ISSN 3071-9658

European Journal of Interdisciplinary Issues

https://www.eujini.org.pl

Bias in AI Admission and Hiring

Taisiia Prykhodko [1] *

[1] *Igor Sikorsky Kyiv Polytechnic Institute (Ukraine). Bachelor's Degree in Information Systems and Technologies. Graduate Assistant, American University.*

* ***Corresponding Author,*** e-mail: prykhodko.taisiiaa@gmail.com

ARTICLE INFO

Research Article

Received:
30 December 2025
Revised:
15 February 2026
Accepted:
7 March 2026
Published online:
25 March 2026

DOI: 10.5281/zenodo.20079910

ABSTRACT

The integration of artificial intelligence into admission and hiring systems offers modernizing efficiency yet entails significant risks regarding the automation of inequality. This paper examines the technical sources of algorithmic bias, including proxy variables, feedback loops and data provenance, within educational and workforce pipelines. Employing a critical synthesis methodology, the study provides a theoretical analysis of the intersection of deep learning architectures, emerging Generative AI (GenAI) capabilities and evolving regulatory frameworks such as the EU AI Act and NYC Local Law 144. The analysis highlights that "fairness through unawareness" is technically insufficient and that the deployment of Large Language Models (LLM) introduces new risks of algorithmic monoculture and sociolinguistic homogenization. Addressing the mathematical impossibility of simultaneously satisfying competing fairness metrics, the paper proposes a responsible AI governance framework. This framework emphasizes Algorithmic Impact Assessments (AIA), continuous disparate impact auditing, and a transition from Human-in-the-Loop to Human-in-Command architectures. The study concludes that responsible implementation requires shifting the role of AI from a gatekeeper to a decision-support scout, ensuring that automated systems meet rigorous standards of accountability and symbiotic human oversight.

KEYWORDS

algorithmic bias, AI recruitment, generative AI, Human-in-the-Loop, HRIS, explainable AI (XAI).

Introduction

The integration of Artificial Intelligence into professional and educational spheres has reshaped the pathways to social mobility. Today access to education and employment is increasingly mediated by algorithmic systems, ranging from predictive scoring models in university admissions to automated resume screeners and psychometric profiling tools in recruitment (Bogen & Rieke, 2018). While the deployment of these technologies is driven by the imperatives of efficiency, scalability and the reduction of subjective human error, their use has exposed a critical systemic vulnerability - the amplification of existing inequalities (Eubanks, 2018).

This paper examines the integration of AI into educational and workforce pipelines and argues that algorithmic bias often represents the codification of historical disparities embedded within training data (Barocas & Selbst, 2016). Beyond general critique, this analysis dissects the technical mechanisms of discrimination, specifically proxy variables, feedback loops and the opacity of deep learning models (Pasquale, 2015). Moreover, work addresses the emerging challenges posed by Generative AI (GenAI), investigating how Large Language Models (LLM) may contribute to algorithmic monoculture and sociolinguistic homogenization in talent selection (Bommasani et al., 2021).

The objective of research is to synthesize distinct strands of algorithmic accountability, quantitative fairness metrics, legal compliance frameworks (such as the EU AI Act) and engineering best practices into a coherent strategy for responsible implementation (Madiega, 2021). By analyzing the mathematical trade-offs of satisfying varying definitions of fairness alongside the practical necessity of decision-making, the paper proposes a shift in perspective. Also seeks to define how institutions can engineer systems that justify their role in evaluating human potential through structured transparency, rigorous auditing and human-AI collaboration.

Geographically, the regulatory analysis is primarily anchored in European (EU AI Act) and North American (NYC Local Law 144) legal frameworks, reflecting the current concentration of specific legislation regarding automated hiring. Functionally, the scope of the study is restricted to external selection processes, specifically admissions and initial talent acquisition, excluding internal workforce management issues such as algorithmic performance review or automated termination.

Finally, given the velocity of GenAI development, the article addresses current model architectures, recognizing that future iterations may introduce novel ethical complexities.

Literature Review

The scholarly discussion on algorithmic bias emphasizes that automated discrimination is a structural rather than a purely technical issue. Pasquale (2015) established the foundational critique of "black box" algorithms, highlighting how their inherent opacity masks the logic of decisions that control access to economic opportunities. This socio-technical critique is expanded by O'Neil (2016), who characterizes biased predictive models as "weapons of math destruction" that reinforce inequality under the guise of objective data.

