Issues in Information Systems
Volume 26, Issue 3, pp. 127-135, 2025

DOI: https://doi.org/10.48009/3_iis_2025_2025_110

Explainable AI in employment decision-making: a systematic review of transparency methods in hiring algorithms

Stephen Fabeyo, *Middle Georgia State University. stephen.fabeyo@mga.edu*

Abstract

Artificial intelligence (AI) tools have increasingly been shaping employment decisions, from resume screening to employment tests and automated video interviews. Concerns about the "black box" nature of these tools have grown as many of these algorithmic models offer little insight into how or why hiring decisions occur. This lack of transparency undermines fairness, accountability, and regulatory compliance, especially in contexts where bias may persist or worsen. Explainable AI (XAI) is emerging as a critical strategy to address these concerns by providing interpretable outputs that reveal the model's logic and decision-making factors. This study presents a systematic review of explainable artificial intelligence (XAI) techniques utilized in hiring and employment-related AI systems. It identifies prominent methods, such as Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), decision trees, and counterfactuals, and evaluates their performance in terms of interpretability, fidelity, and fairness. Additionally, it examines their impact on candidate trust and organizational transparency. The findings suggest while XAI can improve understanding and reduce biases, technical complexity and human interpretability issues often limit its effectiveness. The paper concludes by recommending a human-centred approach to XAI design and the development of stronger regulatory frameworks for fair and equitable hiring practices.

Keywords: explainable AI (XAI), algorithmic hiring, interpretable machine learning, AI fairness, AI governance, HR technology.

Introduction

Artificial Intelligence (AI) has become a transformative force in hiring and recruitment processes (Albaroudi et al., 2024). From algorithmic resume screening to automated video interviews and personality assessments, various stages of talent acquisition now leverage AI to make decisions (Kelly-Lyth, 2021; Moss, 2021). These tools promise efficiency, scalability, and objectivity, helping organizations sift through thousands of applications in seconds. Prominent platforms like HireVue, Pymetrics, and Workday deploy machine learning models to evaluate candidates' skills, behaviours, and cultural fit based on both structured and unstructured data (Hofeditz et al., 2022; Wilkens et al., 2025).

However, this rise in AI-driven hiring brings forth a pressing challenge: opacity. Most high-performing AI models, particularly deep learning and ensemble methods, operate as "black boxes," offering little to no visibility into how decisions are made. This lack of transparency undermines accountability and trust, particularly when candidates are rejected without any clear explanation. The deployment of opaque AI systems in hiring also poses substantial legal and ethical risks under emerging regulatory regimes such as the European Union's General Data Protection Regulation (GDPR), New York City's Automated Employment Decision Tool (AEDT) law, and the ADA in the United States (Moss, 2021; Vogel et al., 2024).

Explainable AI (XAI) is emerging as a proposed solution to these challenges, aiming to render complex models interpretable, thereby enabling users to understand, trust, and contest algorithmic decisions. XAI methods range from inherently interpretable models, such as decision trees, to post-hoc explanation tools like SHAP and LIME, which approximate the influence of features on specific outcomes (Kalasampath et al., 2025; Sogancioglu et al., 2023). However, despite the promise of XAI, significant questions remain about its practical utility in employment contexts. Do these explanations truly enhance understanding for HR professionals and job candidates, or do they merely offer the illusion of fairness without addressing deeper structural inequalities in hiring systems (Wasserman-Rozen et al., 2024)?

This study undertakes a systematic review of XAI techniques in employment decision-making to answer these critical questions. By synthesizing findings from academic research, technical validation studies, and policy analysis, the review aims to assess the use of XAI, the most prevalent methods, and whether they effectively deliver on the goals of transparency, fairness, and trust. In doing so, this systematic review provides a foundation for best practices in deploying interpretable AI in human resource technologies.

Problem Statements

The growing integration of AI into hiring processes has introduced new layers of opacity in employment decision-making. These systems often utilize complex models whose internal logic remains inaccessible to both end-users and the individuals impacted by their decisions. This lack of transparency significantly undermines stakeholders' ability to evaluate the fairness, legality, or accuracy of these decisions. When organizations fail to provide interpretable outputs, job applicants find it exceedingly difficult to understand the reasons behind their rejections, employers cannot effectively audit their systems for bias, and regulators struggle to ascertain whether decisions adhere to anti-discrimination laws.

