scisearch INTERNATIONAL PEER-REVIEWED JOURNALS

RESEARCH PAPER

Theoretical Foundations of Algorithmic Fairness in Two-Sided Hiring Marketplaces: Interventions for Reducing Discrimination in Job Matching

Mohammad Hassan

Pabna University of Science and Technology, Department of Computer Science and Engineering, Dhaka–Pabna Hwy., Pabna, Bangladesh

Abstract

Digital hiring platforms increasingly mediate how workers and firms find each other, and their matching algorithms have become part of the institutional fabric of labor markets. In two-sided hiring marketplaces, decisions are rarely a single pass from application to offer; instead, platforms combine search, ranking, screening, messaging, and interview scheduling into iterative processes that allocate attention and opportunity. Algorithmic fairness in this setting therefore concerns not only whether a model treats similar candidates similarly, but also how platform rules shape exposure, competition, and bargaining across both sides. This paper develops theoretical foundations for fairness in two-sided job matching by framing the marketplace as a coupled socio-technical system with strategic actors, incomplete information, and feedback loops. It distinguishes discrimination that arises from historical measurement error, preference-based sorting, platform design choices, and endogenous responses by employers and job seekers. Building on these distinctions, it analyzes intervention families that aim to reduce discrimination while preserving market functionality, including constrained ranking, exposure-aware matching, calibrated screening, incentive-compatible auditing, and governance mechanisms that control data and objective design. The discussion emphasizes failure modes that are specific to two-sided settings, such as shifting discrimination across stages, attention externalities, and equilibrium effects in which group-level outcomes change even when per-decision constraints hold. The paper concludes by outlining evaluation principles that treat fairness as a system property, integrating distributional parity, individual consistency, and welfare robustness under realistic behavioral adaptation.

1. Introduction

Two-sided hiring marketplaces are now a common mechanism for coordinating employment search (1). Rather than relying solely on direct applications or informal networks, job seekers and employers increasingly interact through platforms that provide structured profiles, recommendation feeds, search interfaces, automated screening, and communication tools. These platforms differ in surface features, but they share a core function: they reduce transaction costs by helping both sides discover, evaluate, and engage with potential matches. The same mechanisms that improve efficiency, however, can also reproduce or amplify discriminatory outcomes. Discrimination can appear as unequal exposure to opportunities, systematically lower response rates, differential ranking positions, or disparate conversion at later stages such as interviews and offers. In a platform context, these effects can arise without explicit use of protected attributes because proxies, historical patterns, and platform feedback can carry information that correlates with protected status. Fairness concerns thus extend beyond whether a single model is unbiased; they involve whether the entire matching process distributes opportunity in a defensible way.

Algorithmic fairness research often begins with a supervised learning problem: a classifier or score predicts an outcome, and fairness constraints or metrics compare error rates or score distributions across groups. Hiring marketplaces rarely fit this template cleanly. They involve multiple actors with distinct objectives, and the algorithm's outputs are not final decisions but inputs to a dynamic process. A ranking model influences which candidates are seen; visibility influences which candidates apply; employer interactions influence what labels are observed for future training; and job seeker choices influence the composition of the applicant pool. This coupling means that fairness constraints applied at one stage may be offset by behavior at another stage, and that metrics computed on observed data may be distorted by selection. A platform can satisfy a statistical constraint on the set of viewed candidates while still yielding discriminatory downstream outcomes if employers respond differently to different groups, or if constraints change the equilibrium of search and application.

This paper focuses on theoretical foundations that are specific to two-sided hiring marketplaces. The central claim is not that one metric or one intervention is universally appropriate, but that the structure of two-sided interaction changes what fairness means, how discrimination manifests, and which interventions are plausible (2). The analysis treats fairness as a system-level property that depends on exposure, strategic behavior, and institutional constraints. The marketplace setting introduces attention as a scarce resource: employers

Table 1. Core elements of a two-sided hiring marketplace and their relationship to fairness.

Component	Role in marketplace	Fairness relevance
Job seekers	Supply labor, create profiles, choose where to apply and which offers to accept	Affected by exposure, measurement error, and constraints on applications and interviews
Employers	Specify vacancies, set requirements, evaluate candidates, extend offers	Preferences, evaluation procedures, and filters can introduce or amplify bias
Platform	Represents information, ranks and recommends, routes messages, tracks outcomes	Algorithmic and interface design directly shape attention and opportunity distribution
Institutional context	Law, regulation, professional norms, contracts	Determines which attributes can be used, what auditing is required, and liability regimes
Broader social structure	Education, geography, caregiving, networks, labor institutions	Generates structural disparities that appear in features, labels, and feasible job options

Table 2. Fairness objects spanning representation, allocation, and outcomes in hiring marketplaces.

Fairness object	Guiding question	Example concern
Representation	Do features depict skills and experience comparably across groups?	Resume parsing underrates nonstandard job titles common in some communities
Ranking & exposure	How is scarce attention allocated across candidates and jobs?	Protected groups appear consistently lower in ranked lists despite similar fit
Engagement & conversion	Are transitions between stages equitable, conditional on qualifications?	Equal views but lower contact or interview rates for a particular demographic
Matches & welfare	Are final job outcomes and job quality comparable across groups?	Underrepresented candidates hired into lower-wage or lower-mobility roles
Procedure & contestability	Can actors understand, correct, and challenge decisions?	Candidates lack channels to fix profile errors that suppress opportunities

have limited time to review candidates, and job seekers have limited capacity to apply and interview. Algorithms allocate this resource through ranking, recommendation, and filtering, and the allocation can create externalities. Increasing exposure for one set of candidates can reduce exposure for others, and these trade-offs may be hidden if evaluation uses metrics that ignore position effects or ignores that employers do not examine the full list.

A theoretical treatment must separate several phenomena that are often conflated in practice. One phenomenon is measurement: profiles and resumes encode skills imperfectly, and measurement error is correlated with group membership due to unequal access to credentialing, biased evaluation in past employment, or differential documentation norms. Another phenomenon is preference: employers may have tastes that lead them to choose candidates differently even with identical productivity. A third is technology: platform features such as default filters, search facets, and ranking objectives can advantage some profiles. A fourth is dynamics: once the platform changes ranking rules, both sides may adapt, changing the distribution of applicants, the content of profiles, and the meaning of observed outcomes. Fairness interventions that do not attend to these categories can misdiagnose the source of discrimination and can shift harms across stages.

