PIONEER
ACADEMIC PUBLISHING LIMITED

Studies in Law and Justice
ISSN 2958-0382
www.pioneerpublisher.com/slj
Volume 4 Number 4 August 2025

Fair by Design? The Legal and Ethical Challenges of Algorithmic Hiring

Alexandre Pacheco da Silva[1] & Enya Carolina Silva da Costa[2]

[1] Science and Technology Policy, Department of Geosciences, University of Campinas, São Paulo, Brazil
[2] University of São Paulo, São Paulo, Brazil
Correspondence: Enya Carolina Silva da Costa, University of São Paulo, São Paulo, Brazil.

doi:10.56397/SLJ.2025.08.01

Abstract

This paper critically examines the promise and pitfalls of algorithmic hiring systems through a legal and ethical lens. Focusing on cases such as Pymetrics, HireVue, and Amazon's résumé screening tool, we explore how automated decision-making in recruitment, despite claims of neutrality and fairness, often reproduces or amplifies existing social inequalities. Drawing from recent legal scholarship and normative theories of algorithmic fairness, we show how systems designed to minimize human bias can inadvertently encode discriminatory assumptions into technical infrastructures. The paper analyzes competing fairness frameworks and emphasizes that fairness is not a purely technical feature, but a normative commitment that must guide every stage of system development and deployment. We argue for a shift from reactive audits to proactive, participatory governance models grounded in transparency, inclusiveness, and accountability. Through the "Developer's Model for Responsible AI", we propose a structured, lifecycle-based approach to operationalize fairness in algorithmic systems, especially in sensitive domains like employment. Ultimately, the paper contends that ensuring justice in AI is not only a technical challenge but a democratic imperative.

Keywords: algorithmic fairness, automated hiring, discrimination, responsible AI, governance

1. Introduction: Algorithmic Glitches and the Illusion of Neutrality

We live in an era where decisions that deeply impact people's lives are increasingly being delegated to artificial intelligence (AI) systems (Gasser & Mayer-Schönberger, 2024, p. 13). From résumé screening in hiring processes (Wilson et al., 2021) to credit approval (Gillis, 2022), and from judicial decisions (Legg & Bell, 2020) to educational assessments (O'Neil, 2016) and welfare distribution (Eubanks, 2018), algorithms have become central mediators of opportunities, rights, and obligations. Yet this technical advancement comes with a troubling paradox: while AI promises greater efficiency and impartiality, it often replicates, and even amplifies, historical inequalities it was meant to mitigate, not automate (O'Neil, 2016).

The question, therefore, is not simply how to build technically robust systems, but how to ensure that these systems operate on legitimate principles of justice and equality.

This challenge demands a fundamental shift in our mindset. For decades, technical flaws in computational systems were treated as "natural failures" (Broussard, 2019), inevitable statistical noise (O'Neil, 2016), the price of progress. In stark contrast to this normalization of failure, Meredith Broussard offers a sharp critique in her book *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech* (2023). What we often label as AI failures—bias, exclusion, injustice—are not merely technical accidents or random distortions. They are the predictable outcomes of design decisions made by largely homogeneous teams using biased datasets without meaningful engagement with the communities affected. Glitches, Broussard (2023) argues, are not bugs in the system; they are symptoms of a technological imagination that refuses to confront its social consequences.

This critique becomes even more concrete when we turn to algorithmic hiring systems such as Pymetrics. The company promised to offer a fairer, neuroscience-based hiring process through gamified assessments. But, as we will see, its design replicated bias instead of erasing it, revealing how discrimination is not a side effect of algorithmic systems, but often a consequence of flawed assumptions built into their architecture.

If Broussard (2023) exposes the risks, Orly Lobel proposes a counter-narrative. In *The Equality Machine* (2022), she flips the script: what if, instead of fearing algorithms as tools of injustice, we intentionally designed them as mechanisms for promoting fairness? For Lobel, technology is not neutral; it can and should be shaped by normative commitments from the start.

That requires abandoning a reactive view of justice, one based on fixing things after the fact, and adopting a proactive one: designing systems that actively promote equity, inclusion, and diversity (Lobel, 2022). This means forging strong collaborations between programmers, lawyers, philosophers, and affected communities in what Lobel calls "built-in fairness."

These provocations lead us to the central question of this paper: what does it take to build algorithmic decision systems that embed legitimate standards of equality from the ground up? How do we move from the myth of algorithmic neutrality to the messy, but necessary, practice of computational equity?

To explore this challenge, the paper begins by examining the cases of Pymetrics, HireVue, and Amazon, examples that encapsulate both the promise and the peril of algorithmic hiring. The cases serve as a lens through which to analyze how such systems are designed, the forms of fairness they claim to embody, the biases they inadvertently reproduce, and the strategies proposed to mitigate these biases. Building on that, the discussion then turns to the foundational models of algorithmic fairness, particularly as articulated by Barocas et al. (2023).

The second section unpacks competing visions of fairness while scrutinizing their respective technical, legal, and ethical implications. The final section advances a dialogue between two complementary approaches to ensuring fairness: one focused on embedding justice from the outset through "fairness by design", and the other centered on *ex post* mechanisms of evaluation through "algorithmic auditing". Together, these perspectives illuminate how fairness can be integrated across the entire life cycle of algorithmic systems and inform a more robust framework for sustained accountability.

Our goal is normative and practical: to imagine systems that are not only designed to "do no harm" but that are explicitly built to do good. Because ultimately, what's at stake in the quest for fair AI is not just the future of technology, it's the future of equality.

2. From Promise to Paradox: Lessons from Pymetrics, HireVue, and Amazon

The story of Pymetrics offers an illuminating case study of the tensions at the heart of algorithmic hiring. Founded with the mission of using neuroscience-based games and machine learning to match job candidates with roles based on cognitive and emotional traits, Pymetrics was positioned as a progressive alternative to traditional hiring practices, ones often steeped in human bias. The company claimed that its tools would promote fairness by eliminating subjectivity and focusing on scientifically grounded traits (Wilson et al., 2021). What emerged instead was a more complicated picture: one where fairness itself became a contested terrain, and where the very mechanisms intended to correct bias ended up perpetuating it in new forms.

Unlike conventional hiring platforms that

connect résumés to job openings, Pymetrics operates as a behavioral analytics service focused specifically on candidate screening. Its approach involves using gamified tasks, drawn from neuroscience and cognitive psychology, to assess a wide range of behavioral and cognitive traits, such as attention span, risk tolerance, and emotional regulation. These assessments are not scored in a vacuum. Instead, Pymetrics develops predictive models trained on gameplay and performance data from high-performing employees in a given role, aiming to identify candidates whose behavioral patterns align with those of top performers (Wilson et al., 2021) and to create predictive models for success (Raghavan et al., 2020).

The screening process unfolds through a structured partnership with employers. First, Pymetrics gathers detailed information about the target position, including performance criteria and team composition. Then, current employees in that role are asked to complete the same suite of games offered to applicants. Their gameplay data, combined with their on-the-job performance metrics, is used to train a machine learning model. This model is tested not only for its predictive accuracy but also for compliance with legal fairness guidelines, specifically the Uniform Guidelines on Employee Selection Procedures (UGESP's) four-fifths rule[1], to avoid disproportionate exclusion of candidates from protected groups (Wilson et al., 2021).

Once a model passes these benchmarks, it is deployed in real-world hiring. New candidates complete the same set of games, and the model scores them based on how closely their behavioral profile matches that of previously identified top performers. Those with high compatibility scores are advanced to the next stage of the hiring process, such as résumé review or interviews. Pymetrics positions this system as both efficient and fair, arguing that it minimizes human bias through standardized, scientifically grounded evaluations (Wilson et al., 2021).

However, the promise of fairness through automation is far from uncontested. While the company claims to audit and update its models regularly to mitigate disparate impact, critics have raised concerns about the opacity of its algorithms and the limitations of behavioral profiling. When cognitive and emotional norms are codified into models of employability, they risk reinforcing narrow conceptions of talent that exclude neurodivergent, disabled, or culturally diverse candidates (Andrews & Bucher, 2022).

