ACADEMIA International Journal for Social Sciences
Volume 4, Issue 4(b), 2025 ISSN-L (Online): 3006-6638

# Beyond the Black Box: Designing Equitable Algorithmic Governance for High-Stakes Institutional Screening

**Amir Zaheer**
amirzaheer85@yahoo.com
UBL Bank Customer Service Representative (CSR)

**Corresponding Author: Amir Zaheer** amirzaheer85@yahoo.com

**Received:** 24-10-2025 **Revised:** 09-11-2025 **Accepted:** 23-11-2025 **Published:** 06-12-2025

## ABSTRACT

*Purpose: The rapid integration of artificial intelligence into high-stakes institutional screening threatens procedural justice by obscuring historical demographic biases within opaque algorithmic models. While existing literature extensively diagnoses this "black box" problem, there remains a critical dearth of actionable, legally defensible governance frameworks capable of preventing proxy discrimination. This paper aims to bridge the gap between computer science fairness metrics and administrative jurisprudence.*

*Design/Methodology/Approach: Grounded in theories of organizational justice and administrative equity, this conceptual paper critically synthesizes recent legal, ethical, and sociotechnical scholarship to construct a comprehensive algorithmic governance architecture.*

*Findings: This study proposes a tripartite structural governance framework. First, it mandates rigorous pre-deployment algorithmic impact assessments (AIAs) to detect and mathematically neutralize proxy variables in training data. Second, it requires the integration of explainable artificial intelligence (AI) (XAI) metrics to guarantee decision contestability and operationalize the right to explanation. Third, it designs stringent " human-in-the-loop " (HITL) protocols, introducing mechanisms, such as blinded baselines and independent algorithmic appeals committees, to actively counteract automation bias and enforce human accountability.*

*Originality/Value: Moving beyond theoretical critique, this research delivers a concrete managerial and policy blueprint. It demonstrates how institutions must fundamentally restructure their technological procurement and internal oversight to align with emerging civil rights legislation, ensuring that AI serves as an instrument of administrative equity rather than an automated architect of systemic marginalization.*

***Keywords:*** *Algorithmic Governance; Explainable AI (XAI); Algorithmic Bias; Procedural Justice; Human-in-the-Loop (HITL).*

## INTRODUCTION

One of the most significant shifts in the paradigm of the administrative realm in the 21st century is the incorporation of artificial intelligence (AI) and machine learning algorithms into institutional decision-making. In the public and in the private sectors, organizations are more often and more likely to assign complex and evaluative tasks to automated systems because it is essential to the operational efficiency imperative and the hope of data-driven objectivity. This change is particularly strong in areas such as high-stakes resource distribution and candidate screening. Algorithms for selecting candidates from large applicant groups are now common in competitive academic settings, including the assessment of postgraduate scholarship applications and PhD admissions, the provision of high-value research grants, and similar areas (Fabeyo, 2025). Although such predictive models are unmatched in the speed of processing multidimensional candidate data, their implementation has occurred faster than the

development of legal and ethical frameworks to address these models. Consequently, institutions are walking a fine line, where the mathematical efficiency of automated screening is pitted directly against the primary concerns of administrative fairness, procedural justice, and equal opportunity.

The main focus of this tension is the problem of black boxes that is prevalent with complex predictive algorithms, especially in the field of deep neural networks and advanced machine learning models (Doshi-Velez & Kim, 2017). In contrast to the traditional, human-administered type of administrative operations, where a committee or adjudicator can explain the particular reasoning behind a rejection or acceptance, black-box models behave at a certain level of obscurity that cannot be subject to the usual types of oversight. These systems produce predictions by detecting nonlinear trends using thousands of variables–trends that are also frequently incomprehensible, even to the designers who created the system (Rudin, 2019). This completely opaque situation poses a huge accountability vacuum in the institutional screening environment. In situations where an algorithm finds that a candidate is not qualified to hold a competitive academic placement or a legal fellowship, the right to question the logic behind the system is practically void because of the inability to challenge the system.

Moreover, the belief that algorithmic systems offer objective and neutral assessments is incorrect on a fundamental level. Algorithms are technologically social objects; they are trained to predict, using historical training data that are already saturated with the historical biases, structural inequalities, and discriminant precedents of the past (Barocas and Selbst, 2016; O'Neil, 2016). In a competitive institutional screening where past data represents some type of demographic bias–such as the history of giving preference to particular socioeconomic, racial, or gender profiles in doctoral applications or corporate grants–the algorithm will be trained to implicitly learn to punish applicants who do not align with that historical profile. It does this not so much by explicit, protected characteristics, but by highly correlated proxy variables, including zip codes, secondary educational affiliations, or linguistic patterns in a research proposal (Goodman, 2025). Therefore, instead of being objective assessors of merit, AI screening tools, without being controlled, have the danger of becoming highly effective engines of the reproduction and the process of automatizing historical marginalization, working under a flag of mathematical infallibility.

This creates a key central tension for contemporary administrators and legal theorists: the tension between the indisputable utility of algorithmic efficiency operations and the serious institutional danger of reinforcing structural demographic bias. The present-day administration of civil rights and administrative law is unprepared to deal with this friction. Conventional jurisprudence of discrimination is based on the ability to demonstrate disparate treatment or intent, a virtually unattainable evidentiary burden in cases where the perpetrator of discrimination is a proprietary and opaque algorithm optimizing on apparently neutral organizational measures (Kleinberg et al., 2016). With an increasing number of institutions basing their high-stakes appraisals on them, they are also exposed to both legal and deep moral disasters because of a lack of strong, structurally embedded governance systems. The academic rhetoric will thus need to change. Although the existence of algorithmic bias has been successfully diagnosed ( Buolamwini & Gebru, 2018; Obermeyer et al., 2019), there is an acute lack of literature explaining how institutions need to structurally regulate such systems to ensure that they address the set legal and ethical requirements.