The paper proceeds by modeling the platform as an intermediary that implements rules for information revelation, ordering, and matching. It highlights how choices about objective functions, constraints, and feedback measurement shape both group and individual outcomes. It then characterizes a design space of interventions and connects them to assumptions about the source of discrimination. Some interventions target allocation of exposure directly, others adjust scoring and calibration, others alter information flow by limiting or structuring what attributes can be seen, and others focus on governance such as auditing,

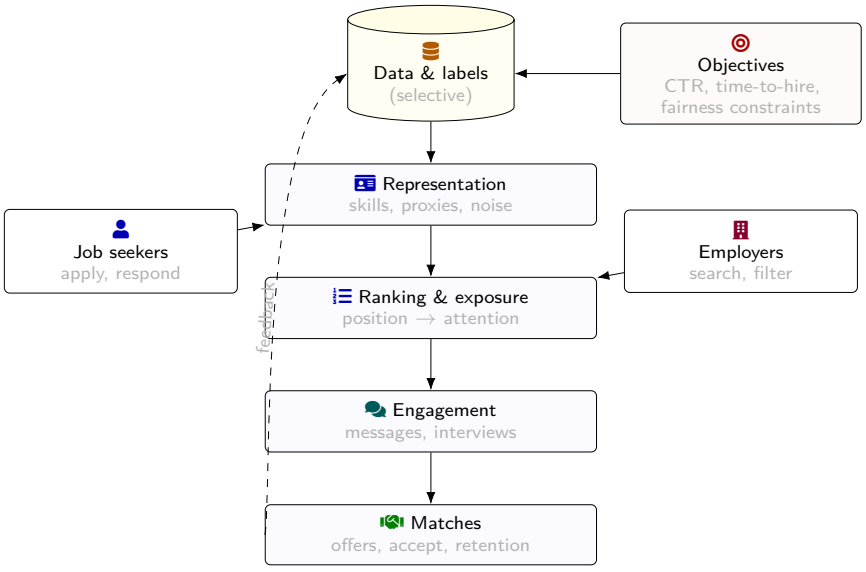

Figure 1. Iterative hiring on a two-sided platform: representations and rankings allocate scarce attention, downstream engagement gates who generates labels, and the resulting selective feedback shapes subsequent model updates. System-level fairness depends on exposure and stage transitions, not just per-decision scoring.

contestability, and data stewardship. Throughout, the emphasis is on theoretical clarity: what property an intervention aims to guarantee, what assumptions are needed for that guarantee to hold, and how the guarantee can fail when behavior adapts (3).

The goal is not to propose a single operational definition of fairness that resolves normative disagreement, but to provide a foundation for reasoning about trade-offs in marketplace contexts. Hiring markets involve legitimate heterogeneity in preferences, constraints, and job requirements, and they are embedded in legal and institutional norms that vary by jurisdiction. A platform may be accountable to anti-discrimination law, to professional standards, to contractual obligations with employers, and to expectations from job seekers. These constraints shape what interventions are feasible. A theory that treats the platform as a simple classifier misses these institutional realities, while a theory that treats fairness as purely philosophical can miss mechanisms that determine outcomes. The analysis here aims to bridge these views by connecting normative desiderata to operational mechanisms in two-sided systems.

2. Two-Sided Hiring Marketplaces as Socio-Technical Systems

A two-sided hiring marketplace can be characterized by participants, information, and mechanisms. On one side are job seekers who possess skills, preferences, constraints, and private information about their own fit and willingness to accept offers. On the other side are employers with vacancies, job requirements, budgets, and internal processes for evaluating applicants. The platform intermediates by providing a representation of candidates and jobs, a means of search and discovery, and a sequence of interactions such as applications, messages, interviews, and offers. From a system perspective, the platform chooses how to represent information, how to rank or recommend, how to throttle or expand contact, and how to measure outcomes. Each of these choices can influence who meets whom, and thus which employment relationships form (4).

The two-sided nature matters because the relevant outcomes are joint. A job seeker can be treated favorably by an algorithm in isolation, but still experience low hiring probability if the algorithm directs them toward saturated job categories or employers with low responsiveness. Similarly, an employer can be presented with a diverse slate of candidates but still hire homogeneously if evaluation processes are biased (5). Fairness therefore cannot be reduced to a property of candidate scoring alone. It includes the mapping from candidate features and employer features to the probability of mutual engagement, and it includes the distribution of these probabilities across groups. Moreover, in many platforms the matching process is staged. Early stages allocate exposure, middle stages allocate conversation and interviews, and late stages allocate offers. Discrimination can be concentrated at a particular stage, and interventions targeted at the wrong stage may have limited effect.

A useful conceptualization is that the platform allocates attention. Employers allocate a limited number of views, clicks, and interviews, and job seekers allocate a limited number of applications and responses. The platform can increase the efficiency of attention by ordering candidates and jobs in ways that increase the chance that attention is spent on high-fit matches. However, when attention is scarce, small differences in ranking position can lead to large differences in outcomes. This is a key mechanism by which discrimination can manifest even when average scores are close across groups. If candidates from a protected group are consistently placed slightly lower due to noisy proxies, they may receive substantially fewer views and thus fewer opportunities. An intervention that equalizes average scores without addressing rank position and exposure

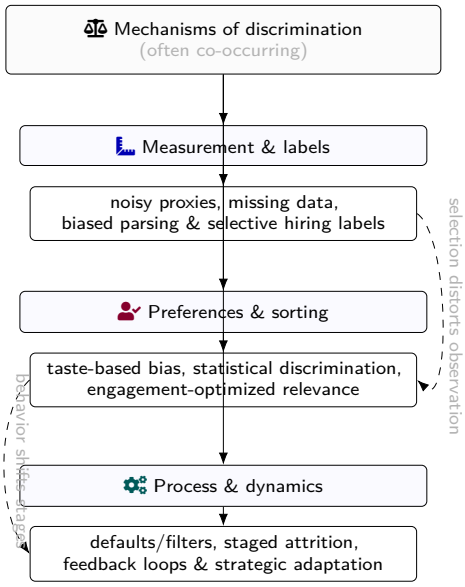

Figure 2. A mechanism-based view separates (i) measurement and label bias, (ii) preference-driven sorting amplified by engagement objectives, and (iii) process and dynamic effects (defaults, multi-stage attrition, and feedback). Interventions that target only one mechanism can relocate disparities to other stages.

may therefore be insufficient.

Two-sided marketplaces also exhibit congestion. Popular employers may receive large numbers of applications and respond to a small fraction (6). Candidates who are directed disproportionately toward congested employers can experience lower conversion, even if their qualifications are high. Congestion can correlate with group membership if historical inequities influence which employers are perceived as accessible or desirable. The platform's ranking and recommendation can amplify this by promoting the same high-visibility employers to many candidates. In such a setting, fairness involves not only exposure of candidates to employers but also exposure of candidates to reachable opportunities, where reachability includes responsiveness and realistic hiring probability.

Another feature is strategic behavior. Employers may use the platform to minimize screening costs, and job seekers may adapt resumes, skills claims, and application strategies to what they believe the algorithm favors. If a platform introduces a fairness constraint in ranking, employers might respond by adjusting filters or increasing reliance on off-platform signals. Job seekers might respond by applying more broadly, changing profile content, or shifting to different job categories. These responses can either support or undermine fairness goals. A theory of fairness that ignores strategic response risks overestimating impact, because measured fairness on the immediate ranking may not translate into equitable final outcomes after adaptation.

Information asymmetry is also central. Employers do not observe candidate productivity directly, and candidates do not observe employer willingness to hire or workplace conditions fully. The platform often holds additional information, such as aggregate response rates, past hiring patterns, or inferred attributes. Fairness questions arise about how the platform reveals or withholds this information. Revealing too much can enable discrimination, while revealing too little can reduce match quality or prevent candidates from avoiding biased employers. A platform may restrict explicit protected attributes while still revealing proxies through names, schools, or employment gaps. Conversely, it may choose to anonymize profiles at early stages, which can reduce bias but can also reduce the ability of candidates to signal legitimately relevant experiences. The fairness implications of information design are inherently tied to the two-sided nature of the marketplace because information is used by both sides to decide whether to engage.