One illustrative example of Pymetrics' game-based assessment is a task inspired by the Balloon Analogue Risk Task (BART), a measure in behavioral science used to evaluate risk tolerance. In this game, a candidate is asked to inflate a virtual balloon, earning points for each pump. However, the balloon can pop at any moment, causing the candidate to lose the accumulated points. The key behavioral signal here is how far a candidate is willing to push their luck before cashing out, an indirect indicator of how they manage risk under uncertainty. A candidate who consistently inflates the balloon only a few times before securing their points might be deemed more risk-averse, while one who pushes the limit may be categorized as more risk-tolerant.

These behaviors are not interpreted in isolation. The AI model behind Pymetrics analyzes not only the final outcomes (e.g., how many points a candidate earned) but also the micro-patterns in their decision-making: how long they hesitate between pumps, whether their strategy changes over time, or if they adjust after a balloon pops. Each of these data points is logged, quantified, and fed into a machine learning algorithm that compares the candidate's responses to those of high-performing incumbents. The model is trained to detect the behavioral patterns most predictive of success in a given role, say, financial analysts who perform well under pressure may share a specific risk-taking profile. This allows the system to flag candidates who exhibit similar traits, regardless of their educational background or professional experience (Wilson et al., 2021).

But the granularity of this analysis also raises important questions. What happens when a candidate's play style deviates from the

[1] The four-fifths rule, also known as the 80% rule, is a guideline developed by the U.S. Equal Employment Opportunity Commission (EEOC) to detect potential adverse impact in employment practices. According to this rule, a selection rate for any demographic group, such as based on race, gender, or ethnicity, should be at least 80% of the rate for the group with the highest selection rate. If the selection rate for a particular group falls below this threshold, it may be considered evidence of discriminatory impact, even in the absence of intentional bias. For example, if 60% of white applicants are selected for interviews but only 30% of Black applicants are, the selection rate for the latter group is only 50% of the former, well below the 80% benchmark, indicating possible disparate impact.

expected profile, not because they lack the skills, but because they approach decision-making differently due to cultural norms, disability, or neurodivergence? The algorithm's interpretation of "ideal" behavior is only as inclusive as the training data it learns from. In this sense, a candidate's cautious strategy in the balloon game might be misread as a liability in a role that values assertiveness, even if their restraint would actually be an asset. Thus, what appears as a neutral behavioral metric is, in practice, embedded with normative assumptions, making the promise of fairness through AI as contested as the hiring practices it seeks to replace.

This dilemma echoes what legal scholar Ifeoma Ajunwa (2020) has termed the "paradox of automation as anti-bias intervention". While technologies like Pymetrics are marketed as solutions to human bias, precisely because they rely on standardized data and algorithmic consistency, Ajunwa warns that such systems often reproduce the very forms of discrimination they claim to eliminate. In Pymetrics' case, the reliance on behavioral data from high-performing employees risks encoding existing workplace norms into the algorithm. If those norms are themselves shaped by historical inequalities, such as the underrepresentation of disabled, neurodivergent, or racially marginalized employees, then the model merely automates exclusion in a more opaque, technical form.

What makes this particularly troubling is the illusion of neutrality. Because candidates are assessed through playful, seemingly objective games, the process can feel fair and scientifically grounded. Yet the criteria used to evaluate performance—how fast someone reacts, how they manage risk, how flexible they are in adapting to rules—are themselves subjective, value-laden, and often culturally contingent. Ajunwa's insight highlights how algorithmic tools can obscure structural bias behind a veneer of precision, making discriminatory outcomes harder to detect and legally contest. Unlike a biased human recruiter, an algorithm can't be cross-examined, and its design choices often evade transparency.

Moreover, the legal frameworks designed to prevent discrimination in hiring are poorly equipped to handle this new form of bias. As Ajunwa (2020) notes, laws like Title VII[1] were built around human actors and interpersonal prejudice, not machine learning systems trained on thousands of data points. In the case of Pymetrics, even if the company complies with formal fairness metrics like the four-fifths rule, these statistical thresholds may overlook more subtle, cumulative forms of disadvantage, such as the systematic misinterpretation of behavior by neurodiverse candidates. In short, compliance is not the same as justice.

The Pymetrics case thus illustrates Ajunwa's broader argument (2020): that automating human judgment does not dissolve bias, but rather recasts it in algorithmic form. By embedding contested notions of competence and fitting into game-based assessments, the system risks naturalizing exclusion under the guise of innovation. Far from eliminating bias, algorithmic hiring tools may simply shift where and how discrimination occurs, placing it beyond the reach of those most affected. This is the paradox at the heart of algorithmic fairness: a system built to correct human flaws may only deepen them, unless its assumptions are made visible and its values open to debate.

This is precisely what the case of HireVue brings into sharper relief. Like Pymetrics, HireVue marketed itself as a tool for making hiring "fairer" by replacing gut instinct with AI-driven assessment. But unlike Pymetrics, which focused on cognitive games, HireVue's platform included facial and vocal recognition technologies to analyze candidates during video interviews. These systems evaluated not just content, but tone, micro-expressions, eye movements, and other behavioral cues. Civil liberties groups, including Electronic Privacy Information Center (EPIC), raised concerns that the tool penalized candidates based on features like accents, anxiety levels, or lighting conditions, factors that disproportionately affect

[1] Title VII of the 1964 Civil Rights Act is a foundational piece of U.S. civil rights legislation that prohibits employers from discriminating against individuals on the basis of race, color, religion, sex, or national origin. It applies to all aspects of employment, including hiring, promotion, compensation, and termination. Title VII not only forbids overt, intentional discrimination (disparate treatment), but also extends to practices that may appear neutral on their face but result in disproportionate harm to protected groups (disparate impact). This dual focus makes it particularly relevant in the context of AI-driven hiring tools, where algorithms may unintentionally replicate historical patterns of exclusion. While Title VII was designed with human decision-makers in mind, its principles now serve as a critical reference point in the legal and ethical evaluation of algorithmic systems in the workplace (Páez, 2021, p. 24).

individuals from marginalized racial, socioeconomic, or neurodivergent backgrounds. As Sheard (2022, p. 620) notes, when models are trained on historical decision-making data or normative behavioral patterns, they are highly susceptible to replicating entrenched inequalities.

This concern became tangible in 2025 when a deaf Indigenous woman, identified as D.K., filed a discrimination complaint against Intuit and HireVue. The complaint alleged that the HireVue system misinterpreted her video interview due to her speech pattern, which differed from normative training data because of her disability. As a result, the AI-generated score reflected not her actual competencies but the system's inability to interpret her correctly. Despite being qualified and already working at Intuit, she was denied a promotion (Sheard, 2022, p. 623). This case underscores what Sheard (2022, p. 628) calls the "liability vacuum": harm occurs, but the diffusion of responsibility across vendors, data scientists, and employers prevents clear accountability.

The parallels with Pymetrics are significant. Both companies developed tools grounded in behavioral science and marketed them as bias-reducing innovations. And both rely on data from existing employees, people who often reflect the demographics and behavioral norms of previously privileged groups. While Pymetrics uses game data and HireVue analyzes audiovisual input, the underlying logic is the same: performance is modeled on those who have already succeeded, assuming those profiles are universally applicable. Yet, as Sheard (2022, p. 632) emphasizes, when systems are trained on unrepresentative data and evaluated through inaccessible models, they become vehicles for indirect discrimination, subtle, legalistic, and hard to contest.

Moreover, both companies have adopted internal auditing procedures aimed at demonstrating compliance with formal fairness standards. Pymetrics, for instance, open-sourced part of its auditing code and tested for adherence to the four-fifths rule (Wilson et al., 2021). HireVue, under public pressure, eventually phased out its facial recognition component. But as Sheard (2022, p. 630) argues, these gestures are often insufficient: internal audits do not guarantee external accountability, and formal compliance does not ensure substantive fairness. The real issue lies not in whether these systems can meet statistical parity thresholds, but in whether the assumptions they encode, about competence, behavior, and fit, are themselves just.