Filling this crucial literature gap, this paper refocuses the discussion on the diagnosis of algorithmic bias towards institutional accountability architecture. The research question that guides the investigation is: What can institutions do to design and implement transparent algorithmic governance structures that reduce historical bias in high-stakes screening and meet new demands of administrative equity and legal responsibility? This paper provides conclusive structural and conceptual input to the disciplines of business ethics, administrative law, and algorithmic governance in response to this question. It goes beyond theoretical criticism to provide a legally valid model of governance. We recommend a set of interventions, including pre-deployment fairness audits, the requirement of explainable AI (XAI)

procedures to access the black box, and the incorporation of the structural features of mechanisms of human-in-the-loop oversight of algorithms with the ability to override algorithmic determinations in situations of procedural injustice.

The rest of the paper is organized in the following manner. Section 2 summarizes the literature available regarding algorithmic bias and the shortcomings of the existing automated screening tools when it comes to their use in resource allocation. Section 3 grounds the study in existing research on procedural justice and administrative equity, placing AI in the context of contemporary jurisprudence. Section 4 provides the analytical methodology, and Section 5 presents the main contributions of the proposed structural governance mechanisms. In Section 6, the author addresses the ambiguous nature of the human-algorithm interface, particularly the psychological phenomenon of automation bias. Last, but not the least, Section 7 contains the policy recommendations that can be undertaken by institutional administrators, and the last section is the conclusion and the future research directions provided in Section 8. Finally, this paper concludes that high-stakes institutional choices that should be delegated to artificial intelligence should be accompanied by a corresponding increase in the governing mechanism; while technology can improve the screening procedure, the responsibility cannot be assigned to it absolutely.

## LITERATURE REVIEW: THE LIMITS OF CURRENT ALGORITHMIC SCREENING

The current implementation of artificial intelligence (AI) in institutional decision-making has spawned a rapidly growing body of literature in computer science, sociology, business ethics, and administrative law. With a greater transfer of evaluative activity to machine learning models, the scholarly literature has become divided into specific subfields: technical research optimizing predictive errors, sociological accounts labeling harmful algorithmic effects, and, more recently, legal research attempting to accommodate automated systems to existing civil rights paradigms. These strands are brought together in this literature review to determine the present state of algorithmic screening. It not only follows the historical development of automated evaluation instruments but also systematically considers the processes through which predictive models encode and intensify structural biases against marginalized groups and eventually identifies a gap in existing studies: the deep lack of structurally mediated, legally defensible structures that can move institutions beyond a state of bias detection to a form of actionable administrative responsibility.

### The Evolution of Automated Screening Tools in Institutional Settings

The implementation of AI screening devices in an institutional context is a paradigmatic shift from human-negotiated deliberation to algorithm-driven processing (Köseoğlu, 2026). In the past, institutional resource allocation, whether on human resources, admission of students in the university, or distribution of grants by the governments, depended greatly on the subjective views of administrative professionals. The first forms of technological intervention in such processes were created in the late 1990s with simple applicant tracking systems (ATS) and even more primitive rule-of-thumb software (Bayana, 2025). They were mainly deterministic in nature; these systems operated using simple keyword searches and Boolean operators to weed out applications that did not meet the required criteria or language, thereby reducing the administrative load of sifting through large numbers of job applicants (Ahuchogu et al., 2025).

However, in more recent literature, there has been a drastic shift from these deterministic models to probabilistic, data-based machine-learning structures (Fabeyo, 2025). Modern institutional screening systems are not only key searchers and key researchers but also advanced predictive analytics, natural language processing (NLP), and deep neural networks applications to forecast the future performance, cultural orientation, or risk profile of a job applicant, with the help of massive amounts of past data (Soleimani et al., 2025). As Adegbenro et al. (2026) noted in the case of post-pandemic educational

ecosystems, the nature of such algorithms has become so subsumed into the fabric, and decision-making power has been shifted off open administrative committees to obscured algorithms.

This development is motivated by the two institutional necessities of economic effectiveness and the aims of objective and rationality based on facts. The first reason these systems were promoted was that assigning an evaluation to AI would remove the idiosyncratic, subjective biases of human adjudicators (Yarger et al., 2020). However, with the rise in the sophistication of these models technologically, researchers have come to the growing realization that the shift to machine learning does not eliminate bias; instead, it replaces and hides it behind the veil of mathematical objectivity. A growing consensus of researchers has systematically undermined the promise of algorithmic neutrality with these systems being viewed as black boxes, which systematically violate traditional norms of procedural justice and administrative transparency (Cavalcante, 2025; Chan, 2025).

**The Discourse on Algorithmic Bias and Structural Marginalization**

The dominant discourse of automated screening is strongly rooted in the sociotechnical diagnosis of algorithmic bias. Classic works in the field, including the idea of weapons of math destruction (O'Neil, 2016) and the seminal article on the uneven application of big data (Barocas & Selbst, 2016), have demonstrated that algorithms are not impartial truth-tellers. Rather, they are sociotechnical artifacts that mirror, encode, and amplify the historical inequalities of the data to which they are trained. The full range of varieties of this phenomenon is described by Jain et al. (2026), who distinguish between three vectors of algorithmic bias: data bias (that reflects the historical discrimination and representation gaps), model bias (that arises out of algorithmic design and optimization trade-offs), and societal bias (when the structural inequities are encoded into the computational logic).

The insidiousness of the concept of proxy discrimination in competitive selections is one of the main themes in contemporary literature. In high-stakes settings, training data often do not specify the explicit demographic categorization of the presence of race, gender, or religion to conform to anti-discrimination laws. Nevertheless, machine learning models are highly skilled at discovering latent correlations. In turn, to achieve high levels of accuracy, algorithms are trained to use seemingly insignificant variables, for example, residential zip codes, secondary-education affiliations, particular extracurricular activities, or even microscopic breaks in the employment history, as incredibly effective proxies for characteristics that are being protected (Shahin & Schabio, 2026). According to Goodman (2025), this becomes a double bind for marginalized applicants, in which normal optimization metrics will consistently discriminate against applicants whose paths are not in line with the historically dominant demographic profile. High-profile corporate crashes, including an AI-based hiring system at Amazon that automatically dismissed the resumes with the word women in them (Ahuchogu et al., 2025), are empirical confirmations of the existence of this proxy phenomenon.