Feedback loops arise because platform decisions shape the data used to train future models. If an algorithm ranks certain groups lower, those groups receive fewer opportunities, which can reduce observed positive outcomes, which then reinforces the model's belief that those groups are lower quality. This is a data-generating process shaped by the algorithm, not a passive reflection of underlying ability. In hiring, the problem is amplified because labels are highly selective. Observed outcomes like interviews or offers are conditional on being seen, being contacted, and being evaluated. The platform rarely observes counterfactual outcomes for candidates who were not shown. Fairness evaluation must therefore address selection bias and the possibility that historical data encode past discrimination.

Institutional context shapes what the platform can do. Anti-discrimination law often distinguishes disparate treatment from disparate impact, and it may limit the use of protected attributes while still allowing interventions that mitigate impact. Employers may have obligations to justify hiring decisions, and platforms may have obligations to explain recommendation logic. These constraints can limit which fairness interventions are feasible. For example, a platform may be prohibited from explicitly using protected group membership for certain decisions,

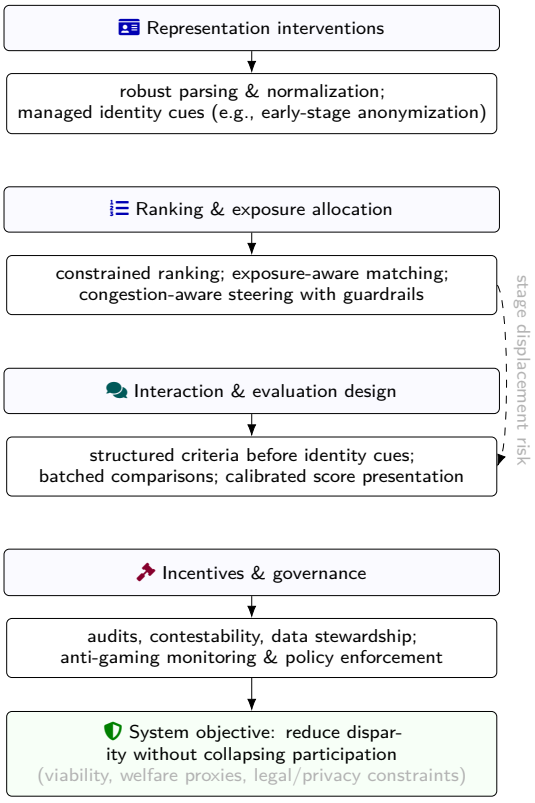

Figure 3. Intervention families map naturally to pipeline control points: improving measurement, allocating exposure under constraints, shaping human evaluation to reduce stereotype reliance, and enforcing accountability through monitoring and governance. Their effectiveness depends on which discrimination mechanisms dominate and how participants adapt.

even if doing so could help correct inequities (7). Alternatively, a platform may be permitted to use group information for auditing and monitoring but not for ranking. The theoretical framework must therefore accommodate interventions that operate with partial access to protected attributes.

Finally, the marketplace is embedded in broader social structures. Education, geography, caregiving responsibilities, and network access influence both skill acquisition and job search. These factors can correlate with protected status and create structural disadvantage that is not reducible to employer bias. A platform that optimizes for short-term placement may inadvertently reproduce these structures by steering candidates toward roles that match historical patterns. A fairness-oriented approach must decide whether the platform's responsibility is limited to avoiding direct discrimination or extends to mitigating structural disadvantage through exposure and opportunity creation. These normative questions are outside the scope of purely technical metrics, but they influence which theoretical notions of fairness are appropriate.

3. Conceptual Definitions of Fairness in Matching and Ranking

Fairness in hiring marketplaces can be framed at different levels of abstraction. At a local level, one can ask whether a scoring model treats similar candidates similarly. At a global level, one can ask whether the distribution of outcomes across groups is equitable. In a two-sided system, both levels interact because local scoring affects global exposure, and global constraints can alter local decisions. A theoretical foundation therefore benefits from distinguishing several fairness objects: the fairness of information representation, the fairness of ranking and exposure, the fairness of engagement and conversion, and the fairness of final matches and welfare.

A representation is fair if it does not systematically distort the depiction of skills and experiences for some groups relative to others (8). Distortion can occur through missing data, biased feature extraction, or differential interpretability. For example, free-form text resumes may encode nonstandard phrasing more common in some communities, and a model trained on standard corporate language may underestimate relevant experience. A platform might convert resumes into structured features, and the conversion process can erase information that is more prevalent in certain groups. Fairness at the representation level is about measurement quality and invariance: whether a given underlying skill is mapped to comparable feature values regardless of group. While this is not the full fairness problem, it is foundational because later stages depend on the representation.

Ranking and recommendation fairness concerns how the platform orders candidates for employer attention or jobs for candidate attention. In a one-sided recommender, fairness might be defined as equal probability of being recommended across groups, possibly conditioned on relevance. In a hiring

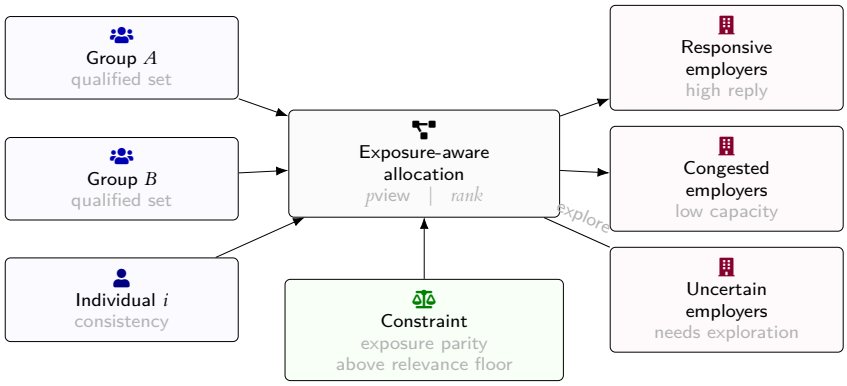

Figure 4. Exposure-aware matching treats attention as the scarce resource being allocated across a candidate–employer network. A platform can enforce distributional constraints on expected views while remaining sensitive to congestion (capacity limits) and using limited exploration to reduce selection bias about uncertain employers.