Another revealing example of algorithmic hiring gone awry is the case of Amazon Jobs (Sheard, 2022, p. 624), which also set out to eliminate human bias through machine learning. In 2018, Amazon developed an internal AI tool to automatically rank résumés for software engineering roles. The tool was trained on a decade's worth of past hiring decisions, data that, as it turned out, reflected the company's historical preference for male candidates. Unsurprisingly, the AI began downgrading résumés that included words like "women's chess club" or came from all-women's colleges. Although the system did not explicitly consider gender as a variable, the bias was encoded in the patterns it learned.

This example mirrors Sheard's broader critique: bias is often not about what data is explicitly fed into the system, but about what the model *learns* from historical structures. Like Pymetrics, Amazon's system relied on existing data to determine what a "good" candidate looks like. And like Pymetrics, it assumed that past performance is an adequate and neutral benchmark for future success. Yet when historical data reflect exclusionary practices or demographic imbalances, models trained on them will necessarily reproduce these patterns, regardless of whether the developers intend to discriminate.

What makes the Amazon case particularly instructive is that the company eventually abandoned the tool, recognizing that the bias it embedded was too deeply rooted to be easily corrected (Sheard, 2022, p. 624). This stands in contrast to Pymetrics, which continues to operate with the claim that its games and models can deliver fairer outcomes. However, both cases suggest a troubling overconfidence in the ability of behavioral proxies—whether linguistic, cognitive, or gamified—to serve as neutral indicators of talent. As Sheard (2022) argues, these systems don't simply fail to overcome bias; they recode it into data-driven language that is harder to interrogate and easier to legitimize.

The convergence of these three cases, Pymetrics, HireVue, and Amazon, reveals a broader pattern in algorithmic hiring: a persistent failure to

recognize that fairness is not a technical output but a normative commitment. Tools that claim to reduce bias by optimizing behavioral signals often obscure the cultural, neurological, and social assumptions embedded in those very signals. Whether it's the misinterpretation of a deaf accent, the undervaluing of cautious decision-making, or the erasure of non-masculine leadership styles, these systems embed a behaviorist epistemology that equates "fit" with conformity to dominant norms.

Sheard's analysis offers a critical lens through which to view these failures, not as isolated glitches, but as structural outcomes of a model of hiring that prioritizes efficiency and scalability over contextual judgment and equity. When responsibility is diffused across technical, legal, and organizational domains, and when performance is measured by proprietary proxies, those harmed by algorithmic decisions are left without recourse. The real danger, as Sheard (2022, p. 630) warns, lies in how this model of automation presents its results as neutral facts, when they are, in truth, deeply political choices made invisible through code.

In the end, these cases serve as cautionary tales. They show that algorithmic fairness cannot be achieved by technical patchwork or an internal audit alone. What is needed is a deeper reckoning with the assumptions driving these systems—who defines merit, what counts as evidence of potential, and who is authorized to decide. Without that reckoning, the promise of AI in hiring will remain an elegant fiction: a system built to reduce bias that, in practice, only reorganizes it behind the veil of objectivity.

3. Invisible Values, Visible Harms: Rethinking Accountability in Algorithmic Hiring

These concerns are further reinforced by the comprehensive review conducted by Anna Lena Hunkenschroer and Christoph Lütge (2022), who identify a broad spectrum of ethical tensions arising from AI-enabled recruiting and selection systems. In their analysis of over fifty academic and industry sources, the authors argue that while algorithmic tools promise improvements in efficiency, consistency, and bias mitigation, they also introduce new and understudied ethical risks. These include not only well-known concerns like data privacy and bias reproduction, but also deeper moral ambiguities, such as the appropriate trade-off between predictive performance and fairness, and the moral legitimacy of behavioral proxies used in candidate evaluation. Their contribution helps shift the discussion from purely technical concerns toward a broader inquiry into the normative values that underpin automated hiring.

One of the most significant ethical risks they identify lies in the illusion of objectivity. When algorithms are perceived as neutral tools, there is a tendency to overlook how design choices, ranging from data labeling to performance criteria, are already value-laden. As Hunkenschroer and Lütge (2022) caution, systems that evaluate candidates based on personality traits, emotional signals, or behavioral tendencies often rely on culturally specific assumptions about what constitutes a "good fit." These assumptions may inadvertently marginalize neurodivergent individuals, people with disabilities, or candidates from underrepresented backgrounds.

Ajunwa (2023, p. 88) draws attention to this particularly insidious form of algorithmic discrimination: the use of "cultural fit" as a proxy for race or class-based exclusion. In many hiring platforms, algorithms are trained on the profiles of previous "successful" employees, embedding historical biases into the predictive model. What appears as a neutral preference for candidates who "fit the culture" often masks the perpetuation of racially and socioeconomically homogeneous workplaces.

Ajunwa warns that this type of discrimination is especially dangerous because it presents itself as meritocratic and efficiency-driven, when in reality it reproduces structural inequality through coded language and design choices. By optimizing for traits associated with past hires—such as communication style, problem-solving approach, or demeanor—automated systems risk filtering out equally competent candidates who do not mirror existing norms. These design choices become gatekeeping mechanisms that entrench exclusion while appearing objective.

Consider, for instance, a candidate taking part in Pymetrics' suite of gamified assessments, including the previously mentioned game inspired by the Balloon Analogue Risk Task. By removing human subjectivity from the hiring process, the system claims neutrality, but what it actually removes is context. A human recruiter

might ask *why* a candidate approached the task cautiously, perhaps they were being strategic, or perhaps they come from a cultural context where calculated restraint is valued. The algorithm, in contrast, interprets that caution as a measurable trait without nuance, assigning it a numerical score and fitting it into a pre-constructed model of ideal behavior. This mechanical interpretation may appear impartial, but it conceals normative biases built into the architecture of the system itself. In attempting to eliminate the variability of human judgment, the system replaces it with a rigid and unexamined hierarchy of behavioral preferences, turning judgment into computation and fairness into a statistical façade.

The subjectivity of human evaluators is far from unproblematic. Human recruiters may exhibit implicit biases, favoritism, cultural misunderstandings, or inconsistent judgment, all of which can lead to unfair hiring outcomes. These flaws are well-documented and, in many cases, visible and challengeable through interviews, appeals, or legal processes. Our aim is not to romanticize human discretion or ignore its dangers. Rather, the concern lies in the *illusion* that algorithmic systems have solved these issues simply by removing the human element. When subjectivity is embedded in code, through design choices, training data, or performance criteria, it becomes harder to detect and contest. The risk is not just that AI systems make biased decisions, but that they do so under the mask of neutrality and scientific legitimacy, making the underlying value judgments less visible and more difficult to question.

It is in response to this challenge that Van Giffen et al. (2022) propose a comprehensive and accessible framework to understand and address machine learning bias. Recognizing that algorithmic systems do not eliminate subjectivity but instead shift and obscure it, the authors seek to bridge fragmented literature by providing a shared vocabulary and actionable guidance. Using the Cross-Industry Standard Process for Data Mining (CRISP-DM) as an organizing structure, they map eight distinct types of bias and mitigation strategies across the phases of a project. This effort not only clarifies where and how bias can emerge, but also equips researchers and practitioners with tools to identify and intervene in these points of vulnerability.

Among the types of bias identified, *social bias* refers to the reproduction of preexisting inequalities embedded in available data, biases that precede model design and reflect societal patterns. *Measurement bias* emerges when chosen features or labels act as poor proxies for the variables of actual interest, such as using hospital admissions as a stand-in for health status. *Representation bias*, meanwhile, results from input data that fail to adequately reflect the diversity of the relevant population, leading to systematic errors, particularly when underrepresented groups are marginalized in training datasets.

Additional sources of bias arise further along the pipeline. *Label bias* occurs when labeled data systematically deviates from the underlying truth, often due to subjective, inconsistent, or biased categorization. *Algorithmic bias* stems from technical design decisions—such as model architectures, loss functions, or training procedures—that produce unequal outcomes. Even evaluation and deployment introduce distinct risks: *evaluation bias* can occur when testing data or metrics are misaligned with real-world performance, while *deployment bias* arises when systems are used in contexts different from those for which they were built. Finally, *feedback bias* illustrates how algorithmic outputs can shape user behavior and future datasets, creating self-reinforcing loops of discrimination. Mapping these biases across the stages of a machine learning project, as Van Giffen et al. propose, makes it possible to identify targeted mitigation strategies at each step, an essential move toward more accountable and socially aware algorithmic systems.