The mechanisms of this "automated inequality" are further dissected by Eubanks (2018) and Benjamin (2019), who demonstrate how high-tech profiling often punishes marginalized populations and replicates racial hierarchies through software. Specifically, in the recruitment sector, Barocas and Selbst (2016) analyze the "disparate impact" of Big Data, noting that algorithms can produce discriminatory outcomes even in the absence of explicit intent.

Technical sources of such bias are frequently attributed to "proxy variables." Bogen and Rieke (2018) and Prince and Schwarcz (2020) illustrate how neutral data points can correlate with protected characteristics, leading to indirect discrimination. Furthermore, the mathematical complexity of solving these issues is highlighted by Kleinberg et al. (2016), who prove the "impossibility theorem"—the fact that different definitions of fairness are often mathematically incompatible.

Current research also addresses the evolving regulatory and technological landscape. Madiega (2021) discusses the framework of the EU AI Act as a critical response to these risks. Finally, as the

field shifts toward Large Language Models, Gebru et al. (2021) emphasize the necessity of "Datasheets for Datasets" to ensure transparency and accountability in the provenance of data used to train the next generation of hiring and admission systems.

Problem Statement

The integration of Artificial Intelligence into admission and hiring processes is primarily driven by the need for enhanced operational efficiency and the reduction of human subjective error. However, the deployment of automated predictive models has revealed a profound structural vulnerability: the capacity of algorithms to institutionalize and scale social inequalities. The core challenge lies in the technical and ethical tension between mathematical optimization and social equity. Automated systems often function as opaque mechanisms that codify historical disparities present in training data, transforming past prejudices into future gatekeeping protocols. Despite emerging regulatory efforts, there remains a significant gap in establishing governance frameworks that can effectively navigate the trade-offs between competing fairness metrics. This situation necessitates a shift toward more transparent, accountable, and human-centered oversight architectures to ensure that AI serves as a support tool for identifying potential rather than a digital barrier to social mobility.

Methods and Materials

This study employs a qualitative, interdisciplinary design utilizing a critical synthesis and conceptual analysis methodology to examine the intersection of algorithmic architecture, legal jurisprudence and social ethics. It is important to note, that this research is theoretical in nature - it does not involve the deployment of novel machine learning architectures, experimental datasets or primary empirical data collection. Instead, the methodology is structured around a theoretical comparative analysis of technical fairness metrics vis-a-vis legal compliance requirements, followed by the development of a normative governance framework.

The analysis draws upon a tripartite corpus of data collected from sources published primarily between 2014 (marking the onset of the Amazon hiring algorithm case) and 2024 (Dastin, 2018). A review of peer-reviewed publications from computer science and engineering databases (IEEE Xplore, ACM Digital Library) was conducted to identify mechanisms of algorithmic bias (proxy variables, feedback loops) and quantitative definitions of fairness (statistical parity, equalized odds) (Mehrabi et al., 2021). Primary legal texts were analyzed to establish the compliance landscape. The geographical scope is restricted to the European Union (GDPR, EU AI Act) and the United States (specifically NYC Local Law 144 and EEOC guidelines), representing the current frontier of AI regulation (U.S. Equal Employment Opportunity Commission, 2023). Data on real-world implementations were aggregated from technical reports, white papers and investigative audits regarding major AI vendors in the HRIS sector (HireVue, Pymetrics) and proprietary internal tools (Amazon).

The study proceeds through three analytical phases:

1. Technical deconstruction - decomposing the architecture of predictive models to isolate specific vectors of bias, including historical data encoding and feature selection.
2. Comparative matrix analysis: evaluating the utility versus risk of Generative AI (GenAI) applications in HRIS. This phase involved classifying AI sub-tasks (scheduling vs. psychometric profiling) based on their potential for disparate impact and operational efficiency.
3. Synthesis and framework design: integrating the conflicting demands of mathematical fairness metrics and legal mandates to engineer a proposed responsible AI governance framework. This phase utilizes a prescriptive approach to define necessary oversight mechanisms, such as Algorithmic Impact Assessments (AIA) and Human-in-Command (HIC) protocols.