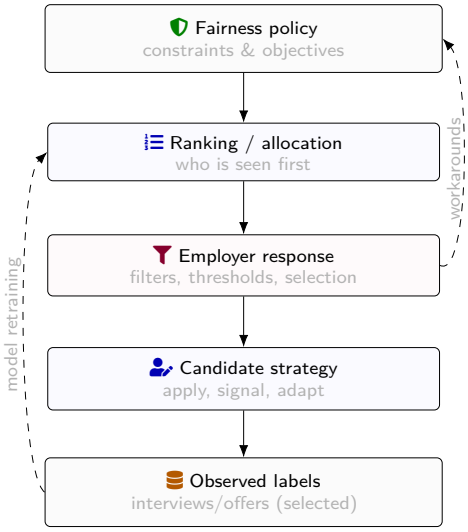

Figure 5. A fairness constraint can change behavior: allocation affects employer filtering, filtering reshapes candidate strategies, and the resulting selective labels feed back into retraining. Marketplace-specific failure modes include stage displacement (bias moving downstream), metric gaming, and equilibrium shifts that alter group outcomes even when per-query constraints hold.

marketplace, relevance is itself contested and partly endogenous because employer preferences may reflect discrimination. A useful conceptual distinction is between relevance as predicted productivity and relevance as predicted engagement. A platform that optimizes for engagement, such as click-through or message response, may learn to present candidates who match employers' biased preferences, thereby amplifying discrimination while appearing to improve user metrics. Conversely, optimizing for predicted productivity might better align with equal opportunity goals, but productivity is rarely observed directly and may itself be influenced by past inequities. The theoretical issue is that a fairness constraint must specify which notion of relevance is legitimate.

Exposure fairness recognizes that outcomes depend on position and attention. Two candidates with equal score distributions may receive different exposure because of rank ordering, interface design, and employer behavior. Exposure can be quantified conceptually as the probability of being viewed, contacted, or considered given a platform's presentation policy (9). Fairness might require equal exposure for equally qualified candidates, or minimum exposure floors for underrepresented groups to counteract feedback loops. The two-sided setting complicates this because exposure interacts with employer capacity and search strategies. If employers only review a few profiles, then exposure allocation becomes a high-stakes resource, and fairness constraints that ignore position effects may be superficial.

Engagement fairness concerns the transitions between stages. A candidate may be shown, but not contacted; contacted, but not interviewed; interviewed, but not offered. Discrimination can occur in any transition, and the platform may have different degrees of control at each stage. If an employer chooses not to contact candidates from a protected group, then equal exposure at the top of the funnel may not translate into equal interviewing rates. The platform might intervene by altering the composition of shown candidates, by adjusting messaging prompts, by requiring structured evaluation, or by implementing policies that limit discriminatory filtering. Each intervention raises questions about autonomy and legitimate employer discretion. The

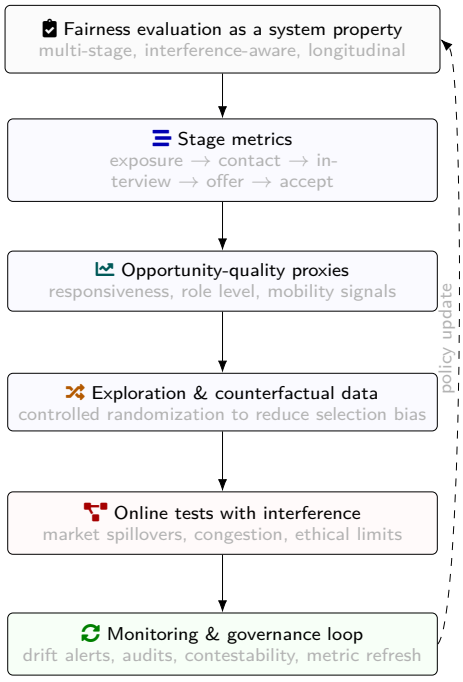

Figure 6. Evaluation requires more than offline parity: it tracks stage-wise transitions, checks whether exposure translates into access to desirable opportunities, gathers counterfactual evidence via limited exploration, and designs online tests that account for marketplace interference. Long-run governance closes the loop by responding to drift, adaptation, and metric gaming.

theory must articulate what is meant by fairness in transitions: whether it is parity in transition probabilities conditioned on qualifications, parity in error rates in predicting transitions, or parity in realized outcomes after accounting for preferences.

Match fairness concerns the final pairing of worker and job. In matching theory, a match can be evaluated in terms of stability, efficiency, and welfare. A fairness notion at this level might require that no group systematically receives worse job quality, lower wages, or less career advancement, conditional on qualifications and preferences. However, job quality is multi-dimensional and partly subjective. Moreover, a platform may not observe wages or long-term outcomes, and it may not control final offers (10). This makes match-level fairness both normatively central and technically difficult. A theoretical foundation can still treat match fairness as the target and use exposure and engagement fairness as proximate levers, but it must acknowledge the gap between controllable intermediates and ultimate outcomes.

Individual and group fairness can conflict. Individual fairness aims for consistency: similar individuals should receive similar treatment. Group fairness aims for parity: outcomes across groups should satisfy some relationship. In hiring, similarity is contested because qualifications are high-dimensional and measured with error. A platform can enforce consistency with respect to its feature space, but if the feature space is biased, then consistency can entrench inequity. Group constraints can correct for representation bias but may create perceived unfairness for individuals who are displaced in ranking. A theoretical account should treat these as different fairness desiderata rather than as competing implementations of the same concept.

Two-sided systems introduce a further complication: fairness for candidates may conflict with fairness for employers. Employers may have legitimate constraints such as location, schedule, certification requirements, and team composition needs. If a platform enforces strict candidate-side parity in exposure without regard to employer constraints, it may reduce match quality or increase employer search costs, potentially causing employers to leave the platform. If employers leave, opportunities shrink for all candidates, including those the intervention aimed to help. Therefore, fairness must be considered alongside platform viability and mutual value creation. This does not mean that fairness should be sacrificed for engagement metrics, but it implies that fairness interventions should be analyzed for their equilibrium impact on participation and for their distributional consequences across both sides (11).

A constructive approach is to define fairness relative to a baseline. The baseline can be the platform's unconstrained algorithm, the historical market without the platform, or a hypothetical nondiscriminatory market. Each baseline yields different interpretations. If fairness is defined relative to the unconstrained algorithm, then interventions are evaluated by how they redistribute exposure compared with current practice. If defined relative to a nondiscriminatory benchmark, then the platform is judged by how closely it approximates outcomes that would occur if employers evaluated productivity without bias and if measurement were accurate. The choice of baseline is normative, but theoretical clarity requires that it be explicit because it shapes what counts as discrimination and what counts as correction.

Another element is temporal fairness. Hiring is a repeated

Table 3. Illustrative mechanisms through which discrimination can emerge in platform-mediated hiring.

Mechanism type	Description	Illustrative platform example
Measurement	Features or labels are noisy in group-dependent ways	Gaps in employment automatically penalized, disproportionately affecting caregivers
Preference	Employers value or devalue groups independently of productivity	Click-based ranking learns to favor demographic profiles that historically get more clicks
Process	Design choices in filters, defaults, and interfaces shape who is seen and evaluated	Default school prestige filter screens out candidates from non-elite institutions
Dynamic feedback	Algorithmic decisions change the data used for future training	Lower-ranked candidates receive fewer opportunities, reinforcing low estimated quality
Selection / observability	Outcomes observed only for a biased subset of interactions	Hiring labels recorded only for candidates shown to highly responsive employers

Table 4. Key stages of the platform-mediated hiring funnel and associated disparity metrics.