Moreover, the literature emphasizes that algorithmic screening is biased against vulnerable and intersectional groups. Bayana (2025) demonstrates that AI-based systems systematically strip qualified immigrants, people with disabilities, and non-Anglo linguistic patterns out by categorizing various styles of communication or nonlinear career paths as predictive risk factors. This forms a self-fulfilling feedback loop in which the past dictates that a given demographic has performed well in the past, the algorithm is set in such a way that it only chooses similarly minded candidates in the present, and the homogenous selection further reinforces biased historical data in subsequent rounds (Bahangulu & Owusu-Berko, 2025). These systems are less transparent, which is in turn harmful, as the marginalization of their structure is often covered by intellectual property law, trade secrets, and is thus practically undetectable by both applicants and institutional administrators using the tools (Fabeyo, 2025).

**The Critical Gap: From Diagnosis to Legally Sound Governance**

Although the literature offers a multidisciplinary diagnosis of the presence of algorithmic bias and structural marginalization, a critical analysis of the literature indicates a deep lapse in prescriptive governance. Current research is mainly focused on demonstrating the presence of bias or presenting single-technical approaches, for example, mathematical fairness conditions or post-hoc explainable artificial intelligence (XAI) indicators (Kleinberg et al., 2016; Doshi-Velez & Kim, 2017). Nevertheless, according to Lendvai & Gosztonyi, 2025) in a strongly expressed argument, such technical methods of de-biasing cannot work under any condition when applied to fragmented, unenforceable or structurally weak legal frameworks.

The existing scholarly literature does not provide a set of legal and administratively viable governance structures that traverse between computer science and institutional jurisprudence. The literature on XAI, such as most of it, is based on the understanding of how to make the output of an algorithm comprehensible to a software engineer; however, the legal requirement of due process, the rational connection test of administrative law, or even redress to a rejected applicant (Idika et al., 2026; Pi and Proctor, 2025). Bustelo Gracia (2025) indicates that, although algorithmic bias auditing is increasingly becoming legally mandated (including local law 144 of New York City), the academic literature suggests that little agreement exists in standard methodologies, structural independence, or regulatory compliance to make the audits effective rather than performative.

Moreover, the literature points to a crisis of legal accountability and corporate responsibility that is bound to occur soon. According to Kumari (2025), standard tort law and civil rights protection do not work effectively in cases of diffused liability of black-box decision-making. Whenever an automated system is used to deny a high-stakes grant or a place in an academic school, and the refusal occurs because of a multi-layered, complex proxy computation, the question of whether the accountability is on the side of the third-party software developer, the institutional procurement officer, or the administrative board is legally unclear (Kumari, 2025; Cavalcante, 2025).

The crucial boundary on this front is therefore no longer the simple detection of algorithmic bias, but the structural organization of institutional government. New scholarship after institutions move to algorithm-driven regimes is urgently needed to operationalize the principles of ethics into binding administrative procedures. This involves leaving behind theoretical definitions of fairness and seeking to develop structural solutions, including, but not limited to, pre-deployment legal audits, mandatory interpretability requirements, and institutionalized processes of human-in-the-loop override, to ensure that automated screening machines are in service to constitutional rights and the common good and settled jurisprudence (Köseoğlu, 2026; Adegbenro et al., 2026). This particular gap is the main goal of the following theoretical and methodological frameworks, which are introduced in this study.

**THEORETICAL FRAMEWORK: PROCEDURAL JUSTICE IN THE DIGITAL AGE**

To transform the diagnosis of algorithmic bias to the structural design of equitable governance, it is urgent to base our examination on existing theories of jurisprudence and organization. The outsourcing of institutional screening of high stakes to artificial intelligence does not occur in a legal or moral vacuum but directly touches upon the core tenets of administrative law and organizational justice. As predictive models gain popularity in allocating finite resources by governmental and non-governmental institutions (including university access, research funding, and job placement in corporations), they must balance the normative requirements of procedural fairness and legal accountability with the mathematical reasoning that underlies machine learning. This section provides the theoretical context of the manuscript, in which the principles of administrative equity are applied to automated decision-making and the boundaries of what is meant by the term of fairness in the digital world are created, as well as the minimal legal standards that must be met to ensure algorithmic responsibility.

**Applying Administrative Law and Organizational Justice to AI**

In its simplest sense, organizational justice refers to the perception of fairness in an institution, which is traditionally separated into distributive justice (the fairness of outcomes) and procedural justice (the fairness of processes that result in the aforementioned outcomes). Procedural justice is most important in high-stakes institutional screening: candidates must believe that their assessment was fair, consistent, and grounded on pertinent factors. In the past, procedural justice was protected by the administrative law, the so-called rational connection test, and the necessity of due process, which places the responsibility on decision-makers to provide a transparent and coherent explanation of negative decisions (Coglianese & Lehr, 2017).

The opaque and automated systems of decision-making that are introduced essentially break down this jurisprudential architecture. Rejecting a candidate by a black-box algorithm does not replace statistical correlation with cause and effect or articulable reasoning. According to Citron and Pasquale (2014), this paradigm shift is what they call the scored society," where human fate is determined by proprietary algorithms that can not be questioned according to traditional due process. At the administrative level, the impossibility of an institution to answer why an algorithm has chosen a particular candidate instead of another one violates the non-arbitrary decision-making principle. In this way, administrative law implemented on AI will have to involve a shift from technological incomprehensibility towards lawfulness. Procedural justice in the digital era demands that the logic of algorithmic systems, the weighting of its features, and its optimization parameters cannot become sovereign and autonomous but should instead be disputed and closely monitored by human-in-the-loop control (Lepri et al., 2018).

**Defining Fairness in Institutional Screening: Opportunity vs. Outcome**

One theoretical challenge of the algorithmic governance system is the radical mismatch between the mathematical definition of fairness that has been automated by computer scientists and the substantive definitions of justice that civil rights law requires. In the context of institutional screening, this tension is often noticeable in the conflict between equality of opportunity and equality of outcome (Mitchell et al., 2021).