Results and Discussion

Integration of artificial intelligence systems into educational and workforce pipelines: technical sources of algorithmic bias

Contemporary AI-based selection systems function primarily as predictive models designed to forecast a candidate's future success based on historical patterns. In modern recruitment, platforms utilize diverse input data, ranging from facial micro-expressions and linguistic complexity to gameplay behavior, to generate an "employability score" (Raghavan et al., 2020). Similarly, in higher education algorithms process academic transcripts, standardized test scores and extracurricular data to predict student retention rates and graduation success (Aulck et al., 2016).

As noted in foundational critiques of algorithmic fairness, AI does not invent discrimination - it formalizes existing historical inequality (Benjamin, 2019). Supervised learning models are typically trained on datasets representing "successful" candidates from previous years. If an organization has historically favored specific demographic groups, whether consciously or due to implicit human bias, the algorithm learns to prioritize characteristics associated with those groups (Figure 1).

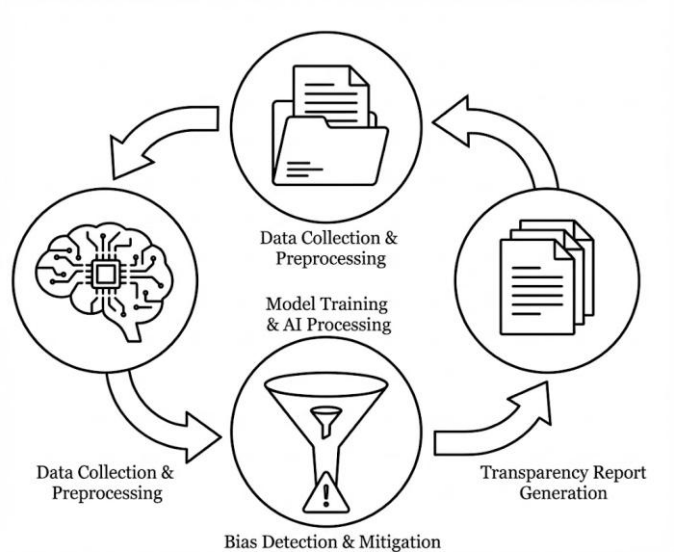

Figure 1. The feedback loop of algorithmic bias. Illustrates how historical data trains the model, which selects candidates, creating new biased data that reinforces the original model

A common misconception in the technical design of such systems is that fairness can be achieved by removing protected attributes (such as race, gender or age) from the dataset, a process known as "fairness through unawareness". However, this approach is technically insufficient due to the existence of proxy variables (indirect indicators) (Prince & Schwarcz, 2020).

It is also important to note, that the technical architecture of these systems often creates self-reinforcing feedback loops. In a typical deployment, the algorithm selects a cohort of candidates and only these selected candidates become sources of performance data that are fed back into the model to update its weights (Figure 2). Consequently, the system remains blind to "counterfactuals" - the potential success of the candidates it rejected (O'Neil, 2016). If the model initially excludes a specific demographic group based on flawed criteria, it never receives data to refute this exclusion. Over time, this narrows the definition of a "qualified candidate", creating a closed loop where the algorithm becomes increasingly confident in its biased predictions, effectively automating the marginalization of underrepresented groups (Ensign et al., 2018).