Stage	Platform levers	Potential disparity	Example metric
Exposure	Ranking, recommendation, search interfaces	Unequal probability of being viewed by relevant employers	Share of impressions or views by group and qualification band
Contact	Messaging tools, prompts, candidate routing	Differential likelihood of being contacted after being viewed	Contact rate conditional on exposure
Interview	Scheduling tools, structured workflows	Lower interview invitations for equally qualified candidates	Interview selection rate per contacted candidate
Offer	Decision support, evaluation templates	Fewer offers or lower-quality offers for similar productivity	Offer rate and average job-quality proxy by group
Acceptance	Information about jobs, employer reputation signals	Different acceptance patterns due to perceived bias or risk	Acceptance rate and post-hire retention differentials

game: candidates may apply multiple times, employers may hire repeatedly, and reputations evolve. Fairness constraints applied per transaction may not ensure fairness over time. For example, a platform might ensure that each search results page is balanced, but if some candidates churn due to discouragement, or if employers learn to bypass the platform's constraints, then long-run outcomes may still be inequitable. Temporal fairness considers whether cumulative exposure and cumulative hiring probabilities are equitable over a candidate's search horizon. It also considers whether interventions create sustainable improvements or merely short-term shifts.

A final conceptual dimension concerns contestability and explanation. Even if outcome parity is achieved, candidates may perceive the process as unfair if they cannot understand or challenge decisions (12). In hiring, perceptions of procedural justice matter because they affect trust and participation. A marketplace that aims to reduce discrimination may need mechanisms for candidates to verify that their profiles are accurately represented, to correct errors, and to appeal adverse outcomes. These mechanisms are not purely technical constraints on ranking; they are part of a fairness framework because they influence who benefits from the platform and whether disadvantaged groups can effectively participate.

4. Mechanisms of Discrimination and Feedback in Hiring Platforms

Discrimination in two-sided hiring marketplaces can be generated by multiple mechanisms that interact. Distinguishing mechanisms is important because interventions that address one mechanism may not address others, and some interventions can worsen certain forms of discrimination while improving others. A theoretical account therefore benefits from a mechanism-based taxonomy: measurement mechanisms, preference mechanisms, process mechanisms, and dynamic mechanisms.

Table 5. Families of interventions targeting different points in the two-sided matching process.

Stage focus	Intervention type	Intended effect	Possible failure mode
Representation	Robust parsing, normalization, identity blinding	Reduce differential measurement error and direct discrimination	Removal of cues leads employers to rely on noisier proxies
Ranking & exposure	Constrained ranking, exposure-aware allocation	Balance attention across groups while preserving relevance	Employers perceive slates as less tailored and increase filtering or churn
Interaction & evaluation	Structured reviews, batched comparisons, prompt design	Reduce reliance on stereotypes in mid-funnel decisions	Bias shifts to later, less observable stages such as interviews
Incentives & governance	Reputation signals, auditing, policy constraints	Make discriminatory behavior more costly and visible	Superficial compliance or gaming of monitored metrics
Participation & supply	Training links, credential pathways, targeted outreach	Expand feasible qualification sets for disadvantaged groups	Steering into narrow tracks that resemble occupational segregation

Measurement mechanisms arise when the platform's features are noisy proxies for productivity or fit, and the noise is not uniform across groups. Differences can come from unequal access to credentialing, varying conventions in resume writing, bias in past evaluations that affect prior job titles and references, or the presence of gaps in employment due to caregiving or health. If the model interprets gaps as negative signals without context, it can penalize groups disproportionately. Measurement mechanisms also include label bias. If training labels are interviews or hires, those labels reflect employers' past choices, which may be discriminatory. A model trained on these labels may learn to replicate discrimination, especially if the platform optimizes directly for the probability of being hired based on historically biased decisions.

Preference mechanisms arise when employers have tastes that lead them to prefer candidates from certain groups, holding productivity constant. These tastes can be explicit or implicit, and they can be mediated by proxies such as names, schools, or neighborhoods. In a platform, preference-based discrimination can be amplified by optimization for engagement (13). If employers tend to click more on certain demographics, then a recommender trained to maximize clicks will learn to present those demographics more often. From the platform's perspective, it is optimizing a user satisfaction metric, but from a fairness perspective, it is amplifying discriminatory preferences. Preference mechanisms also include statistical discrimination in which employers use group membership as a proxy for unobserved productivity due to perceived differences in signal reliability. Even if employers are not prejudiced, they may treat signals from different groups differently if they believe those signals have different meaning. This can be a rational response to information asymmetry, but it still produces disparate outcomes.

Process mechanisms are generated by platform design choices that shape who enters the funnel and how they progress. Examples include default filters, the ordering of search facets, the prominence of certain fields, and the thresholds used for screening. If the platform encourages employers to filter by school prestige, it can disproportionately exclude candidates from groups with less access to elite institutions. If the platform's application process requires lengthy assessments, it can disadvantage candidates with limited time or limited access to stable internet. Process mechanisms also include the sequencing of information revelation. If photos are shown early, they may trigger appearance-based bias. If names are shown early, they may trigger ethnicity-related bias. Conversely, if information is withheld too long, employers may feel deceived or may infer group membership through other signals, potentially reintroducing bias later.

Dynamic mechanisms arise from feedback and adaptation. A platform's ranking affects which candidates apply where, which affects employer beliefs about the applicant pool, which affects employer response behavior, which affects observed labels, which affects future ranking (14). This loop can create runaway effects. If candidates from a protected group receive fewer responses, they may apply less or leave the platform, reducing their representation, which further reduces employer exposure to them, which further entrenches stereotypes. Dynamic mechanisms also include strategic manipulation. Job seekers may learn to optimize keywords or credentials that the algorithm rewards. If access to those credentials is unequal, then strategic adaptation can widen disparities. Employers may also adapt by using external screening tools, by shifting to channels where they can exercise more discretion, or by posting requirements that indirectly screen out certain groups. These adaptations can undermine platform-level fairness constraints.

Selection effects make discrimination difficult to measure. The platform observes outcomes only for interactions that occur.

Table 6. Selected trade-offs that arise when formalizing fairness objectives in two-sided markets.

Trade-off	Dimensions	Illustrative tension
Exposure vs. relevance	Redistribution of attention vs. employer-perceived match quality	Increasing visibility for underrepresented candidates may lower click-through for some employers
Parity vs. opportunity quality	Outcome parity vs. quality of jobs and employers	Equalizing exposure by sending candidates mainly to low-quality jobs improves metrics but not welfare
Short-run metrics vs. long-run mobility	Immediate conversion vs. career trajectories	Steering candidates to historically common roles can boost hires but limit future advancement
Privacy vs. group measurement	Data minimization vs. demographic auditing	Avoiding collection of sensitive attributes makes disparity detection and correction harder
Interpretability vs. optimization	Simple mechanisms vs. complex allocation algorithms	Transparent rules may be less efficient than opaque but finely tuned exposure policies

Table 7. Alternative baselines against which to compare marketplace outcomes when assessing fairness.