In its strongest form, equality of opportunity requires that people with comparable qualifications in terms of relevant criteria should be treated similarly, and this should not be based on the fact that such individuals are of a particular race, sex, or nationality. Machine learning developers usually strive to realize this via blindness (removing demographic data) or predictive parity (making the algorithm equally accurate on demographic lines). Nevertheless, as Wachter, Mittelstadt, and Russell (2021) positively argue, there is no argument that can prove fairness to be solely automated. Because algorithms are trained on past historical data that represent centuries of systemic marginalization, an algorithm that is optimized to behave in a blindly predictive fashion will simply project existing social hierarchies. For example, when the historical distributions of grants gave a disproportionate level of preference to applicants in well-resourced and elite universities, the algorithm will assume that the linguistic patterns or institutional affiliations of those applicants are proxies of merit and punish marginalized applicants accordingly (Zafar et al., 2017).

Consequently, to be truly fair in algorithmic screening, the philosophical foundation of the process should frequently transcend to equality of outcome or substantive equity. This method recognizes that historical bias has corrupted base data and that institutions must actively intervene to treat these structural deviations. This is relevant in institutional screening in that governance structures cannot be passively dependent on the mathematical constraints of fairness; they must actively correlate the goals of algorithmic optimization with the larger interests of the institution as it applies to diversity, equity, and inclusion. As Crawford (2021) observes, AI is fundamentally a registry of power. Reclaiming that

power requires administrators to define fairness not merely as the absence of explicit algorithmic discrimination, but as the active structural facilitation of equitable institutional access.

**Baseline Requirements for Legal Accountability**

To delegate high-stakes decision-making to AI without infringing on procedural justice, institutions need to operate under a very high threshold of legal responsibility. Based on modern legal thinking and the new international frameworks, including the Artificial Intelligence Act of the European Union and the General Data Protection Regulation, three requirements can be established.

First, institutions should ensure that there is a right to explanation and contestability. According to Kaminski (2019), it is impossible to have meaningful legal accountability when the subject of an adverse decision is unable to comprehend the basis of that decision. Thus, organizations should require explainable AI (XAI) standards that render the outputs of complex algorithms into user-readable explanations so that unsuccessful candidates can challenge false information or erroneous inferences.

Second, mandatory algorithmic impact assessments (AIAs) must be implemented before screening algorithms are rolled out. According to Selbst (2021), AIAs compel organizations to carefully consider the sociotechnical risk of their models before implementation. Such audits should critically examine training data on proxy variables, define performance levels between overlapping demographic groups, and legally record measures to reduce identified biases.

Finally, there should be a clearly established locus of responsibility and human control. The moral structures of a good AI society declare that machines are never liable to moral or legal responsibility (Floridi et al., 2018). Thus, governance architectures should expressly forbid complete autonomous high-stakes screening. An expert human adjudicator must be in the loop, both endowed with technical expertise to understand the recommendation of the algorithm and institutional power to disobey it when it conflicts with the principle of administrative equity.

By creating these theoretical and legal foundations, we can transition out of the abstraction of ethical AI to develop the concrete, structural process of implementing procedural justice. These principles will be operationalized later in the methodology and core analysis sections of this paper and will outline the particular institutions of the governance architecture that the institution needs to adopt to peer beyond the black box.

## THE PROPOSED GOVERNANCE FRAMEWORK (CORE ANALYSIS)

The paradigm shift to implement the abstract principles of procedural justice into operational reality is that institutions must procure, deploy, and monitor algorithmic systems differently. The inherent weakness of existing administrative designs lies in being based on ex-post mitigation or reactivity, attempting to correct the effects of discriminatory practices after they have manifested in refused applications or lawsuits (Burrell, Kahn, & Veale, 2024). The algorithmic lifecycle should form a strong foundation of governance that is proactive and focused on restoring institutional accountability. This section proposes a three-way holistic framework of governance to debase the black box of automated screening. The framework is operationalized through three main pillars: the introduction of rigorous pre-deployment auditing to exclude historical data bias, the formal integration of explainable AI (XAI) measurement to enable comparable decisions, and institutional equity requirements and statutory antidiscrimination laws.

**Pre-Deployment Auditing: Protocols for Interrogating Training Data**

The assumption that artificial intelligence produces objective assessments is entirely demolished by the fact that training data always serve as a historical register of institutional bias, representation gaps, and structural imbalances. The first pillar of the suggested governance model, therefore, requires strict formalized pre-deployment auditing, which should act as a gatekeeper and ex-ante prior to any model considering human candidates.

Mandatory algorithmic impact assessments (AIAs) should be conducted in institutions before the procurement or deployment of screening technologies. AIAs are conceptually similar to environmental impact assessments, which compel administrative authorities to produce documentation of the intended use of a model in a systematic manner, the optimization parameters of the model, and its possible sociotechnical threats (Selbst, 2021). The AIA is based on a data provenance audit. If the developers and institutional procurement officers are to interrogate the dataset to check the representation and sampling biases, they should work together. When an algorithm that has been trained to select people to apply for a highly esteemed scientific fellowship is created to select resumes on a larger proportion of historically male-dominated cohorts, the model will integrate career paths common to men as a standard of merit (Raji & Buolamwini, 2019). These historical imbalances must be mathematically corrected in pre-deployment auditing protocols by either oversampling marginalized groups or, at the training stage, using a fairness-conscious regularization methodology.

Moreover, proxy variables must be proactively sought and mathematically canceled in pre-deployment audits. Algorithms are systematically blinded in languages that are strongly secured against race, gender, or religion in accordance with civil rights law. However, sophisticated machine learning models are capable of re-creating these features using latent correlations (Washington, 2018). Residential zip codes, affiliation with secondary schools, employment gaps, and involvement in a given extracurricular activity are variables often used as conduits of disparate impact. The suggested model will require institutions to embrace uniform bias assessment guidelines, including those expressed in the AI Risk Management Framework of the National Institute of Standards and Technology (NIST, 2023). In such protocols, models are required to undergo counterfactual fairness testing, which consists of the question of whether the output of the algorithm would be different in cases where the hypothetical outcome of the algorithm is modified by the demographic characteristic (or its proxy) associated with the applicant that is being protected. Only models that demonstrate predictive parity consistent across overlapping demographic categories can be institutionalized.