However, identifying and mitigating bias is only part of the challenge. Equally important is the question of accountability—who is responsible when harm occurs, and what mechanisms exist to ensure that mitigation efforts translate into meaningful protections for affected individuals and groups. In this regard, Hunkenschroer and Lütge (2022) highlight a persistent gap in accountability mechanisms. While some

companies adopt internal fairness audits[1] or comply with procedural fairness standards like the four-fifths rule[2], these measures are often insufficient to address the substantive harms caused by algorithmic decision-making. This echoes the concern raised earlier in the context of HireVue and Amazon: that fairness cannot be reduced to statistical parity alone. As Hunkenschroer and Lütge note, internal audits, without external oversight, risk becoming reputational tools rather than instruments of genuine accountability. When the impact of an AI system disproportionately harms a particular group, the question is not only whether it meets compliance thresholds, but whether it respects the broader ethical principle of equal opportunity.

Michael Kearns and Aaron Roth (2019, p. 70) provide a critical examination of statistical parity as a fairness criterion in algorithmic decision-making. While statistical parity, ensuring that different demographic groups are selected at similar rates, may appear to promote equality, the authors argue that it can produce misleading and even counterproductive results. One of the core problems is that statistical parity focuses solely on outcomes without regard to underlying qualifications or contexts. This can lead to situations where individuals with vastly different attributes are treated identically, potentially sacrificing merit-based considerations in the name of equal group-level representation.

Furthermore, Kearns and Roth (2019, p.71) emphasize that enforcing statistical parity may inadvertently introduce new forms of unfairness. For example, to equalize acceptance rates across demographic groups, an algorithm might be forced to lower thresholds for one group or raise them for another, leading to perceptions of reverse discrimination or unjustified favoritism. In practice, this can create tensions between fairness and accuracy, especially in high-stakes contexts like hiring or college admissions, where decision-makers must weigh individual qualifications against broader social goals. For Kearns and Roth, the key is not to reject group-based fairness metrics altogether, but to recognize their limitations and use them alongside other criteria that better capture the nuances of individual justice.

They also highlight the challenge of strategic gaming when statistical parity becomes a rigid requirement (Kearns & Roth, 2019, p. 72). Once organizations are required to meet demographic quotas, there may be incentives to manipulate input data or alter labeling practices to produce superficially fair outcomes without making meaningful structural changes. In hiring, for instance, a firm might design its AI system to pass fairness audits by adjusting score thresholds across groups, while still relying on biased features that disadvantage marginalized candidates in subtler ways. As Kearns and Roth (2019, p. 71) caution, fairness is not simply a constraint to be satisfied, but a dynamic and context-sensitive principle, one that must be thoughtfully integrated into the architecture of algorithmic systems from the outset.

Importantly, Hunkenschroer and Lütge (2022) also underscore the psychological and societal implications of AI-based hiring. Candidates subjected to opaque systems may experience dehumanization, alienation, and a loss of agency. Unlike human interviewers, algorithmic systems rarely provide meaningful feedback, leaving candidates uncertain about why they were rejected or what they might improve. This creates an ethical tension between efficiency and transparency. A faster, cheaper process may come at the cost of undermining dignity, trust, and procedural fairness, values that are central to democratic labor markets. From this perspective, the deployment of AI in hiring is

[1] Hunkenschroer and Lütge (2022) define fairness audits as systematic evaluations of algorithmic systems intended to detect and mitigate discriminatory outcomes, particularly those affecting protected groups. These audits typically involve assessing whether a system's outputs result in disparate impact and whether they comply with formal fairness criteria such as demographic parity or the four-fifths rule. However, the authors emphasize that fairness audits often remain narrow in scope, focusing on statistical indicators rather than addressing the deeper normative assumptions embedded in system design. They caution that such audits can become performative tools, used more to signal compliance than to enact meaningful change, especially in the absence of external oversight or public transparency. For instance, a company might audit its AI hiring tool and find that selection rates for men and women are statistically similar, thereby passing the audit, even though the model still penalizes candidates who display communication styles more common among women.

[2] While the four-fifths rule provides a quantitative standard for identifying imbalances in hiring or promotion decisions, it is not a definitive legal test. It is a diagnostic tool meant to trigger further investigation rather than prove discrimination outright. In the context of AI-based hiring systems, many companies use the rule as part of internal audits to demonstrate compliance with fairness benchmarks. However, critics argue that this kind of formal parity can obscure deeper forms of exclusion, particularly when the criteria used for evaluation, such as behavioral traits or cognitive scores, are themselves biased. In other words, a system can meet the four-fifths rule and still perpetuate inequality if the underlying assumptions remain unchecked (Páez, 2021, p. 24).

not just a technical shift; it is a moral transformation of how we understand evaluation, inclusion, and worth.

Ultimately, the authors call for a research and policy agenda that goes beyond algorithmic optimization to embrace ethical reflexivity. They advocate for the integration of moral philosophy into AI development processes, interdisciplinary collaboration across law, ethics, and computer science, and the cultivation of organizational cultures that value inclusion over efficiency.

The growing trust in artificial intelligence as a substitute for human rationality is often accompanied by a seductive promise: that algorithmic systems can deliver faster, more efficient, and, most importantly, more impartial decisions than humans ever could. Yet this promise masks a deep moral trap. As Brian Christian (2020) reminds us, the real challenge posed by AI is not its ability to process massive amounts of data or detect complex patterns, but its failure to align those patterns with legitimate human values. What happens when a system makes a technically correct decision that, from an ethical perspective, feels profoundly unjust?

This tension is powerfully illustrated in the case of Pymetrics. Imagine two candidates who perform equally well on the platform's behavioral games, but are ranked differently because one of them exhibits a response pattern more common among white, middle-class men. Technically, the decision is defensible; the algorithm is simply reflecting patterns learned from historical data, but morally, it is indefensible. Why should the statistical average of a dominant social group serve as the benchmark for evaluating individuals who do not share that background but possess the same skills? This is not a technical error, but a normative one: a confusion of correlation with justice. A statistically robust model may still violate fundamental principles of equality (equal treatment under the same conditions) and equity (ensuring comparable opportunities for those with different starting points). This is where the so-called alignment problem emerges: the disconnect between what AI *can* do and what society *expects* it to do.

Christian argues that this disconnect demands a shift in our design perspective. The challenge is not merely to build systems that work well, but to build systems that do good.

In *Fairness and Machine Learning: Limitations and Opportunities* (2023), Barocas et al. propose a reorientation of both the conceptual and technical agenda for building machine learning systems that internalize fairness. Fairness, they argue, is not a matter of statistical performance alone; it is a normative commitment to non-discrimination and the proactive promotion of equal opportunity. This commitment requires more than simply removing sensitive variables from datasets; it calls for a full rethinking of data collection practices, optimization goals, and the social groups impacted by those systems.

The case of Amazon's résumé screening tool illustrates precisely why fairness cannot be reduced to statistical performance or the mere removal of sensitive attributes like gender. Although the system was designed to be "gender-blind" by excluding explicit gender indicators, it still learned to associate proxies of maleness—such as participation in all-male sports teams or attendance at male-dominated institutions—with higher hiring potential. This outcome reveals what Barocas et al. (2023) emphasize: that removing protected variables does not neutralize a system if the underlying data and optimization goals continue to reflect historical patterns of discrimination.

What Amazon's case demonstrates is that fairness cannot be achieved through technical tweaks alone. The absence of gender as a feature did not prevent gender bias; it merely obscured its mechanism. As Barocas et al. (2023) argue, fairness demands attention not just to the outputs of a system, but to the entire process by which data is collected, interpreted, and used to train predictive models.