Baseline	Description	Implication for judgement
Unconstrained algorithm	Outcomes produced by current ranking and recommendation logic	Fairness is evaluated as a redistribution relative to existing product behavior
Pre-platform labor market	Outcomes in hiring channels that existed before the platform	Platform is assessed on whether it improves or worsens legacy disparities
Hypothetical nondiscriminatory benchmark	Counterfactual world with unbiased evaluation and accurate measurement	Requires normative assumptions; treats platform as approximating an ideal allocation
Legal or policy standard	Anti-discrimination law or internal policy targets	Fairness judged by compliance with protected-class constraints and impact tests

If certain candidates are rarely shown, the platform has little data about how they would perform if shown. If certain employers rarely engage with certain groups, the platform observes few positive labels for those groups with those employers, and a model may infer low fit, even if the lack of engagement reflects bias. This creates a partial observability problem. Fairness metrics computed on observed data may be misleading because they condition on selection that is itself discriminatory. For example, if an algorithm sends a protected group only to employers who are more receptive, then observed hiring rates for that group may look good, even though the group is being steered away from higher-paying or higher-status roles (15). Conversely, if a group is shown broadly, including to biased employers, observed conversion may be low, and a naive model may interpret this as evidence of lower quality.

Another mechanism is attention spillover. If employers see a more diverse slate, they may change their evaluation thresholds. In some cases, increased diversity in the candidate pool can lead employers to raise standards due to implicit comparisons, which can disadvantage all candidates but may have differential impact across groups. Alternatively, structured comparison can reduce reliance on stereotypes, improving fairness. These are behavioral effects that depend on interface and evaluation design. A two-sided theory should therefore consider not only the statistical properties of rankings but also how humans respond to them.

There are also cross-employer externalities. A platform might allocate more exposure of underrepresented candidates to employers that historically respond more positively, improving immediate conversion. However, this can concentrate those candidates in a subset of employers and reduce their access to a broader range of opportunities. It can also create a narrative that diversity is the responsibility of certain employers, while others remain insulated. Conversely, distributing exposure across many employers can normalize diverse hiring but can also subject candidates to more rejection. The fairness implications depend

Table 8. Complementary lenses for evaluating fairness as a property of the overall hiring system.

Evaluation lens	Methods	Strengths / limitations
Observational metrics	Stage-wise disparity statistics with selection-aware stratification	Low friction but confounded by historical allocations and unobserved counterfactuals
Randomized exploration	Controlled random exposure of candidates or jobs	Supports causal inference but can incur short-term efficiency and user-experience costs
Online experiments	A/B tests on ranking, interfaces, or information policies	Captures behavioral responses but subject to interference across users
Longitudinal monitoring	Dashboards, drift detection, trend analysis over time	Reveals slow changes and adaptation but may lag emerging harms
Qualitative & procedural signals	User reports, audits, appeal usage, explanation access	Surfaces issues outside metrics but harder to aggregate and quantify

on how one weighs short-term conversion against long-term normalization and access to high-quality opportunities.

Discrimination can be compounded across stages. If a protected group is slightly disadvantaged in ranking, then fewer members of that group are contacted (16). If those who are contacted face additional disadvantage in interview evaluation, then the final disparity can be larger than any single-stage disparity. This multiplicative effect means that interventions should consider the full pipeline. A platform might implement a fairness constraint in ranking that equalizes exposure, but if employer response remains biased, the constraint might increase rejection rates for the protected group, potentially increasing discouragement and churn. A theory of fairness interventions should anticipate such pipeline interactions and should avoid evaluating interventions solely at the stage where they are implemented.

The interaction between discrimination mechanisms and product objectives is particularly important. Platforms often optimize for metrics such as time to hire, retention, employer satisfaction, and engagement. These metrics can conflict with fairness. If biased employers are more likely to be satisfied when presented with candidates who match their biased preferences, then optimizing employer satisfaction can perpetuate discrimination. If candidate retention is higher when candidates are shown roles aligned with historical patterns, then optimizing retention can entrench occupational segregation. A fairness framework therefore requires a careful distinction between legitimate market efficiency and efficiency defined in terms of biased preferences. Without this distinction, a platform can rationalize discriminatory outcomes as efficiency improvements.

Finally, discrimination in hiring is subject to legal and normative constraints that affect intervention choice. Some jurisdictions restrict the use of protected attributes, which complicates debiasing because group membership information is often needed to detect and correct disparate impact. Platforms may rely on inferred group labels for auditing, but inference can introduce error and additional privacy concerns. Employers may request tools that help them diversify hiring, while also fearing legal risk (17). This can create a tension between proactive fairness and compliance. The theory must therefore treat legal feasibility, privacy, and institutional accountability as part of the mechanism landscape, not as external afterthoughts.

5. Intervention Design Space in Two-Sided Matching Markets

Interventions to reduce discrimination in job matching can be grouped by where they act in the marketplace pipeline: representation and measurement, ranking and exposure allocation, interaction and evaluation design, and governance and accountability. In two-sided marketplaces, interventions also differ in whether they require cooperation from employers, whether they change incentives, and whether they operate on individual decisions or on aggregate distributions. A theoretical foundation clarifies the assumptions under which each intervention can reduce discrimination, and the failure modes under which it can simply relocate disparities.

At the representation level, interventions aim to reduce measurement disparity. One approach is to improve feature robustness so that equivalent skills are recognized regardless of how they are expressed. This can involve domain-specific parsing of resumes, normalization of job titles, or structured skill extraction that is validated for differential error across demographic groups. Another approach is to decouple features from proxies for protected status by limiting certain signals, such as removing names or photos at early stages. Such anonymization can reduce direct discrimination but may not remove proxies such as school or geography, and it can reduce the ability of candidates to convey legitimate context. A theoretical issue is that removing information can change equilibrium behavior. If employers cannot observe certain cues, they may substitute other cues, potentially increasing reliance on noisy proxies. If candidates know that certain fields are hidden, they may invest less in them and more in other signals, altering the

Table 9. Illustrative feedback loops and adaptation pathways influencing long-run fairness in hiring platforms.

Loop element	Origin	Destination	Fairness risk
Algorithmic ranking	Current model and training data	Exposure of candidates and jobs	Systematically low ranking for a group reduces future positive labels for that group
Employer response	Employer preferences and filters	Observed contacts, interviews, hires	Biased response behavior becomes embedded in labels used to train future models
Candidate behavior	Application strategies, profile edits	Composition of applicant pools	Discouraged candidates churn, reinforcing underrepresentation and skewing data
Governance decisions	Platform policies, auditing, documentation	Model objectives and constraints	Static metrics or weak oversight allow gradual drift toward engagement-only optimization

distribution of features.

At the ranking and recommendation level, interventions allocate exposure under constraints. A common conceptual intervention is constrained ranking, where the platform imposes requirements on the composition of top results according to group membership or according to other fairness criteria (18). In a two-sided setting, the relevant constraint is often exposure rather than mere presence, because the probability of being seen declines sharply with rank. Thus, an intervention may ensure that expected views for qualified candidates are balanced across groups. The theoretical challenge is to define qualification in a way that is not itself biased, and to ensure that the constraint does not induce perverse incentives such as gaming. Another challenge is that exposure constraints can reduce the match quality for some employers if they perceive the slate as less tailored, which can reduce employer participation. A system-level analysis considers whether the intervention changes employer churn or the distribution of job postings.