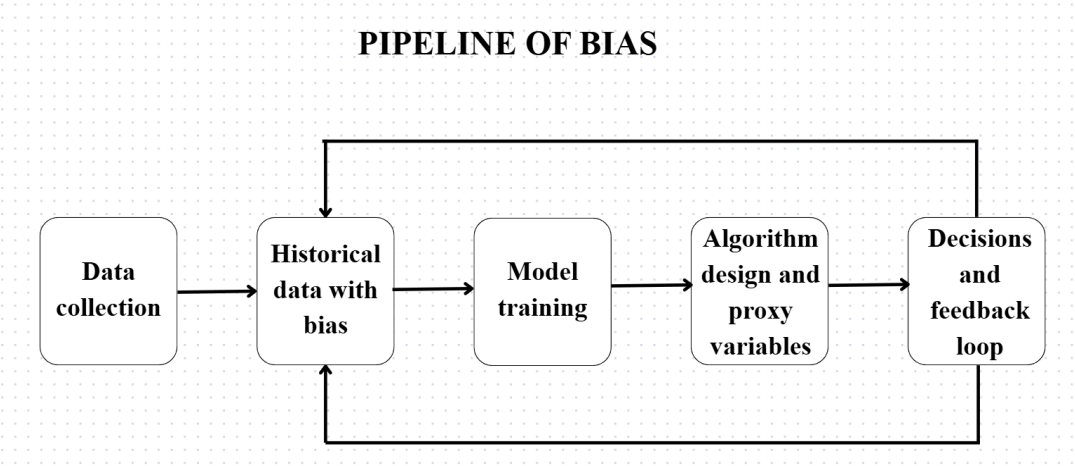

Figure 2. Pipeline of bias diagram

An illustrative case study of such a bias failure is Amazon's internal hiring system (2014-2018). The system was trained on resumes submitted over ten years, a dataset dominated by male candidates. Consequently, the model learned that male candidates were preferable. It actively penalized resumes containing the word "women's" and downgraded graduates of two women's colleges.

This case demonstrates that merit, when defined based on historical data, is often erroneously conflated with demographic resemblance to the existing status quo.

Quantitative fairness metrics and ethical-legal constraints - transparency, explainability and compliance requirements

A critical finding in algorithmic accountability research is that fairness represents a series of mutually exclusive value judgments (Chouldechova, 2017). There is no single mathematical equation for "fairness" - instead, researchers utilize competing metrics, each optimizing for distinct ethical objectives (Table 1).

Table 1. Comparison of algorithmic fairness metrics

Metric	Definition	Ethical goal	Technical limitation
Statistical parity	Acceptance rates must be equal across demographic groups	Diversity + representation	May reduce utility by ignoring qualification disparities
Equalized odds	True positive and false positive rates must be equal across groups	Accuracy + equal opportunity for qualified candidates	Difficult to achieve if base rates differ significantly
Predictive parity	The probability of success given a score is equal across groups	Calibration + consistency of the score meaning	Can lead to disparate impact if one group is historically disadvantaged

Statistical Parity requires that positive outcomes be distributed at equal rates across different demographic groups, regardless of the underlying ground truth labels in the data (Dwork et al., 2012). While this aligns with the goal of increasing diversity, critics argue it may compromise model utility by ignoring legitimate disparities in qualification distributions.

In contrast to statistical parity, the equalized odds metric focuses on predictive accuracy. It necessitates those true positive rates (sensitivity) and false positive rates be equivalent across all groups (Hardt et al., 2016). This ensures that qualified candidates from minority groups possess the same probability of selection as qualified candidates from the majority group.

The final metric, predictive parity, demands that the probability of actual candidate success be consistent across all groups without bias. Mathematical proofs, such as the impossibility theorems presented by Kleinberg et al., demonstrate that, except in rare instances where base rates are identical, it is mathematically impossible to satisfy these metrics simultaneously (Kleinberg et al., 2016). Consequently, the selection of a metric is a policy decision regarding which type of error is deemed more ethically acceptable.

The complexity of modern deep learning models often precipitates the "black box" phenomenon, where the internal decision-making logic remains opaque even to its developers. In high-stakes domains such as employment and education, such opacity is ethically untenable. To address this challenge, the field has pivoted toward Explainable AI (XAI), prioritizing interpretability alongside accuracy. Techniques such as SHAP (Shapley Additive exPlanations) values enable auditors to discern which specific features (years of experience vs employment gaps) drove a specific prediction (Lundberg & Lee, 2017). Moreover, counterfactual explanations are becoming a standard requirement for providing actionable feedback to rejected candidates.

Legal frameworks are rapidly adapting to enforce these technical requirements. The regulatory focus has shifted from intentional discrimination to disparate impact (unintentional adverse effects). Articles 13-15 and 22 of the General Data Protection Regulation (GDPR) establish a right to explanation for automated decisions, effectively restricting the deployment of uninterpretable black-box models within European jurisdictions (Goodman & Flaxman, 2017). NYC Local Law 144, implemented in 2023, specifically mandates that automated employment decision tools (AEDT) undergo independent bias audits prior to use, establishing New York as the first jurisdiction to require quantitative proof of fairness in hiring (New York City Council, 2021). The EU AI Act categorizes AI systems used in employment, workforce management and educational access as high risk (European Commission, 2021). This designation imposes strict obligations regarding data governance, record-keeping, transparency, and human oversight. Consequently, regulatory compliance extends beyond data privacy. It now necessitates the continuous auditing of model outputs to ensure adherence to defined fairness thresholds and legal non-discrimination standards.