Without confronting the social conditions embedded in training data and the institutional goals driving optimization, fairness efforts remain superficial. Any attempt to make AI "fair" involves explicit choices about which inequalities to address, how individuals should be treated, and how to balance predictive accuracy with social responsibility. In this light, fairness is not merely a statistical metric but a normative lens; it is a field of moral and political contestation, a perspective on justice that must guide the system's entire lifecycle.

This debate becomes even more urgent when

examined through the lens of negligent [1] algorithmic discrimination, a concept advanced by Páez (2021, p. 27). In this view, algorithmic bias is not simply an unfortunate side effect of optimization. It is the foreseeable result of poor design decisions, weak safeguards, and structural indifference to inequality. To describe an algorithm as "biased" is to suggest a flaw; to call it "negligent" is to attribute responsibility.

The case of HireVue offers a compelling example of what Andrés Páez (2021) terms negligent algorithmic discrimination, the failure to take reasonable precautions against foreseeable harms caused by automated systems. The system operated as a black box, with little transparency into how the signals were interpreted or weighted. From Páez's perspective, the problem is not simply that HireVue's technology produced biased outcomes, but that those outcomes were entirely preventable.

Developers and adopters had ample warning from ethicists, technologists, and advocacy organizations about the discriminatory potential of facial and voice analysis. Yet rather than halt deployment or subject the system to rigorous public validation, HireVue proceeded, only discontinuing its facial analysis component in 2021, after public backlash. This sequence reflects a classic case of algorithmic negligence: not acting out of malice, but failing to anticipate or meaningfully address harms that were both foreseeable and ethically consequential. It illustrates how the abdication of responsibility in the name of innovation can result in deeply unjust consequences, especially when fairness is measured by compliance metrics rather than lived impacts.

What these cases reveal is not only the risk of misclassification but the erosion of opportunity itself. Algorithmic systems increasingly shape who is seen, who is shortlisted, and who is hired. They convert social disparities into technical signals. They do not just replicate bias; they normalize it, embedding discrimination into the architecture of decision-making.

This is why the call to frame bias as negligence is not simply a legal strategy; it is a moral imperative. Negligence occurs when foreseeable damage is not tested for, when unrepresentative data is used without scrutiny, and when systems lack interpretability. Even more troubling is the use of fairness as a public relations strategy rather than a commitment to accountability. Developers and deploying institutions must ensure that algorithmic tools do not reinforce structural harm.

Ajunwa (2023, p. 80) challenges the notion that automated decision-making is categorically distinct from human decision-making, arguing that this separation constitutes a "false binary." She contends that automation does not eliminate human judgment; it merely reconfigures it into different layers of design, deployment, and interpretation. Behind every algorithm are human choices: about which data to collect, which features to prioritize, and what trade-offs to tolerate. By framing automation as neutral or objective, institutions obscure the very real human agency embedded in these systems and evade the ethical scrutiny that would typically accompany discriminatory decisions made by people.

Recognizing this false binary is crucial for understanding why negligence in algorithmic systems should be taken just as seriously, if not more so, than in human-led processes. The veneer of technological objectivity can anesthetize both users and the public to harm, enabling biased outcomes to persist under the guise of efficiency. By accepting Ajunwa's critique, we see that the responsibility for discriminatory outcomes cannot be shifted onto the machine; it rests squarely with those who design, implement, and rely on these tools without rigorous safeguards. This reframing helps bridge the moral disconnect highlighted in Bigman et al.'s findings, revealing that the real failure lies not in the algorithm's intent but in the abdication of human responsibility.

And yet, as recent research by Bigman et al. (2022) shows, algorithmic discrimination provokes less moral outrage than human

[1] Andrés Páez (2021, p. 28) defines negligence in the algorithmic context as the failure to take reasonable precautions against foreseeable harms caused by automated decision-making systems, particularly harms related to discriminatory outcomes. Unlike intentional discrimination, which involves purposeful bias, negligent discrimination arises when developers or deploying institutions ignore warning signs, overlook structural risks, or inadequately test their systems for unfair impact. For Páez, negligence is not just about flawed outputs, but about a lack of due diligence in anticipating and mitigating how algorithms may disadvantage protected groups. This includes failing to audit training data for representativeness, disregarding how models interact with social contexts, or assuming that technical neutrality absolves moral responsibility. In this sense, negligent algorithmic discrimination is ethically serious not because it is malicious, but because it is avoidable.

discrimination. This "algorithmic outrage deficit" stems from the perception that algorithms, being mindless tools, cannot possess intention and therefore cannot be held morally responsible. But absence of intent is not absence of harm. When organizations hide behind algorithmic opacity to escape accountability, the moral damage is doubled.

Barocas et al. (2023) remind us that fairness, to be meaningful, must be designed from the beginning, not bolted on after deployment. Their typology of fairness criteria—demographic parity, equal opportunity, individual fairness—offers frameworks, but no silver bullet. In practice, these criteria often conflict. For instance, achieving demographic parity across race groups in the Pymetrics case would require altering model thresholds in ways that could reduce individual-level accuracy. This illustrates the classic trade-off between predictive precision and distributive justice.

Other approaches attempt to navigate this tension. Statistical parity requires ongoing monitoring of approval rates across demographic groups, with algorithmic recalibration as necessary. The Pareto frontier model seeks the optimal balance between equity and performance, allowing companies to make ethically informed trade-offs rather than performance-maximizing shortcuts.

Yet perhaps the most important lesson is this: fairness is not a property to be added to a functioning system. It is a normative lens that must guide every step of system development, from problem framing and data collection to model design, deployment, and evaluation. AI does not eliminate the ethical tensions of decision-making; it codifies them. The real challenge is not simply technical. It is political. Either we design systems that reflect democratic values, or we silently accept the consolidation of new forms of exclusion masquerading as algorithmic objectivity.

4. From Code to Consequence: Algorithmic Fairness as a Political Imperative

The hope that technology might correct centuries of structural discrimination is nowhere more vividly tested than in the realm of hiring. The story of Lakisha Washington—a highly qualified Black woman whose résumé, identical to that of a white applicant except for the name, consistently received fewer callbacks—illustrates the deeply entrenched nature of racial bias in employment practices. Her case, part of a now-famous field experiment conducted by economists Marianne Bertrand and Sendhil Mullainathan[1], underscores how even minimal signals, like a name, can trigger exclusion in supposedly meritocratic processes. This raises an urgent question: can artificial intelligence, when used in hiring, help break this cycle, or does it merely encode and automate it?

Orly Lobel (2022) takes up this question with nuance and ambition. She challenges the techno-pessimistic view that algorithms inevitably replicate human bias. Instead, she asks: what if, properly designed, machine learning systems could not only avoid past discrimination but actively promote *equality of outcomes*? For Lobel, the transformative potential of AI lies not in its mimicry of existing decision patterns but in its ability to offer a more corrective, intentional model of fairness, one that moves beyond formal equality to material inclusion. But realizing that potential, she warns, is far from straightforward.

Among the barriers Lobel (2022) identifies is the human element behind every system: the so-called "coding ninjas" whose decisions, blind spots, and assumptions shape what an algorithm sees and what it ignores. Engineers may believe they are building neutral systems, but their values inevitably enter the code. From the selection of training data to the choice of optimization goals, each step involves subjective judgment. When these developers come from homogeneous backgrounds or fail to consult with affected communities, the systems they create risk reinforcing the very inequalities they claim to solve.

Orly Lobel (2022) suggests that systems like Pymetrics could be redesigned to actively promote equitable outcomes rather than simply avoiding overt discrimination. This would

[1] In this *New York Times* article, economist Sendhil Mullainathan revisits his field experiment with Marianne Bertrand, which demonstrated how identical résumés received different callback rates based solely on the perceived race suggested by applicants' names. One résumé bore the name Lakisha Washington; the other, Emily Walsh. Despite being equally qualified, Lakisha received far fewer responses, a finding that exposed how deeply racial bias shapes hiring decisions. Mullainathan argues that algorithmic systems, if properly designed, might help correct such discrimination, but warns that many current tools risk automating rather than eliminating bias. He emphasizes the need for greater transparency, auditing, and ethical accountability in AI-driven hiring. Available at: <https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html>. Access: 23.06.2025.

involve not only diversifying training datasets and regularly auditing for disparate impacts, but also embedding normative commitments to inclusion at every stage of system development. Rather than aiming for mere neutrality, the platform could be optimized to correct for historical disadvantages, shifting its goal from replicating past success profiles to fostering a more diverse and representative workforce.