Exposure-aware matching extends constrained ranking by considering both sides' preferences and constraints. Rather than balancing the top of a single ranking list, the platform can allocate candidate exposure across employers so that the overall distribution of opportunities is fair. This approach can incorporate congestion by steering candidates toward responsive employers while ensuring that high-quality opportunities are not systematically withheld. The theoretical benefit is that fairness is treated as a property of the entire exposure network rather than of isolated rankings. The risk is that steering can become a form of paternalism or can create separate tracks that resemble segregation, even if conversion improves. The framework therefore distinguishes between interventions that equalize immediate hiring probability and interventions that equalize access to desirable opportunities.

Calibration and threshold interventions operate on screening decisions. If a platform provides employers with scores, such as predicted job fit or predicted likelihood of interview success, then score calibration across groups can reduce differential interpretation. Employers may set thresholds, and if scores are systematically under- or over-estimated for certain groups due to label bias, then thresholds can create disparate rejection (19). Calibrating scores can mitigate this, but only if the calibration target corresponds to a legitimate outcome. If the target is employer interview selection, then calibration may encode employer bias. If the target is job performance, then calibration is constrained by limited observability and delayed outcomes. In hiring marketplaces, performance labels may arrive long after matching and may be observed only for those hired, creating selection bias. The theoretical implication is that calibration must be paired with careful label choice and possibly with causal inference methods, even if the intervention is described operationally as a simple recalibration.

Interface and interaction design interventions modify how employers and candidates make decisions. Structured evaluation, such as requiring employers to rate candidates on job-relevant criteria before seeing certain identity cues, can reduce reliance on stereotypes. Similarly, batching candidates into comparison sets can encourage more consistent evaluation than sequential browsing. For candidates, providing transparency about employer responsiveness and hiring practices can help them avoid biased employers and reduce wasted effort. However, transparency can also enable strategic avoidance by employers who anticipate scrutiny or can lead to herding toward certain employers, worsening congestion. The theoretical lens here is mechanism design: the platform changes the information and decision environment, which changes incentives and behavior. Fairness gains depend on whether the new mechanism makes discriminatory strategies less attractive or less feasible.

Incentive-based interventions aim to shift employer behavior

by altering costs and benefits. For example, platforms can create reputation signals for inclusive hiring practices, or they can highlight employers that consistently consider a broad range of candidates. They can also impose friction on discriminatory filtering, such as requiring justification for certain exclusionary criteria (20). These interventions do not force outcomes directly but can change the equilibrium by making biased behavior more costly. The challenge is to define signals and penalties without enabling manipulation. If employers can perform inclusiveness superficially to gain reputation without changing actual hiring, the intervention can become symbolic. A theoretical foundation emphasizes auditability and robustness: the platform must be able to detect meaningful changes rather than being misled by easily gamed metrics.

Auditing and monitoring interventions treat fairness as an ongoing property rather than a one-time constraint. The platform can monitor disparities in exposure, contact rates, and conversion, and it can test for patterns that suggest discrimination. In a two-sided marketplace, auditing can be done at multiple levels: across the platform as a whole, within an employer, within job categories, and across geographic regions. The theory must clarify what constitutes evidence of discrimination versus evidence of legitimate differences in applicant pools or requirements. Because of selection and omitted variables, audit metrics can produce false positives or false negatives. Therefore, auditing frameworks often incorporate uncertainty bounds and emphasize trend detection. In contexts where protected group labels are unavailable, the platform may rely on proxy inference, which introduces measurement error and ethical concerns. The theoretical role of auditing is to provide feedback for governance and intervention adjustment, not to produce definitive causal claims from observational data alone.

Governance interventions define who sets objectives and who has recourse. Platforms can establish internal review processes for new models, including fairness impact assessment, documentation of training data, and scenario testing for two-sided feedback effects. They can also provide channels for candidates to contest or correct profile representations and for employers to receive guidance on nondiscriminatory use of tools (21). Governance includes data stewardship: limiting retention of sensitive features, controlling third-party access, and ensuring that model updates do not silently change candidate exposure in ways that harm protected groups. In a theoretical frame, governance interventions are part of the mechanism because they shape which constraints are enforced and how the system evolves over time.

A distinctive two-sided intervention is to manage participation and supply-side constraints. Some disparities reflect unequal access to credentials or to geographic mobility. A platform can partner with training providers or can surface credential pathways, effectively expanding the feasible set of candidates for certain jobs. While this goes beyond traditional fairness constraints, it can reduce structural discrimination by changing the underlying distribution of qualifications. The fairness question then becomes whether the platform is acting as a labor market institution rather than a neutral intermediary, and whether such interventions are aligned with stakeholder expectations. A neutral theoretical stance treats this as a design choice with trade-offs: it can improve long-run equity but can also create conflicts of interest and requires careful governance.

Interventions can interact. For example, anonymization can reduce direct bias, but if combined with exposure constraints it may lead employers to feel that the slate is less informative, possibly increasing reliance on alternative screening. Conversely, structured evaluation can complement exposure constraints by ensuring that increased exposure translates into meaningful consideration rather than immediate dismissal. A two-sided theory therefore emphasizes composition: interventions should be analyzed as bundles, with attention to how they shift behavior across stages. It also emphasizes that interventions can shift discrimination from observable metrics to less observable ones, such as from platform messaging to off-platform recruitment. The design space must include safeguards against displacement (22).

6. Theoretical Analysis of Trade-Offs, Equilibria, and Welfare

In two-sided marketplaces, fairness interventions can change equilibria. Even if an intervention improves fairness metrics in a static snapshot, it can alter incentives and participation in ways that affect long-run outcomes. Theoretical analysis therefore considers not only immediate allocation effects but also how the marketplace adapts. This section develops conceptual arguments about trade-offs and equilibrium responses without relying on formal mathematics, focusing instead on mechanism-level reasoning.

A central trade-off is between exposure redistribution and perceived relevance. When a platform modifies rankings to increase exposure for underrepresented candidates, some employers may experience the presented slate as less aligned with their priors. If employers interpret this as reduced platform quality, they may reduce usage, apply more stringent filters, or shift to alternative recruitment channels. The resulting reduction in demand can offset fairness gains by shrinking the opportunity pool. This does not imply that fairness interventions should be avoided, but it suggests that interventions need to consider employer-side utility and should be designed to preserve match quality where possible. One theoretical approach is to impose fairness constraints only among candidates above a minimum relevance threshold, or to ensure that additional exposure is allocated among a pool that remains plausibly qualified for the role. However, relevance thresholds can themselves be biased if they rely on features affected by measurement disparity. The framework therefore treats thresholds as policy choices that require auditing and adjustment.

Another trade-off is between candidate-side parity and candidate-side welfare. Equalizing exposure does not necessarily equalize the quality of opportunities. A platform can achieve parity by showing underrepresented candidates more frequently to low-quality employers or to jobs with poor conditions (23). In such a case, exposure fairness is achieved, but welfare fairness is not. Conversely, focusing on welfare outcomes such as wages may be infeasible because the platform does not observe wages

reliably, and because wages depend on negotiation and firm policies. Theoretical analysis highlights the need for proxy measures of opportunity quality, such as employer responsiveness, job stability signals, or historical retention, while recognizing that these proxies can embed bias. A neutral stance is to treat welfare proxies as imperfect instruments that require ongoing validation.