Comparative matrix analysis of generative artificial intelligence integration within human resource information systems (HRIS)

The evolution of algorithmic selection has shifted toward the integration of Generative Artificial Intelligence (GenAI) and Large Language Models (LLM). This approach fundamentally alters the architecture of Human Resource Information Systems (HRIS). While traditional machine learning models relied on structured data points, GenAI processes unstructured semantic data, offering new capabilities in candidate engagement and assessment (Table 2).

Table 2. Risk-utility matrix of GenAI integration in HRIS

Application domain	GenAI role	Utility level	Risk level	Primary risk factor
Operational automation	Scheduling, FAQs, standard letters	High	Low	Accessibility issues
Resume summarization	Parsing CVs, matching semantics	High	Medium	Context stripping + summarization bias
Psychometric profiling	Personality analysis, video interviews	Low	Critical	Sociolinguistic bias + hallucinations

Administrative and operational automation is characterized by high utility and low risk. Automated scheduling, chatbots for procedural inquiries (FAQ) and the generation of standardized rejection or offer letters represent broad areas of application. LLMs utilize Retrieval-Augmented Generation (RAG) to access company policy documents and provide factual responses (Lewis et al., 2020). This represents the most successful integration of GenAI. By automating low-stakes administrative tasks, HRIS reduces the "time-to-hire" metric without subjecting a candidate's professional merit to algorithmic judgment. The risk of bias here is minimal, limited primarily to accessibility issues rather than decision-making discrimination.

Semantic search and resume summarization also possess high utility. However, the risk profile increases. Applications include summarizing long CV into standardized profiles and matching candidates to job descriptions based on semantic meaning rather than exact keywords. The operating mechanism relies on vector embeddings to transform candidate profiles and job descriptions into a high-dimensional space to identify semantic proximity. While this improves upon rigid keyword matching (which penalizes candidates for using different terminology), it introduces summarization bias (Zhang et al., 2024). LLM-trained on standard Western corporate data may strip

away the unique, culturally specific context from a resume to fit it into a standard template. Candidates with non-traditional backgrounds may undergo homogenization, where their unique value propositions are diluted during the summarization process.

The final implementation presented is automated interviewing and psychometric profiling. However, this approach yields low utility and a critical level of risk. The area of application here involves conversational AI agents conducting first-round interviews or analyzing candidate text for personality traits (the Big 5). The mechanism is based on sentiment analysis and psycholinguistic profiling. This category presents significant ethical challenges. GenAI models are susceptible to sociolinguistic bias (Blodgett & O'Connor, 2017). They often correlate professionalism with specific dialects and sentence structures. A comparative review indicates that these systems may penalize non-native speakers and neurodivergent candidates whose speech patterns deviate from the training distribution.

Moreover, the risk of model hallucinations, where the AI system generates non-verifiable assertions about a candidate or fails to correctly interpret pragmatic language phenomena such as sarcasm, implicit meaning, or contextual nuance, undermines the reliability of such systems for high-stakes decision-making scenarios (Ji et al., 2023).

A distinct finding in the analysis of GenAI in HRIS is the risk of algorithmic monoculture. As candidates use GenAI to write resumes and employers use GenAI to parse them, the recruitment ecosystem enters a closed feedback loop. This results in a paradox: while the system becomes more efficient at processing text, it may become less effective at identifying diverse talent or outliers who drive innovation. The comparative matrix demonstrates that successful integration into HRIS requires a human-in-the-loop architecture. GenAI excels as a synthesizer of information (document summarization, scheduling), but has significant limitations as an arbiter of qualification. Thus, the technical imperative is to restrict the role of GenAI to decision-support functions rather than autonomous decision-making.

Engineering a responsible AI development framework with built-in human oversight and governance mechanisms

Prior to model training, organizations are required to conduct an Algorithmic Impact Assessment (AIA). This is a preventive protocol designed to identify risks before they are codified.

Following the “Datasheets for datasets” methodology, developers must document the origin of training data (Gebru et al., 2021). This includes explicit statements regarding demographic representation and the acknowledgment of historical inequities embedded within the labels. Instead of testing solely for accuracy, engineers must engage in red teaming - deliberate attempts to break the model by inputting synthetic profiles of protected groups to reveal discriminatory patterns. To mitigate the risks associated with black box decisions, the operational architecture must shift from full automation to human augmentation. The algorithm should output a recommended score with confidence intervals and key contributing factors (utilizing XAI tools such as SHAP), rather than a binary reject/accept decision.