Sanchez-Monedero et al. (2020) estimate that over 83 percent of employers in the United States utilize some form of algorithmic hiring tool, ranging from resume screeners to video assessments, and despite growing adoption, AI hiring tools provide users with no sufficient explanation regarding selection criteria (Binns et al., 2018). These statistics underscore the urgency of integrating robust explainability frameworks into hiring systems. XAI methods offer a potential remedy by generating interpretable representations of algorithmic decisions. However, these methods vary widely in their usability, fidelity, and effectiveness, and are inconsistent or ad hoc in their application across hiring platforms. Moreover, XAI explanations may create a false sense of fairness without genuinely improving decision quality or mitigating discrimination, raising fundamental concerns about the adequacy of XAI as a tool for ethical governance and legal compliance in algorithmic hiring (Wasserman-Rozen et al., 2024). The problem is thus twofold: not only are many hiring algorithms non-transparent, but the tools meant to explain them may themselves be insufficiently interpretable or fair.

Purpose of the Study

This study aims to provide a comprehensive synthesis of existing research on XAI as applied to hiring and employment decision-making systems. Specifically, it seeks to:

- Identify dominant XAI techniques used in hiring contexts
- Evaluate their interpretability, fidelity, and fairness
- Examine the legal, ethical, and usability implications of these methods.

This study provides policymakers, employers, and regulatory bodies with valuable insights to inform the development of ethical AI hiring practices and regulations, thereby ensuring equitable employment opportunities for candidates. The research question for this study is presented below:

RQ: *What are the most effective Explainable AI (XAI) techniques used in employment-related algorithms, and how do they impact transparency in candidate evaluation?*

Literature Review

XAI refers to techniques and frameworks that render machine learning decisions interpretable to human users. It is a response to the "black-box" nature of models like neural networks and ensemble methods (Adadi & Berrada, 2018; Deck et al., 2024; Kalasampath et al., 2025). XAI methods can be categorized broadly into:

- Intrinsic methods: Inherently interpretable models (e.g., decision trees, linear models)
- Post-hoc methods: Explanations derived after model training (e.g., LIME, SHAP)

Moreover, a distinction exists between model-specific methods, designed for particular algorithms, and model-agnostic methods, which are applicable across models. For instance, LIME and SHAP are model-agnostic, offering both local and global interpretability. In contrast, decision trees and attention-based networks are model-specific and often easier to visualize and audit (Kalasampath et al., 2025; Lundberg & Lee, 2017; Ribeiro et al., 2016).

Applications of XAI in HR and Hiring Systems

In the employment domain, XAI methods illuminate predictions in resume screening, personality assessments, and interview scoring. SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are the most commonly used methods (Lundberg & Lee, 2017; Ribeiro et al., 2016). For instance, SHAP actively identifies which attributes (e.g., education level, skill set) most influence candidate scoring, enabling HR professionals to audit fairness (Sogancioglu et al., 2023). Commercial systems, such as Pymetrics and HireVue, incorporate various forms of explainability. Pymetrics uses neuroscience games combined with interpretable models, while HireVue initially used facial analysis and now emphasizes transparency in scoring verbal and behavioural cues (Moss, 2021; Kelly-Lyth, 2021; Vogel et al., 2024). However, trade-offs persist between model accuracy and interpretability. While interpretable models, like decision trees, offer clarity, they often underperform compared to complex models, such as neural networks (Hofeditz et al., 2022).

Interpretability, Fairness, and Trust

User trust in AI systems hinges not just on performance but on the perceived fairness and clarity of outputs. Studies show XAI can improve understanding and acceptance of AI recommendations, especially when paired with domain-specific explanations (Cecil et al., 2024; Deck et al., 2024). However, explainability alone may not mitigate overreliance on incorrect predictions, suggesting a gap between technical transparency and real-world decision utility. Moreover, research warns against oversimplified or misleading explanations that can foster "false fairness" and the illusion of ethical validity without actual mitigation of bias (Wasserman-Rozen et al., 2024). As such, the interpretability of an XAI output must undergo evaluation in terms of human comprehension and its effects on behaviour and trust.

