discrimination) AND (gender

OR women OR woman OR female OR intersectional*) AND (hiring OR recruitment OR employment OR "human resource*" OR "credit scor*" OR lending OR loan* OR borrow*))

Web of Science Core Collection - Delphi/governance query

TS=((Delphi OR "expert consensus" OR "structured consensus") AND ("artificial intelligence" OR algorithm* OR XAI) AND (fairness OR ethics OR governance OR audit* OR accountability OR gender))

Scopus - primary domain query

TITLE-ABS-KEY(("explainable artificial intelligence" OR XAI OR "interpretable machine learning" OR fairness OR "algorithmic audit*" OR "AI audit*" OR "algorithmic accountability" OR bias OR discrimination) AND (gender OR women OR woman OR female OR intersectional*) AND (hiring OR recruitment OR employment OR "human resource*" OR "credit scor*" OR lending OR loan* OR borrow*))

Scopus - Delphi/governance query

TITLE-ABS-KEY((Delphi OR "expert consensus" OR "structured consensus") AND ("artificial intelligence" OR algorithm* OR XAI) AND (fairness OR ethics OR governance OR audit* OR accountability OR gender))

- Apply document type: article, review, conference paper where peer-reviewed; language: English or Spanish; publication years: 2021-2026.
- Deduplicate first by DOI, then normalized title, then author-year-title manual review.
- Record all reasons for full-text exclusion using a controlled taxonomy.
- Perform backward and forward citation searching for every included review and framework article.
- Any subsequent systematic review should reproduce the search through authenticated Web of Science and Scopus subscriptions, preserve the corresponding search and screening records, and document protocol amendments without retrospectively altering the predefined eligibility criteria.

Appendix B
Proposed design for a subsequent formal Delphi study

Table B1. Proposed Questionnaire Domains, Sample Statements, and Rating Dimensions

Domain	Sample statement	Rating dimensions
Governance	The intended decision purpose is legitimate, necessary, and sufficiently narrow.	Relevance; clarity; feasibility
Stakeholders	Women affected by the system participate in defining material harms and remedies.	Relevance; feasibility
Data	The audit must test representativeness, missingness, label bias, and gender proxies.	Relevance; evidence threshold
Intersectionality	Gender results must be reported with relevant intersections where lawful and statistically defensible.	Relevance; feasibility
Model performance	Predictive metrics must be reported overall and by subgroup with uncertainty.	Relevance; minimum evidence
Fairness metrics	No deployment decision should rely on a single fairness metric.	Agreement; rationale
Error asymmetry	False-negative and false-positive harms must be valued in the domain context.	Priority; threshold
Explanations	Local explanations must be tested for fidelity and stability.	Relevance; evidence threshold
Recourse	Counterfactual recommendations must be actionable and feasible for women.	Relevance; feasibility
Human oversight	Reviewers must have time, authority, competence, and evidence to challenge the model.	Relevance; minimum control
Mitigation	Interventions must be compared on a performance-fairness-explainability frontier.	Relevance; feasibility
Decision gate	Severe unexplained intersectional harm is a hard-stop condition.	Agreement; risk tolerance
Vendor governance	Vendors must provide sufficient evidence for independent audit.	Relevance; contractual feasibility
Appeals	Affected persons must receive reasons, correction rights, human review, and a resolution timeline.	Relevance; feasibility

Domain	Sample statement	Rating dimensions
Monitoring	Fairness and explanation drift must trigger predefined investigation and rollback rules.	Relevance; threshold

In a subsequent formal Delphi process, Round 1 would include open-text prompts: Which women are most likely to be invisible in the current audit? Which variables may act as gender proxies? Which explanation would be misleading or infeasible? What harm should automatically suspend the system? What evidence would change your judgment? Round 2 would return anonymized distributions and rationale themes. Round 3 would finalize mandatory, recommended, and exploratory criteria.

Appendix C
Minimum audit evidence package

Table C1. Minimum Audit Evidence Package

Evidence domain	Required artifact
Governance	Approved use case, prohibited uses, accountability matrix, risk classification, vendor obligations.
Data	Datasheet, provenance, sample composition, missingness, labels, proxy analysis, lawful basis, retention.
Model	Model card, version, training procedure, hyperparameters, intended use, limitations, subgroup performance.
Fairness	Metric rationale, thresholds, uncertainty, intersectional results, matched-profile and counterfactual tests.
XAI	Global and local methods, fidelity, stability, user comprehension, actionability, feasibility, privacy/security.
Human process	Decision rubric, training, override authority, override analysis, time and workload, escalation.
Mitigation	Baseline comparison, interventions tested, trade-off frontier, residual risk, independent validation.
Recourse	Notice, reasons, data correction, human review, appeal outcome, resolution time, remedy.
Monitoring	Drift metrics, trigger levels, complaints, appeals, outcomes, re-audit schedule, rollback plan.
Expert-informed validation and decision process	Participant profile, conflicts, response rates, feedback, item history, agreement and disagreement, dissent register.