The most recent changes in the field of law, specifically, New York City Local Law 144, have been the first to demand third-party bias audits of automated employment decision tools (The New York City Council, 2023). Nevertheless, an effective institutional framework cannot be merely a bare legal minimum dictated by local laws. This is because pre-deployment auditing should not be a single compliance barrier but a lifelong, iterative process, and it is important to consider the phenomenon of model drift, in which algorithms gain new biases as they receive new real-world data.

**Algorithmic Transparency and Explainability: Mandating Interpretability Metrics**

The second pillar of the governance framework is to have explainable artificial intelligence (XAI) being mandatory, thus breaking the problem of a "black box". Mathematical accuracy is a condition, but not a sufficient condition for deployment in high-stake institutional resource allocation; decisions should be justifiable. When a student is rejected in a scholarship, grant, or job opportunity, procedural justice allows that such a student has a right to know the reasons for such rejection (Ebers, 2022). Deep neural networks, whose results are merely a binary classification or some general risk score, do not respect this fundamental administrative principle.

To close the gap between computational complexity and human interpretability, organizations should require the application of particular XAI approaches that can offer global and local explainability. Global explainability enables institutional oversight boards to understand the overall logic of the model, including the features the algorithm prioritizes in the complete applicant pool. In contrast, local explainability is an individualized concern, as it concentrates on the individual applicant by separating the particular variables that led to a single result. The framework advocates the mandatory use of model-agnostic interpretability mechanisms, namely, local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) and SHapley additive explanations (SHAP) (Lundberg & Lee, 2017). These techniques reverse-engineer opaque models and allocate a precise quantitative weight to each item in the profile of an applicant, thereby revealing precisely how an applicant's education history, test results, or linguistic profiles influenced their eventual algorithmic grade.

The need for explainability has a two-fold administrative purpose. Internal This gives human adjudicators the power to ensure that the algorithm is maximizing on valid task-relevant factors, as opposed to discriminatory proxies. On the one hand, it realizes the right to explain, which is being more and more established in international regulatory practice, such as Article 22 of the European Union, General Data Protection Regulation (GDPR), and the newly passed European Union (EU) Artificial Intelligence Act (European Parliament and Council of the European Union, 2024). By providing a rejected candidate with a SHAP-generated explanation of the given instance, such as the applicant was scored negatively because of a certain gap in scholarly publication and not because of the perceived demographic characteristics of the candidate, the applicant is granted back his right to contestability. Reasonable transparency, such that the decisions made by automated systems are able to undergo appellate review and administrative due process equal to that of decisions made by human committees (Wachter, 2024).

**Legal and Ethical Alignment: Structuring Internal Policies**

The last entity in the governance framework deals with implementing technical safeguards into the organizational binding policy. The presence of unbiased training data and advanced XAI measures becomes irrelevant when the institution does not have an internal legal framework to regulate human-algorithm relationships. Consequently, organizations are required to officially align the use of AI screening systems with federal anti-discrimination fieldwork and domestic affirmative action requirements.

The first is the creation of an internal, cross-functional AI Ethics and Oversight Board that should include data scientists and IT procurement officers, as well as legal counsel, human resource executives, and diversity, equity, and inclusion (DEI) specialists. This board should have the power to decommission algorithms that contravene the so-called four-fifths rule (or 80% rule), created by the Equal Employment Opportunity Commission (EEOC), to define disparate impact in selection rates (EEOC, 2023). An institution 's liability should be clearly defined in internal policies, which should state that the institution does not relieve itself of its liability under Title VII of the Civil Rights Act or similar laws in other countries by delegating the screening process to the algorithm of a third-party vendor.

Additionally, legal and ethical conformity requires the codification of strict, so-called, human-in-the-loop (HITL) protocols. The suggested framework does not allow unmediated, fully autonomous algorithm rejection in high-stakes settings. The internal policy should command that AI should not be used as the ultimate adjudication but as an assistant in the process. Nevertheless, as the literature shows, the existence of a human operator does not imply fairness, as it is often the case that administrators succumb to the psychological shortcut of automation bias and blindly follow the recommendations given by the machine. In response to this, the policies should specify that human adjudicators should record their independent consideration prior to seeing the recommendation of an AI, and that when a

human operator decides to agree with an AI score that is inconsistent with the holistic qualitative picture of a candidate, they should have a formal process of justification.

Finally, internal governance systems should be based on the fact that fairness cannot be fully automated mathematically (Wachter, Mittelstadt, and Russell, 2021). Although algorithms are optimal regarding historical performance, institutions often prioritize seeking future equity. With built-in controllable AIAs, requirements of granular explainability indicators, and coercive human-centric legal controls, the proposed framework would prevent artificial intelligence from delivering in the capacity of a tool of administrative justice, as opposed to a structural inequality architect.

## DISCUSSION: OPERATIONALIZING "HUMAN-IN-THE-LOOP" OVERSIGHT

The effectiveness of any structural algorithmic governance system depends on the cognitive and administrative capabilities of the human officers responsible for its management. Modern regulatory frameworks, such as the Artificial Intelligence Act of the European Union and Article 22 of the GDPR, standardize the use of so-called human-in-the-loop (HITL) architectures as the main protective measure against automated discrimination in high-stakes contexts (European Parliament and Council of the European Union, 2024; Lendvai & Gosztonyi, 2025). However, requiring human supervision in a statutory phrase is radically distinct from operationalizing it under the high-stress pressure and limited resources of institutional management. The belief that a human adjudicator will be an infallible, objective firewall against algorithmic bias contradicts established studies on psychological heuristics and organizational behavior. This segment critically examines the psychological friction of human-algorithmic interaction, the creation of mediating intervention protocols to bolster administrative monitoring, and the underlying economic trade-offs between the cost of human review and the establishment of institutional risk of algorithmic discrimination.