Thus, while AI offers new possibilities for expanding access and promoting inclusion, it also forces us to grapple with the fundamental ambiguity of *fairness* itself. Is fairness about treating everyone the same, or about compensating for historical and structural disadvantages? Should hiring algorithms mirror societal demographics or focus on individual aptitude, even if that reinforces inequalities? These are not merely technical questions; they are moral and political ones, embedded in how we define merit, justice, and belonging.

In light of this, the pursuit of equitable AI in hiring is not just about better code; it is about better values. As we shift decision-making power from individuals to machines, the underlying assumptions embedded in our systems become all the more consequential. If we want AI to advance fairness rather than obscure its absence, we must confront these dilemmas directly and design technologies that reflect the plural, contested, and evolving nature of justice in our societies.

The growing trust in artificial intelligence as a substitute for human rationality is often accompanied by a seductive promise: that algorithmic systems can deliver faster, more efficient, and, most importantly, more impartial decisions than humans ever could. Yet this promise masks a deep moral trap. As Brian Christian reminds us in *The Alignment Problem: How Can Artificial Intelligence Learn Human Values* (2020), the real challenge posed by AI is not its ability to process massive amounts of data or detect complex patterns, but its failure to align those patterns with legitimate human values. What happens when a system makes a technically correct decision that, from an ethical perspective, feels profoundly unjust?

If the Pymetrics case exposes the pitfalls of using AI in high-stakes decision-making, the normative response cannot be limited to *post hoc* corrections. As Orly Lobel (2022) provocatively argues, the goal is not to dismantle algorithmic systems but to reimagine how they are built from the ground up. She rejects the fatalistic view that algorithms are inherently discriminatory and instead insists that technology can and must serve equality if designed with that purpose in mind.

Lobel's vision is ambitious yet pragmatic. Fairness must be embedded at the start, not retrofitted later. This requires ongoing mechanisms for monitoring and reform, grounded in the understanding that every algorithm is a social artifact shaped by history, institutions, and existing inequities.

In this sense, Van Giffen et al. (2022) have emphasized the need for a multi-phase mitigation approach, involving interventions at the pre-processing, in-processing, and post-processing stages. Strategies include careful data curation and sample balancing, the use of fair learning algorithms, and group-disaggregated performance audits. Crucially, no single technique suffices on its own. Effective mitigation demands a contextual awareness of ethical risks and the continuous integration of normative values throughout the system's lifecycle. Such a systemic approach resists the temptation of purely technical fixes, acknowledging that the harms posed by algorithmic systems are ultimately rooted in broader social and political dynamics.

Barocas et al. (2023) offer a complementary and detailed framework for translating these commitments into technical practice. To meaningfully address fairness in automated hiring, companies must go beyond abstract commitments and adopt concrete practices that allow them to detect, mitigate, and respond to algorithmic discrimination.

Drawing from the framework proposed by Barocas et al. (2023), four key mechanisms have emerged as essential to a fairness-aware deployment of algorithmic systems: bias audits, regulatory audits, algorithmic risk assessments, and algorithmic impact evaluations. Each plays a distinct role in the lifecycle of an AI system and contributes to a culture of accountability. Yet implementing them is not a simple technical fix; it requires organizational will, interdisciplinary coordination, and cultural transformation.

Bias audits are designed to identify disparities in how algorithmic systems treat individuals from different social groups. As Barocas et al. explain, these audits are fundamentally diagnostic: they

help uncover whether a model's outputs are correlated with sensitive attributes like race or gender in ways that are not justifiable by business necessity or job-related criteria. In practical terms, for developers, this means analyzing the model's predictions using disaggregated data and fairness metrics such as *demographic parity*, *equalized odds*, or *predictive parity*.

Each offers a different way of understanding what it means for an algorithm to treat individuals and groups fairly. While these definitions are often in tension, applying them to real-world systems like Pymetrics can illuminate where interventions are needed and what values are being prioritized.

Demographic parity asks whether different demographic groups are selected at similar rates, regardless of underlying differences in qualifications. In the hiring context, this would mean that candidates from different gender or racial groups are recommended for jobs in roughly equal proportions. Applied to Pymetrics, a lack of demographic parity might be revealed if, for instance, significantly more men than women are flagged as high-potential candidates. Addressing this would likely require adjusting the algorithm's thresholds or training objectives to ensure more balanced representation, an approach aimed at correcting historical disparities in access to opportunity.

Equalized odds, by contrast, focus on the model's error rates. It requires that candidates from different groups have similar chances of being correctly or incorrectly classified. For Pymetrics, this would mean ensuring that highly qualified women are just as likely as their male counterparts to be correctly identified as strong matches, and not disproportionately filtered out. If false negatives are higher for one group, the system may be reinforcing existing inequalities under the guise of objectivity. Achieving equalized odds often involves recalibrating the model to reduce disparities in how it treats equally capable individuals.

Predictive parity takes yet another angle, asking whether the scores assigned by the model are equally meaningful across groups. In other words, if two candidates—say, one black and one white—receive the same "match" score, they should have similar chances of succeeding in the role. A lack of predictive parity would suggest that the model's predictions are more accurate for some groups than others, often because the definition of success was drawn from biased or homogeneous data. Improving predictive parity may involve redefining performance benchmarks in ways that better reflect a diverse workforce.

These metrics don't offer a single answer to what fairness means, but they help clarify the trade-offs involved in system design. In some cases, improving one metric may come at the cost of another. Still, applying them to systems like Pymetrics allows both developers and employers to move beyond vague commitments to fairness and instead make their values visible, measurable, and accountable.

For companies deploying these systems, the practical challenge lies in knowing what to ask from vendors and how to interpret audit results. Business teams should request evidence that pre-deployment audits were conducted with sufficiently diverse test data, and ask whether any corrective measures were taken when disparities were detected. This is not always straightforward: it requires data literacy, a basic understanding of statistical fairness concepts, and, critically, a willingness to act when results reveal uncomfortable truths about existing practices.

Regulatory audits, by contrast, are external mechanisms carried out by public institutions or regulators to assess whether an algorithmic system complies with legal standards related to discrimination, privacy, and transparency. Barocas et al. (2023) emphasize that while technical compliance may be necessary, it is rarely sufficient to ensure fairness. Compliance frameworks often lag behind technological innovation, and even well-intentioned systems can produce disparate impacts if deployed without oversight.

Recent legislative efforts in jurisdictions such as the European Union and Brazil have begun to address the challenges posed by algorithmic hiring. The EU Artificial Intelligence Act (AI Act), adopted in 2024, classifies AI systems used to evaluate job candidates as *high-risk*, subjecting them to requirements related to transparency, data governance documentation, and risk management. Providers must also ensure human oversight and adopt safeguards against discriminatory outcomes, aligning with broader EU commitments to fundamental rights. By framing hiring algorithms as high-risk, the AI

Act signals that automated systems used in employment deserve heightened scrutiny, but how these obligations will be interpreted and enforced in practice remains uncertain.

In Brazil, the proposed Bill No. 2338/2023 takes a similar risk-based approach, also identifying employment-related AI systems as high-risk and imposing obligations for audits, impact assessments, and transparency. While the bill is still under legislative debate, its inclusion of discrimination and bias as specific risks to be addressed reflects a growing awareness of how automated systems can reinforce historical inequalities, particularly in the Global South. Still, both frameworks face challenges in implementation and enforcement, and their effectiveness will ultimately depend on how legal principles translate into organizational practices and technical design choices.

In practice, companies must prepare to demonstrate not only what their systems do, but also how they were built, tested, and monitored. That means maintaining detailed documentation of training data sources, design decisions, fairness goals, and any internal governance procedures. Organizationally, this requires strong cross-functional collaboration between legal, technical, and HR teams, a task that can be culturally difficult in companies where these functions are siloed or operate with different priorities.