Legal and Ethical Implications of (Un)explainable Hiring Systems

Several regulations, such as the EU's GDPR Article 22 and the NYC Bias Audit Law, require automated decisions, particularly in employment, to be explainable and subject to human oversight (Article 22 GDPR. Automated Individual Decision-Making, Including Profiling | GDPR, 2019; *The New York City Council -*

File #: Int 1894-2020, n.d.). The legal rationale is grounded in the right to explanation, due process, and redress. Yet, current XAI tools often fall short of meeting these standards. As Leben (2023) argues, many XAI methods do not generate the kind of "meaningful information" required for legal compliance, particularly concerning counterfactual or causal reasoning. Ethically, reliance on partially interpretable systems can create new forms of inequality if explanations are accessible only to experts. Furthermore, without mechanisms for appeal or human intervention, XAI risks legitimizing opaque or flawed systems rather than correcting them (Wasserman-Rozen et al., 2024).

Methodology

This review employed a systematic review methodology to synthesize the academic literature about artificial intelligence (AI) and explainable AI (XAI) within employment contexts. Although research on XAI in hiring remains in its infancy, several seminal technical contributions have emerged. Among these, the model-agnostic explanation techniques LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) are the most consistently cited across both empirical investigations and theoretical analyses, and they underpin a substantial portion of XAI applications in recruitment and selection. The review adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Moher et al., 2009) and encompassed literature published between 2016 and 2025. Searches were conducted across IEEE Xplore, the ACM Digital Library, arXiv, SSRN, Google Scholar, Scopus, and Web of Science, using combinations of the terms "explainable AI," "hiring," and "human resources."

Selection Criteria

From an initial pool of over 70 peer-reviewed articles and policy documents, 22 met the full inclusion criteria. A thematic coding approach was used to extract data on explainability methods, interpretability outcomes, fidelity to underlying models, trust measures, and regulatory implications. Articles were included based on the following criteria:

- Peer-reviewed or officially published academic work
- Direct focus on AI and XAI in human resource technology or algorithmic hiring systems
- Demonstrated technical, empirical, or legal evaluation of XAI methods
- Relevance to transparency, fairness, and regulatory compliance in hiring

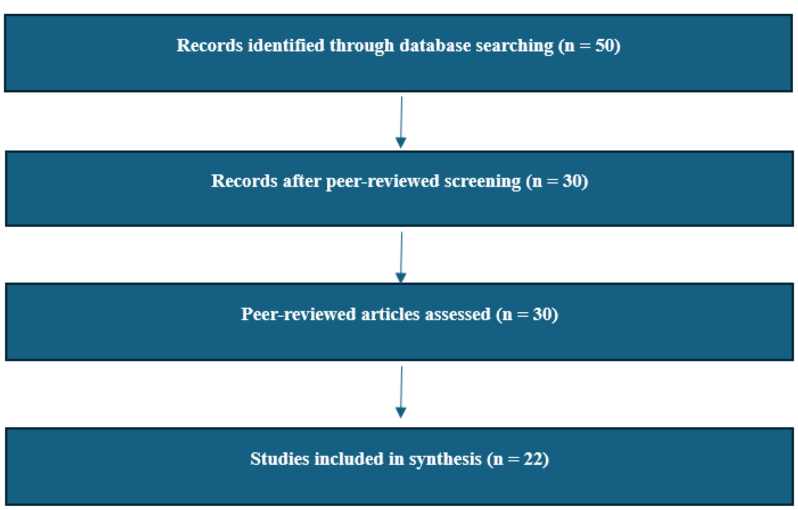

Figure 1. PRISMA Flow Diagram of Study Selection Process. Adapted from Moher et al. (2009).

Analysis and Results

Findings from the Systematic Review

While scholarly interest in explainable AI (XAI) for employment decision-making is expanding, the volume of peer-reviewed research specifically addressing its application in hiring contexts remains limited. Despite the broader growth of XAI publications, studies focused on human resources and algorithmic hiring constitute a small but emerging subfield. The relative scarcity of targeted research highlights an urgent need for interdisciplinary inquiry bridging technical, legal, and organizational perspectives.