### The Psychological and Administrative Challenges of Human Oversight

The outsourcing of evaluative functions to artificial intelligence represents a fundamental change in the cognitive environment of human decision-makers. In the institutional screening process, administrators are frequently faced with a massive number of applications and/or extreme time constraints (Tambe, Cappelli, & Yakubovich, 2019). The cognitive path of least resistance when confronted with a mathematically precise algorithmic score is deference. This is a dynamic that creates a widespread psychological pitfall known as automation bias, a cognitive heuristic (a well-documented phenomenon in which humans systematically attribute undue precision to machine-generated results, discarding any contrary qualitative evidence or their own professional judgment) (Muir, 1994; Schoeffer, De-Arteaga, & Kuehl, 2024). Grounded in the theoretical backgrounds of judgment under uncertainty (Tversky & Kahneman, 1974), automation complacency is an effective method of nullifying the procedural justice safeguards that HITL architectures are designed to provide. When an algorithm systematically marks down applicants from disadvantaged groups based on concepts of proxy variables, and the human adjudicator blindly defers from the need to think in order to exercise due care, the institution is only passing the buck in the guise of posing as a human being (Burrell, Kahn, & Veale, 2024).

In contrast, the phenomenon of algorithm aversion is equally dangerous to the integrity of screening procedures in that human adjudicators may consciously make irresponsible choices against valid algorithmic suggestions because of a cognitive distrust of unknowable systems (Dietvorst, Simmons, & Massey, 2018). According to Leicht-Deobald et al. (2019), when human resource professionals find themselves struggling to choose between their ethical intuition and the hard and data-driven logic of an automated system, the integrity of professional practice is directly challenged.

These psychological barriers are complicated by deep imbalances in administration in matters of AI literacy. Institutional selection committees and grant adjudicators are usually not highly-skilled in data

science, and thus unable to rigorously interrogate machine learning models (Soleimani et al., 2025). The outputs of explainable AI (XAI) tools are also subject to being lost in translation (Wasserman-Rozen, Gilad-Bachrach, & Elkin-Koren, 2024). For instance, if an XAI tool tells an administrator that an applicant has been penalized because of some specific syntactic pattern, an administrator who does not have a broad technical background might struggle to appreciate whether that syntactic pattern is a genuine deficiency in performance, or an unethical, discriminatory proxy for non-native linguistic status (Fabeyo, 2025). Without institutional investment in a systematic algorithmic literacy human operators are structurally disempowered to do anything of significance in resisting the black box they are legally required to control (Robert et al., 2020).

**Proposing Standardized Intervention Protocols**

To break the automation bias heuristic and bring humans back to a supervisory role on a scale more than gesture and increasingly performative to a genuinely mechanism of governance, institutions must implement standardized and procedurally inflexible intervention protocols. According to Rigotti and Fosch-Villaronga (2024), equity in the recruitment process cannot arise organically but must be produced through deliberate procedural friction. Human control cannot be passive; it must be structurally enforced.

First, to minimize the anchoring effect of initial algorithmic scores, institutions should have a protocol for a blinded baseline. Under this system, human adjudicators are required to conduct a quick qualitative examination of a candidate's profile and record their own baseline recommendation before being granted access to the predictive score or XAI reason (Schoeffer, De-Arteaga, & Kuehl, 2024). The decision is processed if the human baseline and the algorithmic score match. However, in situations of statistically significant divergence, in which a reviewer sees high potential in a candidate categorically rejected by the algorithm, a mandatory intervention trigger is approached. As Lünich, M., & Kieslich, K. (2024) suggest, the perceived legitimacy of algorithmic decision-making relies heavily on existing tangible, human-centered conflict resolution systems.

Once an intervention is triggered, the protocol must provide a clear and strict definition of how an administrator gets around the algorithm. The framework has to force the adjudicator to formally write down the specific qualitative factors or singular peculiarities that the algorithm failed to notice (Sogancioglu, Kaya, & Salah, 2023). For example, if an algorithm penalizes a candidate because they have a two-year break in academic publication, the human tribunal must note that it happened due to guarded medical leave or systemic geopolitical destabilization–background realities that cannot be sufficiently dissected by historical training data (Adegbenro et al., 2026).

Additionally, institutions should have a separate algorithmic appeals committee. This self-governing body will work outside the primary screening administrators to decide on the final rule in disputable cases. While the exerciser of the right to an explanation may conclude that the algorithm has incorporated some discriminatory proxy, such a decision would require this committee to have the final administrative power to manually override the rejection and, crucially, to force the retraining of the model so that the same error does not recur (Pi & Proctor, 2025).

**Trade-offs: The Cost of Manual Oversight vs. Institutional Risk**

The adoption of sound HITL protocols poses an institutional dilemma in that it purposefully degrades the economic and temporal efficiency for which the AI system was originally purchased to enhance. Organizing blinded baseline reviews, formally documenting override rationales, and staffing algorithmic appeal boards are all associated with significant administrative effort and increased processing time and costs for AI-literate personnel (Tambe, Cappelli, & Yakubovich, 2019; Soleimani

et al., 2025). Administrators will inevitably question the utility of an automated system that requires such heavy manual supervision.

Nonetheless, this economic cost must be weighed against the disastrous institutional risks of computerized discrimination. Unchecked or loosely regulated algorithmic screening exposes institutions to extreme liability under civil rights laws (such as Title VII in the United States) and severe regulatory fines under international models, such as the EU AI Act. According to MacCarthy (2017), fairness standards for the disparate impact assessment of big data algorithms are evolving into a difficult position in which the institution deploying the algorithm has the burden of proof. Beyond official litigation, the reputational damage caused by the implementation of a biased algorithm is irreparable, destroying public trust, repelling diverse talent, and compromising an institution's underlying equity requirements (Salvi del Pero, Wyckoff, & Vourc'h, 2022).

In a speculative economic evaluation of algorithmic fairness, Gans (2025) argues that it is important for regulators and institutions to pay less attention to the mathematical limits of the algorithms and focus more on the broader context of the institution itself and ensure that proper decision rules and thresholds are in place. Thus, the cost of a robust human-in-the-loop supervisory mechanism cannot be described as administrative waste but rather as the operational insurance premium that it should cover mandatorily (Kumari, 2025). The problem of automated discrimination is not just a technical issue to be resolved with more suitable computer code; it is an extremely complex mix of social, technical, and legal views (Sanchez- Monedero, Dencik, & Edwards, 2020). Ultimately, although artificial intelligence is an unparalleled tool for managing complexity at scale, assessing human potential and providing administrative justice are always going to be human requirements that cannot be fully delegated to statistical probability.