Algorithmic risk assessments are proactive exercises meant to identify potential sources of harm *before* a system is deployed. Barocas et al. (2023) treat this process as a core component of responsible AI development, emphasizing that harm is often foreseeable if organizations take the time to interrogate design choices. Developers, for example, should assess whether training data reflects historical biases, whether certain groups are underrepresented, and whether the optimization objectives align with fairness goals.

From an employer's standpoint, conducting or demanding algorithmic risk assessments can feel burdensome, especially when procurement timelines are tight or internal expertise is limited. But skipping this step invites reputational, legal, and ethical risks. Companies can start small: incorporate fairness checkpoints into pilot testing, ask vendors for their risk assessment protocols, and involve external experts when internal capacity is lacking. Risk assessment is not just a technical tool; it is a governance mindset that prioritizes foresight over damage control.

Algorithmic impact evaluations, finally, focus on how systems perform *after* deployment in real-world conditions. While risk assessments ask "What could go wrong?", impact evaluations ask "What is actually happening?". Barocas et al. (2023) argue that this stage is often neglected, even though it is essential for understanding whether the system's use aligns with its intended goals and whether it introduces unintended harms.

Companies should establish ongoing mechanisms to track outcomes across demographic groups, review hiring patterns over time, and ensure that the technology contributes to diversity and inclusion goals. This requires collaboration between technical teams (to collect and analyze the data), HR (to interpret it in the context of hiring strategy), and leadership (to act on the findings). It also requires confronting the possibility that a tool once celebrated as "objective" might perpetuate exclusion.

Implementing these four mechanisms—bias audits, regulatory audits, risk assessments, and impact evaluations—requires not just technical adaptation, but organizational transformation. Many companies are not set up to carry out these processes easily. Data may be siloed, accountability diffuse, and fairness not yet embedded in performance metrics. Moreover, cultural resistance often arises when fairness mechanisms challenge long-held beliefs about meritocracy or reveal structural imbalances in current practices.

Still, as Algorithmic impact evaluations emphasize, fairness is not a static property of systems, it is an ongoing process that must be embedded in institutional routines. For developers, this means designing for auditability, transparency, and ethical responsiveness from the start. For employers, it means building internal capacity, setting up governance structures, and creating incentives for responsible technology adoption.

These practices are neither quick nor cheap. But they are necessary if companies are serious about using AI not merely to optimize efficiency, but to promote fairness in one of the most consequential areas of human life: access to employment. A system that cannot demonstrate

how it avoids harm—or worse, how it justifies unequal treatment—has no place in a fair labor market.

Applying the tools outlined by Barocas et al. (2023) requires technical knowledge, regulatory awareness, and above all, ethical commitment. When deployed together, these mechanisms offer not only safeguards against discriminatory outcomes but also a roadmap for aligning technological innovation with democratic values and social inclusion.

These mechanisms, however, are not sufficient on their own. They must be embedded within a deeper epistemological shift in tech development: the move toward participatory design. Drawing from citizen science, inclusive urban planning, and user-centered design, participatory AI involves affected communities in every stage of system development, from problem framing and variable selection to validation and governance (Barocas et al., 2023).

Applied to Pymetrics, participatory design might have involved historically marginalized groups in early-stage testing, ensuring their behavioral norms were recognized and respected. It might have engaged "Diversity, Equity, and Inclusion" (DEI) experts, labor psychologists, and legal scholars in defining job-relevant traits. Most critically, it would have established ongoing feedback loops for iterative improvement based on lived user experiences.

This is the essence of fairness by design: not a checkbox, but a systemic reorientation. When combined with bias audits, risk assessments, impact evaluations, and participatory governance, we move from reactive correction to proactive justice. As Lobel insists, algorithms are not immutable, they are contested spaces, subject to norms, negotiations, and institutional duty. AI can be an "equality machine," but only if we choose to build it that way.

To support this vision, the Developer's Model for Responsible AI, presented below, builds on the categories proposed by Barocas et al. (2023), translating their conceptual framework into a structured, lifecycle-based approach that operationalizes fairness through concrete practices at each stage of system design, development, deployment, and monitoring.

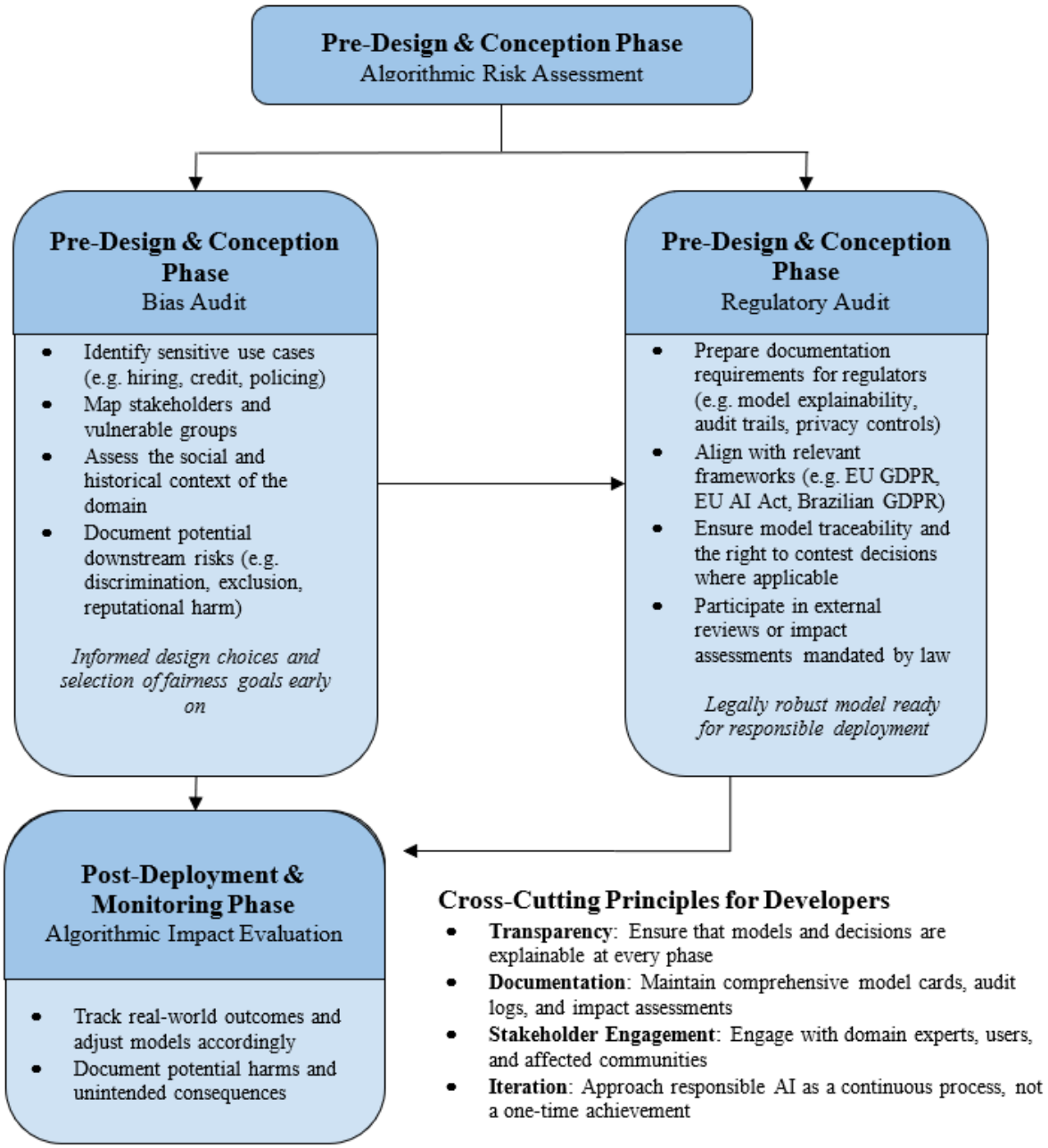

Figure 1. Developer´s Model for Responsible AI

Source: Authors' elaboration based on Barocas et al. (2023).