Theme 1: Dominant XAI Techniques in Hiring Algorithms

Across the reviewed literature, the most frequently implemented XAI methods in hiring systems include SHAP, LIME, decision trees, counterfactuals, anchor explanations, and rule-based models (Hofeditz et al., 2022; Kalasampath et al., 2025; Leben, 2023; Sogancioglu et al., 2023).

- SHAP was highlighted for its ability to generate both global and local interpretations of model behaviour. It has been successfully applied in identifying proxies for gender and race in recruitment datasets (Sogancioglu et al., 2023).
- LIME provides simplified local approximations that help users understand the influence of specific features on individual predictions, particularly in screening tasks (Kalasampath et al., 2025).
- Counterfactual explanations, although less commonly applied, have gained traction in fairness audits, offering actionable insights by indicating how inputs would need to change to produce different outcomes (Leben, 2023).
- Rule-based systems and decision trees offer higher transparency but are limited in scalability and predictive power (Hofeditz et al., 2022).

Theme 2: Interpretability vs. Fidelity

A significant challenge in implementing XAI within employment contexts is achieving a balance between interpretability and fidelity, which refers to the extent to which simplified models accurately reflect the behaviour of more complex algorithms. Naturally interpretable models, such as decision trees, offer clarity but often fall short in their predictive power, particularly in high-dimensional data environments typical of hiring scenarios (Hofeditz et al., 2022; Deck et al., 2024). Furthermore, Deck et al. contend while inherently interpretable models provide intuitive frameworks for tracing decision-making processes, they do not adequately capture the intricate social nuances present in large datasets, which is vital when considering fairness.

Post-hoc methods such as SHAP and LIME continue to be prominent in XAI toolkits due to their versatility. However, these explanations rely on local approximations that often do not generalize effectively across diverse data segments, particularly marginalized or intersectional subgroups (Deck et al., 2024; Cecil et al., 2024). Gade et al. (2020) further highlight XAI explanations are commonly validated using global performance metrics, which overlook subgroup-specific distortions in decision-making, which is a concern that becomes increasingly significant in legally sensitive areas, such as hiring. Moreover, as Moss (2021) points out, XAI tools like SHAP can yield explanations that reflect majority-behaviour norms, thus rendering them unreliable for candidates with disabilities or neurodivergent conditions, whose features diverge from those patterns. Gade et al. also note current XAI systems insufficiently model heterogeneity in candidate features, making them fragile when applied to non-normative applicant profiles.

Hence, while fidelity and interpretability are both essential, their interaction with fairness mandates a subgroup-sensitive validation framework—a recurring theme in recent critical literature (Deck et al., 2024; Pinto et al., 2025; Leben, 2023).

Theme 3: Candidate Trust and Organizational Transparency
Trust in algorithmic hiring decisions depends on both technical accuracy and how explanations are framed and delivered to human users.

- Research indicates users tend to over-rely on AI recommendations, even when presented with explanations (Cecil et al., 2024).
- Others report diversity-aware explainability features (e.g., highlighting fairness criteria or model adjustments) can positively influence trust and decision quality (Deck et al., 2024; Wilkens et al., 2025).

However, scholars like Wasserman-Rozen et al. (2024) argue current XAI explanations often serve ecosystem legitimacy more than user comprehension or empowerment, thereby potentially undermining due process protections.

Discussion of Findings

The reviewed literature confirms while XAI technologies offer tools for illuminating complex hiring algorithms, they may often fall short of ensuring genuine transparency or fairness in practice. A clear gap exists between technical explainability and end-user comprehension, particularly among job candidates.

Technical Potential vs. Real-World Limitations
SHAP and LIME offer mathematical explanations but do not inherently guarantee ethical or regulatory adequacy. Explanations often fail to meet legal standards of "meaningful information," especially in light of GDPR Article 22 or the NYC AEDT law (Article 22 GDPR. Automated Individual Decision-Making, Including Profiling | GDPR, 2019; *The New York City Council - File #: Int 1894-2020*, n.d.). Moreover, real-world deployment reveals explainability can be misused to lend superficial legitimacy to flawed systems, a concern raised by both legal scholars and disability advocates. For example, Moss (2021) describes how AI tools in video interviews disproportionately penalize neurodivergent candidates, despite being marketed as bias-free.