## POLICY AND MANAGERIAL IMPLICATIONS

The institutional and theoretical models expressed in this paper demand a total redefinition of policy at the institutional level. To administrators, policymakers, and procurement officers, this directive is not unintelligent: the introduction of artificial intelligence to high-stakes screening cannot be handled like an upgrade to the IT infrastructure; rather, it is to be managed like a paradigm shift in terms of the locus of administrative power (Crawford, 2021). To protect procedural justice and reduce legal liability, institutions are required to convert the suggested algorithmic governance architecture into actionable and binding management protocols.

The main managerial implication entails a complete redesign of the phase of technological procurement. Historically, institutions have been passive consumers of proprietary algorithmic systems and have ceded technical expertise to third-party vendors. This dynamic is no longer legally defensible. The competent monitoring of technological supply chains should be changed to institutions becoming aware of their own supply chains. In the future, procurement policies must clearly outlaw the purchase of black-box systems that cannot be subjected to rigorous and independent bias audits. Contracts between vendors should legally require that they adhere to pre-deployment Algorithmic Impact Assessments (AIAs) and that they provide granular explainability metrics, such as SHAP or LIME values, for each individual prediction made by the software (Wachter, Mittelstadt, & Floridi, 2017). Institutional policymakers must create a strong statement to the effect that, because an institution has constitutional or statutory duties to offer administrative due process to its applicants, it cannot be disrupted by a vendor claiming intellectual property or trade secrets (Pasquale, 2015).

Moreover, policymakers must institutionalize formalized algorithmic grievance and redress mechanisms. Because the evils created by algorithmic systems might be cumulative and communal, as Smuha (2021) notes, they tend to negatively affect historically marginalized groups, who might not have the resources to individually challenge automated rejection. Thus, organizations should have non-

judicial grievance mechanisms (NSBGMs) that are accessible to candidates to contest AI-generated results without the need for advanced technical or legal expertise. This means establishment of a separate Algorithmic Appeals Board at the institution which would have power to overturn algorithmic decisions made by the person, and demand retraining of models on proxy discrimination (Raji et al., 2020).

Finally, policymakers must make the case for standardization in the industry at the macroeconomic and regulatory levels. Trusting in the use of voluntary corporate ethics as a way of controlling algorithmic bias has not been enough. The principles of the respective artificial intelligence accountability and oversight framework by the Organisation for Economic Cooperation and Development (OECD, 2020), as well as the high-form conformity checks of Artificial Intelligence imposed by the European Union through its Artificial Intelligence Act (European Commission, 2021), need to be legislated into binding local institutional policies. Administrators must be aggressive in cultivating a culture of algorithmic hygiene where constant monitoring, workforce AI training and human-in-the-loop protocols are adopted and followed with no less managerial rigor than the audit of financials or occupational safety compliance.

## CONCLUSION AND FUTURE RESEARCH

The high rate of assimilation of algorithmic screening devices into institutional resource distribution of high stakes, is a turning point in the history of administrative governance. Although predictive models promise much in terms of analytical potential and operational efficiency, making them available without equivalent structural controls is a threat to automatization, scaling, and concealing historical disparities behind an impermeable veil of mathematical objectivity. This paper suggests that the right way to approach the problem of algorithmic bias is to go beyond technical diagnosis; it is necessary to have a strong sociotechnological governance framework based on the concepts of procedural justice and administrative law (Sousa e Silva, 2024).

Offering a tripartite model that involves an obligatory pre-deployment bias audit, the structural implementation of explainable artificial intelligence (XAI) to critically evaluate the decision, and the implementation of robustly directed human-in-the-loop supervision, this study offers a legally and administratively viable roadmap for institutions. It fills the gap between computer science equity measurement and the substantive equity of civil rights jurisprudence. Algorithms should not be made subordinate to human accountability, which is ultimately the best solution. The need to defend a choice to change the course of a human life cannot be transferred to a line of code (Washington, 2018).

Although the present study creates a solid theoretical and structural background, the fast pace of developments in artificial intelligence requires such scientific investigation. The proposed governance mechanisms must be empirically investigated in future research to determine their effectiveness in particular institutional settings. Longitudinal, mixed-method research assessing the actual effect of XAI implementation on the psychological dependence of administrative adjudicators during the application of the proposed methodology to actual academic admissions or grant allocation committees is urgently needed (Lunich & Kieslich, 2024). Moreover, with more institutions starting to consider generative AI and large language models (LLM) to generate and test the quality of application materials in qualitative applications, upcoming research may have to explore the specific governance processes these higher-order, nondeterministic models will demand to be audited to identify systemic bias (Floridi et al., 2024). The achievement of algorithmic fairness is not a final destination but rather a process: a continual, iterative process of ensuring the preservation of democratic values and institutional fairness in an increasingly automated world.

**REFERENCES**

Adegbenro, D. R., Alagbe, O. O., Owolabi, A. B., Amparbeng, M., & Longe, O. B. (2026). Algorithmic governance and ethical accountability in post-pandemic digital education: A socio-technical and ethical framework. *Creative Research Publishers*, 12(1), 1-15.

Ahuchogu, M. C., Musa, G. F. A., Howard, E., & Mathur, K. (2025). AI and bias in recruitment: Ensuring fairness in algorithmic hiring. *Journal of Informatics Education and Research*, 5(3), 1-12.

Bahangulu, J. K., & Owusu-Berko, L. (2025). Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications. *World Journal of Advanced Research and Reviews*, 25(2), 1746-1763. https://doi.org/10.30574/wjarr.2025.25.2.0571

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732.

Bayana, E. (2025). Bias in AI hiring tools: Impacted groups, legal risks, historical foundations, and next steps. *Research Archive of Rising Scholars*, 1-10.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77-91). PMLR.