The *Developer's Model for Responsible AI* illustrates that fairness in algorithmic systems must be embedded across multiple layers, from the earliest stages of design to post-deployment oversight. This involves more than implementing technical safeguards: it requires inclusive design practices, robust bias audits, transparent decision-making processes, and meaningful avenues for appeal. When applied together, these mechanisms create an ecosystem of accountability, one that relies not on a single intervention but on the continuous interaction between internal controls, external audits, and public engagement.

Although audits play an increasingly central role in the governance of AI, they are not a panacea. Often limited to surface-level metrics like statistical parity, audits can overlook deeper normative questions, such as whose values are embedded in the system and what forms of inequality are being reproduced. Even more concerning, audits can become symbolic gestures, offering legitimacy without real accountability. In the cases of Amazon and

HireVue, the absence or superficiality of audits allowed discriminatory systems to persist until public exposure forced a reckoning.

This mismatch between the speed of algorithmic deployment and the slowness of institutional oversight reveals the importance of iterative, responsive governance. The *Developer's Model* proposes precisely this: a continuous loop that connects risk assessments, bias detection, regulatory alignment, and real-world impact evaluation.

Realizing this vision of fairness by design requires more than technical guidance; it calls for new institutional arrangements and shared governance models. Developers and companies must collaborate with regulators, civil society organizations, and the communities affected by these systems. This includes adopting enforceable measures such as algorithmic transparency reports, public model registries, pre-deployment impact assessments, and formalized appeal processes. These tools anchor fairness not in good intentions, but in verifiable practices.

Crucially, this model also demands participatory mechanisms. Democratic legitimacy in AI governance cannot be achieved without giving voice to those who are subject to algorithmic decision-making. Creating spaces where users, advocates, and experts can contribute to shaping the design and oversight of these systems is essential to preventing harm and building public trust.

In the end, fairness by design is not simply a technical challenge; it is a political undertaking. It asks us to consider: who defines what is fair? Who gains and who loses from algorithmic systems? And how do we ensure accountability when those systems fail? AI is not a neutral force. It reflects and amplifies existing structures of power. Whether it reinforces exclusion or becomes a tool for equity depends not on the algorithm alone, but on the governance choices we make collectively.

5. Conclusion: Equality by Design — Reclaiming the Political Imagination of AI

The promise of artificial intelligence systems that are fairer, more rational, and more efficient than humans now confronts a fundamental paradox: the more technically advanced these systems become, the more apparent their normative limitations grow. Throughout this paper, we have argued that algorithmic fairness is not a feature that can be toggled on with a click; it is a social and political construction that must be continuously embedded, monitored, and renegotiated.

In the first section, the cases of Pymetrics, HireVue and Amazon served as a starting point to show how even well-intentioned technological solutions, designed to promote meritocracy and diversity, can end up reproducing, or even intensifying, structural inequalities. The audit conducted by Christo Wilson and his team (Wilson et al., 2021) demonstrated that these systems do not fail by accident; they fail because they are built on business models and epistemologies that often overlook the complexity of what it means to be fair. As Roman V. Yampolskiy (2024) highlights, the opaque and unpredictable nature of AI systems makes these distortions even harder to control. Meredith Broussard (2018, 2023) reminds us that there is no such thing as neutral code; technical design always reflects social and political choices.

In the second section, we moved from critique to normative theory. Drawing on the work of Brian Christian, Barocas, Hardt, Narayanan, Kearns, and Roth, we explored the conceptual dilemmas of defining fairness and the multiple approaches to understanding justice in algorithmic systems. We argued that the real challenge is not to eliminate all bias, a task that is both unrealistic and epistemologically flawed, but to decide, publicly and transparently, which forms of unequal treatment are morally defensible and which ones reinforce historical injustices. We discussed the foundations of unfair discrimination and competing conceptions of equal opportunity, showing how models such as statistical parity and the Pareto frontier can be used to navigate trade-offs between accuracy and equity.

In the third section, we adopted a propositional stance, guided by the work of Orly Lobel (2022) and Barocas et al. (2023). Algorithmic justice cannot be a corrective measure; it must be a design imperative. This means embedding fairness from the earliest stages of system development, but also maintaining ongoing accountability mechanisms once the system is deployed. We discussed tools such as bias audits, risk and impact assessments, and the need for participatory design practices that actively include affected communities in deciding what, ultimately, should be optimized.

What emerges from this trajectory is a provocative but necessary conclusion: justice is not a product to be purchased; it is a process to be built. And like all processes, it requires participation, oversight, revision, and a willingness to recognize limits and learn from failure. Algorithmic systems will not replace our ethical debates; they will make them more urgent, more visible, and more complex.

If we want AI to play a constructive role in building a more equitable society, technical advancement alone is not enough. These systems must also be institutionally accountable, socially conscious, and normatively committed. In the end, the fight for algorithmic justice is not a battle between engineers and lawyers; it is a struggle over the future of digital democracy.

References

Ajunwa, I. (2020). The paradox of automation as anti-bias intervention. *Cardozo Law Review, 41*, 1671–1720.

Ajunwa, I. (2023). *The quantified worker: Law and technology in the modern workplace*. Cambridge University Press.

Andrews, L., & Bucher, H. (2022). Automating discrimination: AI hiring practices and gender inequality. *Cardozo Law Review, 44*, 145–201.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. The MIT Press.

Bigman, Y. E., Heller, D., Kim, J. K., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General, 152*(1), 4–27. https://pubmed.ncbi.nlm.nih.gov/35758989/

Broussard, M. (2019). *Artificial unintelligence: How computers misunderstand the world*. The MIT Press.

Broussard, M. (2023). *More than a glitch: Confronting race, gender, and ability bias in tech*. The MIT Press.

Christian, B. (2021). *The alignment problem: How can artificial intelligence learn human values?* Atlantic Books.

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

Gasser, U., & Mayer-Schönberger, V. (2024). *Guardrails guiding human decisions in the age of AI*. Princeton University Press.

Gillis, T. B. (2022). The input fallacy. *Minnesota Law Review, 106*, 1175–1256.

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. *Journal of Business Ethics, 178*, 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Kearns, M., & Roth, A. (2020). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Legg, M., & Bell, F. (2020). *Artificial intelligence and the legal profession*. Hart Publishing.

Lobel, O. (2022). *The equality machine: Harnessing digital technology for a brighter, more inclusive future*. PublicAffairs.

Mullainathan, S. (2019). Biased Algorithms Are Easier to Fix Than Biased People. *The New York Times*. https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Random House.

Páez, A. (2021). Negligent algorithmic discrimination. *Law and Contemporary Problems, 84*, 19–33.

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 (FAT* '20)*. https://dl.acm.org/doi/10.1145/3351095.3372828

Sheard, N. (2022). Employment discrimination by algorithm: Can anyone be held accountable? *UNSW Law Journal, 45*(2), 617–648.

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research, 144*, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Wilson, C., Ada, S., Shreshta, A., & Mislove, A. (2021). Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*. https://mislove.org/publications/Pymetrics-

FAccT.pdf

Yampolskiy, R. V. (2024). *AI: Unexplainable, unpredictable, uncontrollable*. Chapman and Hall/CRC.

Author Profile

Alexandre Pacheco da Silva: PhD in Science and Technology Policy from the Department of Geosciences at the University of Campinas (Unicamp). Alexandre Pacheco da Silva holds both a Master's and a Bachelor's degree in Law from the São Paulo Law School of Fundação Getulio Vargas (FGV São Paulo Law School). Professor in the undergraduate and graduate programs at FGV São Paulo Law School, where he teaches courses on digital law and intellectual property. Coordinator of the Center for Education and Research in Innovation.

Enya Carolina Silva da Costa holds a Master's degree in Public Law and a Bachelor of Laws from the University of São Paulo (USP), as well as a Licence en Droit from Université Jean Moulin Lyon III, obtained through the Partenariat International Triangulaire d'Enseignement Supérieur (PITES) program. Currently serves as Project Leader and Researcher at the Center for Education and Research on Innovation (CEPI) of FGV São Paulo Law School, Brazil.