Human-in-the-Loop and Socio-Technical Design
Several studies advocate for human-in-the-loop models that combine algorithmic efficiency with human judgment. Hofeditz et al. (2022) and Wilkens et al. (2025) demonstrate how human feedback loops can improve AI-based assessments when paired with transparent, diversity-sensitive interfaces. From a socio-technical lens, XAI is not just a technical fix but a governance challenge. Effective implementation requires aligning tools with organizational values, legal obligations, and inclusive design principles (Gade et al., 2020; Pinto et al., 2025).

Regulatory Readiness and Policy Implications
Despite growing regulatory interest in explainability, enforcement mechanisms are weak. As Kelly-Lyth (2021) and Wasserman-Rozen et al. (2024) point out, many XAI systems remain legally opaque or non-compliant due to vague standards and limited oversight. A multi-stakeholder framework is necessary, involving developers, employers, legal bodies, and civil society, to ensure XAI in hiring serves transparency without compromising justice.

Implications of Findings

The findings of this review have clear implications for the design, deployment, and regulation of XAI tools in employment decision-making systems.

For Developers
XAI systems must prioritize user-centred design, encompassing not only technical accuracy but also delivering explanations that are accessible and meaningful to a broad range of users, including HR staff, applicants, and auditors. Developers should engage with interdisciplinary teams, including ethicists, lawyers, and affected communities, during system design (Leben, 2023; Wilkens et al., 2025).

For Employers
Organizations must consider interpretability as a core selection criterion when procuring AI tools. While high-performing black-box models may promise short-term efficiency, they risk legal and reputational harm if they fail to provide adequate transparency. Investing in systems with built-in fairness metrics and explainability can mitigate risk and improve organizational trust (Sogancioglu et al., 2023; Moss, 2021).

For Policymakers
Regulators must move beyond vague demands for "meaningful information" toward clear technical and procedural guidelines for what constitutes sufficient explainability. This includes defining standards for auditability, explanation interfaces, and fairness documentation. Policies should enforce the integration of human review mechanisms and mandate the disclosure of bias audit results (Kelly-Lyth, 2021; Vogel et al., 2024; Wasserman-Rozen et al., 2024).

Conclusion, Limitations of the Study, and Recommendations for Future Research

XAI has emerged as a vital area of research for improving transparency, fairness, and accountability in AI-driven employment decisions. This review highlights methods such as SHAP, LIME, and counterfactual explanations are increasingly applied in commercial hiring systems. While these tools provide a foundation for interpreting complex models, their real-world effectiveness remains constrained by both technical and human-centred limitations. One major finding is the persistent gap between algorithmic explanation and human comprehension. Many current XAI techniques produce outputs that are mathematically precise yet semantically ambiguous to lay users (Leben, 2023; Cecil et al., 2024). Furthermore, XAI tends to perform better at meeting institutional demands for legitimacy than actually empowering affected individuals (Wasserman-Rozen et al., 2024). For example, while decision trees and rule-based models are more interpretable, they are often set aside in favour of less transparent but higher-performing models, undermining the core goal of transparency (Hofeditz et al., 2022; Wilkens et al., 2025).

Limitations of this review include:

- **Limited scope of literature**: While efforts were made to include diverse databases and sources, the availability of peer-reviewed empirical studies specifically on XAI in hiring remains sparse, especially for underrepresented regions and marginalized identity groups.
- **Rapidly evolving field**: As XAI techniques and legal regulations evolve, some of the findings may become outdated, necessitating ongoing review.
- **Focus on technical validation**: Many studies prioritize model fidelity or accuracy but overlook how real users interpret and act upon explanations (Sogancioglu et al., 2023).
- **Underrepresentation of disabled and neurodiverse applicants**: Tools such as HireVue have shown bias against applicants with disabilities, a demographic often ignored in technical audits; it is not clear how XAI can mitigate this underrepresentation (Moss, 2021).