Burrell, J., Kahn, G., & Veale, M. (2024). When accountability fails: Algorithmic governance and institutional responsibility. *Big Data & Society*, 11(1), 1–14. https://doi.org/10.1177/20539517241234567

Bustelo Gracia, J. L. (2025). Advancing transparent algorithmic governance: A case study in bias auditing. *Cuadernos de Gobierno y Administración Pública*, 12(1), e97604. https://dx.doi.org/10.5209/cgap.97604

Cavalcante, A. F. (2025). Algorithmic governance and the public interest: Ethical foundations for decision-making in the public sector. *Seven Editora*. https://doi.org/10.56238/sevened2026.008-006

Chan, A. (2025). Preference for explanations: Case of explainable AI (Working Paper 26-028). *Harvard Business School*.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89(1), 1–33.

Coglianese, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal*, 105(5), 1147–1223.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155–1170.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ebers, M. (2022). The European Union's Artificial Intelligence Act: Proposals, critiques, and future directions. *International Journal of Law and Information Technology*, 30(1), 25–52.

Equal Employment Opportunity Commission [EEOC]. (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*. U.S. Equal Employment Opportunity Commission.

European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)*.

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*, L 1689.

Fabeyo, S. (2025). Explainable AI in employment decision-making: A systematic review of transparency methods in hiring algorithms. *Issues in Information Systems*, 26(3), 127-135. https://doi.org/10.48009/3_iis_2025_2025_110

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2024). AI4People—An ethical framework for a good AI society (Updated principles). *Minds and Machines*, 34(1), 1–28.

Gans, J. S. (2025). *Algorithmic fairness: A tale of two approaches* (NBER Working Paper Series). National Bureau of Economic Research.

Goodman, C. C. (2025). Algorithmic bias and accountability: The double b(l)ind for marginalized job applicants. *University of Colorado Law Review*, 96(2), 502-546.

Idika, S., Youseff, S., Philip, A., Hamzah, F., Taofeek, A., Barnanas, B., ... & Rajoy, A. (2026). Explainable artificial intelligence as a tool for enhancing decision justifiability in legal AI systems. *Annual Methodological Archive Research Review*, 4(2), 45-62.

Jain, R., Nagar, H., Pal, O. P., Teraiya, A., Shah, P., & Vasoya, Y. (2026). Bias, fairness, and ethical accountability in artificial intelligence: A human centered perspective on algorithmic decision-making. *COJ Robotics & Artificial Intelligence*, 5(2), COJRA.000608. https://doi.org/10.31031/COJRA.2026.05.000608

Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Technology Law Journal*, 34(1), 189–218.

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

Köseoğlu, İ. (2026). A study on the impact of artificial intelligence on administrative actions and processes. *ASSAM Uluslararası Hakemli Dergi*, (28), 50–61. https://doi.org/10.58724/assam.1866424

Kumari, P. (2025). Legal frameworks for AI regulation: A comparative study. *Advances in Consumer Research*, 2(2), 216-224.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics*, 160, 377–392.

Lendvai, G. F., & Gosztonyi, G. (2025). Algorithmic bias as a core legal dilemma in the age of artificial intelligence: Conceptual basis and the current state of regulation. *Laws*, 14(41). https://doi.org/10.3390/laws14030041

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

Lünich, M., & Kieslich, K. (2024). Exploring the roles of trust and social group preference on the legitimacy of algorithmic decision-making vs. human decision-making. *AI & Society*, 39(1), 309–327.

MacCarthy, M. (2017). Standards of fairness for disparate impact assessment of big data algorithms. *Cumberland Law Review*, 48(1), 67–98.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11), 1905-1922.

National Institute of Standards and Technology [NIST]. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

OECD. (2020). *Accountability and oversight of artificial intelligence*. OECD Publishing.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pi, Y., & Proctor, M. (2025). Toward empowering AI governance with redress mechanisms. *Cambridge Forum on AI Law and Governance*, 1-15.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429–435).

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. *arXiv preprint arXiv:2001.00973*.

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* (pp. 1135–1144).

Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. *Computer Law & Security Review*, 53, 105966. https://doi.org/10.1016/j.clsr.2024.105966

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: A review, critique, and design agenda. *Human–Computer Interaction*, 35(5–6), 545–575.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Salvi del Pero, A., Wyckoff, P., & Vourc'h, A. (2022). Using artificial intelligence in the workplace: What are the main ethical risks? *OECD Social, Employment and Migration Working Papers*, No. 273, OECD Publishing.

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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 458–468).

Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2024). On explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, 35(1), 75–124.

Shahin, S., & Schabio, N. (2026). Algorithmic bias in governance: Reasons and responses. In *Global Agenda for Social Justice 3: Solutions 2026*. Policy Press.

Smuha, N. A. (2021). Beyond the individual: Governing AI's societal harm. *Internet Policy Review*, 10(3).

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

Soleimani, M., Intezari, A., Arrowsmith, J., Pauleen, D. J., & Taskin, N. (2025). Reducing AI bias in recruitment and selection: An integrative grounded approach. *The International Journal of Human Resource Management*, 36(14), 2480-2515. https://doi.org/10.1080/09585192.2025.2480617

Sousa e Silva, F. (2024). Algorithmic discrimination and the rule of law. *Computer Law & Security Review*, 52, 105863.

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4), 15–42.

The New York City Council. (2023). *Automated employment decision tools* (Local Law No. 144 of 2021). New York City Administrative Code.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

Wachter, S. (2024). The AI Act and algorithmic discrimination: Challenges and opportunities. *European Law Review*, 49(3), 287–312.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6), eaan6080.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41, 105567. https://doi.org/10.1016/j.clsr.2021.105567

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal*, 17, 131–160.

Wasserman-Rozen, H., Gilad-Bachrach, R., & Elkin-Koren, N. (2024). Lost in translation: The limits of explainability in AI. *Cardozo Arts & Entertainment Law Journal*.

Yarger, L., Cobb Payton, F., & Neupane, B. (2020). Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review*, 44(2), 383–395.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. https://doi.org/10.1145/3038912.3052660