To prevent automation bias (where humans passively accept AI outputs), the interface must require recruiters to explicitly justify why they agree or disagree with the AI recommendation, particularly in borderline cases.

High-stakes decisions, especially rejections, must undergo human review. The framework mandates that no candidate be rejected solely by an automated system without a mechanism for human appeal (Figure 3).

Models are not static - they suffer from concept drift, where their accuracy degrades as real-world conditions change. Real-time dashboards must monitor acceptance rates across demographic groups. If the ratio falls below the four-fifths rule (or other legal thresholds), the system must activate an automatic circuit breaker, pausing decision-making until the anomaly is investigated.

To break the cycle of self-reinforcing bias, organizations must periodically inject exploration data, deliberately hiring a small cohort of candidates, who were rejected by the algorithm, to generate new ground-truth data and verify if the model’s predictions were accurate.

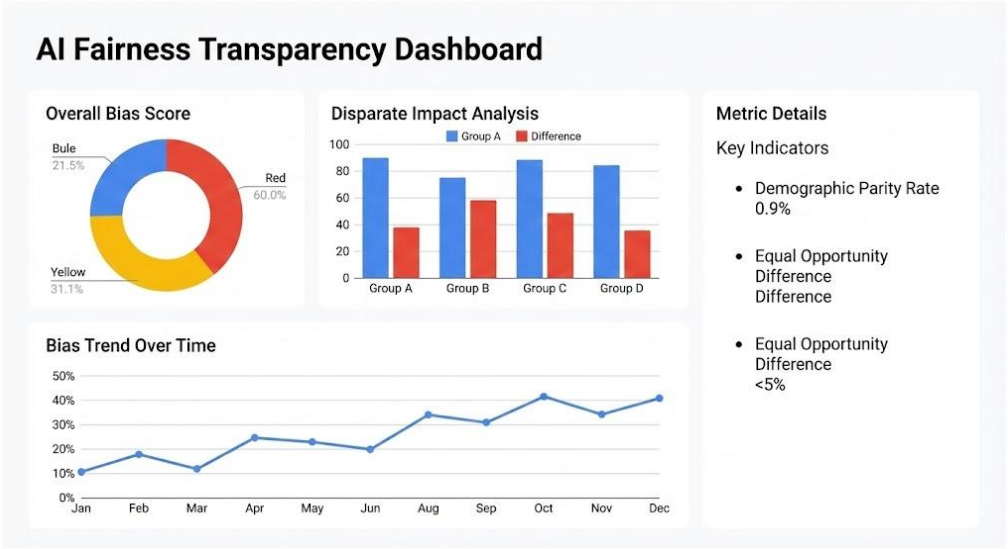

Figure 3. Fairness dashboard mockup. Generative Artificial Intelligence was used to create this image

Future directions of machine-assisted talent selection: human-AI collaborative decision architectures

As the limitations of fully autonomous algorithmic selection become evident, the focus of research and development is shifting toward Human-AI Collaboration (HAIC). This paradigm transcends mere automation (the replacement of human labor) and moves toward augmentation (the enhancement of human judgment). The future of talent selection lies in "Hybrid intelligence" architectures, where the computational power of AI in processing multidimensional data is coupled with the contextual reasoning and ethical intuition of human decision-makers.

The most promising operational models utilize a dynamic division of cognitive labor. Instead of a linear handoff, where AI filters and human interviews, future architectures will function as symbiotic loops. The algorithm acts as a navigational aid, flagging patterns that human cognitive limits might miss (for example, identifying a talented coder from a non-target university based on open-source contributions rather than degree prestige). The human retains full control over the trajectory, verifying the rationale behind the AI's flag and assessing interpersonal nuances, such as leadership potential or cultural contribution, that remain unquantifiable.

If the AI assigns a low score to a candidate whom a human recruiter rates highly, the system must initiate a resolution dialogue, requiring the human to articulate their intuition or the AI to display its evidence.