Recommendations for future research and practice include:

- Developing standardized frameworks for evaluating explanation quality, usability, and fairness across stakeholder groups.

- Integrating human-in-the-loop mechanisms to contextualize algorithmic outputs within broader decision processes (Hofeditz et al., 2022).
- Expanding fairness audits to include intersectional and disability-focused evaluations (Moss, 2021).
- Encouraging regulatory bodies to issue clearer guidelines on the definition and sufficiency of explanations, including mandatory disclosure of bias metrics and explanation formats (Kelly-Lyth, 2021; Vogel et al., 2024).

References

Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A comprehensive review of AI techniques for addressing algorithmic bias in job hiring. AI, 5(1), 383–404. https://doi.org/10.3390/ai5010019

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Article 22 GDPR. Automated individual decision-making, including profiling | GDPR. (2019, October 14). *GDPR*. https://gdpr-text.com/read/article-22/

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's Reducing a Human Being to a Percentage": Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

Cecil, R., Yang, S., & Vu, D. (2024). *Explainability does not mitigate the negative impact of AI-based decisions*. ACM FAccT Conference.

Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). A critical survey on fairness benefits of explainable AI. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1579–1595. https://doi.org/10.1145/3630106.3658990

Gade, K., Geyik, S., Kenthapadi, K., Mithal, V., & Taly, A. (2020). Explainable AI in Industry: Practical challenges and lessons learned. *Companion Proceedings of the Web Conference 2020*, 303–304. https://doi.org/10.1145/3366424.3383110

Hofeditz, L., Kalischko, J., & Weinhardt, C. (2022). Applying XAI to an AI-based system for candidate matching: A controlled experiment. *Decision Support Systems*, *159*, 113770. https://doi.org/10.1016/j.dss.2022.113770

Kalasampath, S., et al. (2025). *A Literature Review on Applications of Explainable Artificial Intelligence (XAI)*. ACM Transactions on Interactive Intelligent Systems.

Kelly-Lyth, A. (2021). Challenging biased hiring algorithms. *Oxford Journal of Legal Studies*, *41*(4), 899–928. https://doi.org/10.1093/ojls/gqab006

Leben, D. (2023). Explainable AI as evidence of fair decisions. *Philosophy & Technology*, *36*(1), 12. https://doi.org/10.1007/s13347-022-00589-3

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *In Advances in Neural Information Processing Systems (NeurIPS),* 30.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1016/j.jclinepi.2009.06.005

Moss, H. (2021). Screened out onscreen: Disability discrimination, hiring bias, and artificial intelligence. *Denver Law Review*, *98*(4), 775–805. https://ssrn.com/abstract=3906300

Pinto, G., Mello, C., & Garcia, A. (2025). Explainable AI in Labor Market Applications: *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, 1450–1457. https://doi.org/10.5220/0013384100003890

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to “solve” the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 458–468. https://doi.org/10.1145/3351095.3372849

Sogancioglu, G., Kaya, H., & Salah, A. A. (2023). Using explainability for bias mitigation: A case study for fair recruitment assessment. *Proceedings of the ACM International Conference on Multimodal Interaction*, 548–552. https://doi.org/10.1145/3577190.3614170

The New York City Council—File #: Int 1894-2020. (n.d.). Retrieved May 13, 2025, from https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CText%7C&Search=

Wasserman-Rozen, H., Gilad-Bachrach, R., & Elkin-Koren, N. (2024). Lost in translation: The limits of explainability in AI. *Cardozo Arts & Entertainment Law Journal*, *42*(2), 391–437. https://ssrn.com/abstract=4567383

Wilkens, U., Lutzeyer, I., Zheng, C., Beser, A., & Prilla, M. (2025). Augmenting diversity in hiring decisions with artificial intelligence tools. *The International Journal of Human Resource Management*. https://doi.org/10.1080/09585192.2025.2492867

Vogel, M., Chertoff, M., Wiley, J., & Kahn, R. (2023). Is your use of AI violating the law? An overview of the current legal landscape. New York University Journal of Legislation & Public Policy, 26, 1029.