Current regulatory frameworks emphasize a Human-in-the-Loop (HITL) approach, which often devolves into a rubber-stamping exercise due to time pressures. The future direction is the "Human-in-Command" model. In this architecture, the AI does not output a decision, but rather a hypothesis. This shift redefines the algorithm's role from a gatekeeper to a scout, ensuring that outliers, often the most innovative candidates, are not filtered out by rigid statistical methods.

The integration of these architectures will necessitate a new set of professional competencies. The recruiter of the future will be an active auditor of algorithms. HR professionals will require training to interpret confidence intervals and recognize when a model is relying on proxy variables. As routine screening becomes automated, the human role will pivot toward high-touch relationship building and complex ethical adjudication.

Ultimately, the collaborative architecture aims to resolve the "Fairness-Utility Trade-off." AI provides consistency (treating similar cases similarly), while humans provide context (understanding why a case might be dissimilar). True fairness emerges from the dialectic between computational consistency and human oversight.

Conclusion

The analysis presented in this paper demonstrates that the quest for a perfectly fair AI in admissions and hiring is mathematically impossible and involves complex social dynamics. Algorithmic bias represents a reflection of deep-seated structural inequalities that predictive models, by design, tend to formalize. This examination of Generative AI further reveals that without rigorous governance, the deployment of LLM in HRIS threatens to create an algorithmic monoculture, prioritizing homogenized profiles.

However, the impossibility of technical perfection does not absolve institutions of the responsibility for ethical improvement. By shifting the focus from blind automation to ethical engineering, organizations can harness the efficiency of AI while safeguarding human dignity. The proposed governance framework, anchored in pre-deployment impact assessments, continuous post-deployment auditing and Human-in-Command architectures, offers a pathway to reconcile the utility of automation with the mandates of equal opportunity.

Finally, the legitimacy of AI in determining human opportunities depends on a fundamental principle - automated systems must operate under a framework of accountability. This legitimacy is established not through black-box efficiency, but through structured transparency, auditability and the recognition that, in matters of human potential, probabilistic outputs should serve as decision support rather than definitive judgments.

References

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *arXiv*. https://arxiv.org/abs/1606.06364

Barocas, S., & Selbst, A. D. (2016). Big Data's disparate impact. *California Law Review*, (104), 671–732. https://doi.org/10.15779/Z38BG31

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity. https://www.politybooks.com/bookdetail?book_slug=race-after-technology-abolitionist-tools-for-the-new-jim-code--9781509526390

Blodgett, S. L., & O'Connor, B. (2017). Racial disparity in natural language processing: A case study of social media African-American English. *arXiv*. https://arxiv.org/abs/1707.00061

Bogen, M., & Rieke, A. (2018). *Help wanted: An examination of hiring algorithms, equity and bias*. Upturn. https://www.upturn.org/reports/2018/hiring-algorithms/

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. https://arxiv.org/abs/2108.07258

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). https://doi.org/10.1145/2090236.2090255

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, (pp. 160–171). https://proceedings.mlr.press/v81/ensign18a.html

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police and punish the poor*. St. Martin's Press.

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, (29). https://arxiv.org/abs/1610.02413

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*. https://arxiv.org/abs/1609.05807

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, (33), 9459–9474. https://arxiv.org/abs/2005.11401

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, (30). https://arxiv.org/abs/1705.07874

Madiega, T. (2021). *Artificial Intelligence Act*. European Parliamentary Research Service. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI%282021%29698792

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), 1–35. https://doi.org/10.1145/3457607

New York City Council. (2021). *Local Law 144 of 2021: Automated employment decision tools*. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page

O'Neil, C. (2016). *Weapons of math destruction: How Big Data increases inequality and threatens democracy*. Crown.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://www.hup.harvard.edu/books/9780674970847

Prince, A. E., & Schwarcz, D. (2020). Proxy discrimination in the age of Artificial Intelligence and Big Data. *Iowa Law Review*, (105), 1257. https://scholarship.law.umn.edu/faculty_articles/682/

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (pp. 469–481). https://doi.org/10.1145/3351095.3372828

U.S. Equal Employment Opportunity Commission. (2023). *Select issues: Assessing adverse impact in software, algorithms, and Artificial Intelligence used in employment selection procedures under Title VII of the Civil Rights Act of 1964.*

Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2024). How language model hallucinations can snowball. *arXiv*. https://arxiv.org/abs/2305.13534