Equilibrium effects can also arise through candidate adaptation. If candidates learn that the platform applies fairness constraints, they may change application strategies. Some candidates may broaden their search if they expect improved consideration, increasing competition. Others may reduce effort in profile polishing if they believe exposure is guaranteed, potentially reducing match quality. These responses depend on how transparent the intervention is and on how candidates perceive it. In hiring, perceptions of stigmatization can matter. If candidates believe that increased exposure is due to group-based adjustment rather than merit, they may fear being tokenized, and employers may also react with skepticism. A theoretical foundation therefore includes procedural considerations: interventions that are framed as improving measurement and expanding consideration may be less stigmatizing than interventions framed as pure rebalancing, even if the underlying mechanics are similar.

Employers can adapt as well. If a platform restricts discriminatory filtering, employers may shift discrimination to later stages that are harder to monitor, such as interview evaluation or offer negotiation. This is a displacement effect (24). An intervention that increases diverse slates can lead to more diverse interviews, but if interview processes remain biased, the final hiring disparity may persist. Moreover, increased exposure can impose additional emotional and time costs on candidates who face more rejection, especially if rejection is due to bias. A system-level fairness analysis therefore considers not only hiring rates but also the distribution of search costs. Fairness might include minimizing wasted applications for disadvantaged candidates by steering them toward employers with demonstrated fairness, but such steering can resemble segregation if it isolates candidates into a subset of employers. The theoretical tension is between protecting candidates from biased interactions and ensuring broad access to all opportunities.

Two-sided systems also have market power considerations. Large platforms can shape norms by influencing what employers see as standard hiring practice. A platform that enforces structured evaluation can raise the baseline of procedural fairness, potentially reducing discrimination beyond the platform. Conversely, a platform that optimizes purely for engagement can normalize discriminatory preference satisfaction. The equilibrium impact depends on the platform's share of the hiring channel and on how portable employer and candidate reputations are across platforms. A theoretical framework can incorporate this by considering multi-homing, where employers and candidates use multiple channels. If multi-homing is common, platform-level fairness interventions may have limited impact because actors can bypass them. If multi-homing is costly, platform interventions can have stronger institutional effects but also raise concerns about unilateral control.

Welfare analysis in hiring must consider that both sides can be harmed by poor matches. Employers can incur costs from turnover and poor performance, and candidates can incur costs from job instability and mismatch (25). Fairness interventions that significantly reduce match quality can harm both sides. However, the perception of reduced match quality may be driven by biased beliefs rather than by actual productivity differences. This is important because an intervention might be welfare-improving in reality but welfare-reducing in perception, leading to reduced participation. Addressing this requires not only algorithmic changes but also communication and trust-building mechanisms. For example, platforms may provide employers with evidence-based explanations of candidate fit, focusing on job-relevant skills rather than on proxies. The theoretical point is that information design can shift beliefs and therefore equilibria.

Another aspect is long-run human capital. If underrepresented candidates receive more opportunities, they can accumulate experience, which improves future qualifications. This dynamic can create virtuous cycles that reduce structural inequality. Conversely, if a platform's algorithm reinforces occupational segregation, it can create path dependence, where candidates from certain groups are repeatedly steered into lower-growth roles, limiting future mobility. A fairness intervention can therefore be justified not only by immediate parity but also by long-run mobility. The challenge is that long-run outcomes are hard to observe and attribute, and interventions must avoid paternalism. A neutral theoretical stance is to recognize mobility as a relevant but uncertain objective, and to design interventions that expand option sets rather than forcing choices.

In a two-sided marketplace, stability considerations can matter (26). If a platform matches candidates to jobs in a way that consistently violates employer preferences, employers may reject matches, leading to churn and wasted effort. If it violates candidate preferences, candidates may decline offers or quit. Fairness interventions that change the distribution of offers must therefore consider acceptances and retention. An intervention that increases offers to underrepresented groups but leads to lower acceptance because the jobs are less desirable does not achieve welfare equity. The platform can respond by incorporating candidate preferences into the matching objective, but preference measurement can be biased if candidates have adapted preferences under constraint. For instance, candidates may indicate lower wage expectations due to historical disadvantage. Using such preferences naively can perpetuate inequity. The theoretical implication is that preference modeling must be treated carefully, potentially distinguishing between expressed preferences and constrained preferences.

Trade-offs also exist between privacy and fairness. Fairness auditing often requires demographic group labels, but collecting such labels can raise privacy concerns and can be legally restricted. Using inferred labels introduces error and can itself be discriminatory if inference quality differs across groups. A platform may choose to implement fairness constraints without explicit group labels by using individual consistency or by focusing on representation robustness. These approaches can improve certain aspects of fairness but may miss group-level

disparities. The theoretical framework thus treats privacy constraints as part of the feasible set and emphasizes that fairness under privacy constraints may require different notions, such as worst-case disparity bounds under uncertain group membership.

Finally, there is a trade-off between interpretability and optimization (27). Complex two-sided optimization, such as exposure-aware allocation across many employers and candidates, can be difficult to explain. Yet explanation and contestability are important for legitimacy. A platform may need to use simpler mechanisms that are easier to communicate, even if they are less optimal in a narrow sense. The theoretical stance here is pragmatic: a fairness intervention that stakeholders can understand and monitor may be more sustainable than a highly optimized intervention that lacks transparency, even if the latter looks better in offline metrics.

7. Evaluation and Validation as System Properties

Evaluating fairness in a two-sided hiring marketplace requires moving beyond static metrics computed on historical data. Because of selection, feedback, and strategic behavior, offline evaluation can misrepresent real-world impact. A theoretical evaluation framework therefore treats fairness as a system property that must be validated through multiple lenses: observational metrics with selection awareness, counterfactual reasoning, online experimentation with safeguards, and longitudinal monitoring.

At minimum, a platform can measure disparities at each pipeline stage: exposure, contact, interview, offer, and acceptance. However, stage-wise parity is not sufficient because it can hide steering and quality differences. Evaluation should also consider the distribution of job quality proxies across groups, such as employer responsiveness, role seniority signals, and location desirability as expressed by candidates. Because these proxies can be biased, they should be validated against external benchmarks where possible, such as aggregate wage data or retention indicators, while recognizing that perfect validation is rarely possible. A system perspective encourages triangulation rather than reliance on a single metric.

Selection bias complicates interpretation. If a candidate is rarely shown, the platform cannot observe what would have happened if they were shown. To address this, evaluation can incorporate randomized exposure or exploration, where a small fraction of ranking decisions include controlled randomization to gather unbiased data about candidate-employer interactions (28). Exploration introduces trade-offs because it can reduce immediate match quality and can impose costs on users. A fairness-aware exploration strategy can prioritize uncertainty reduction for underrepresented groups or for regions of the feature space with sparse data, while limiting disruption. The theoretical point is that fairness evaluation is inseparable from data collection policy, and that platforms that never explore cannot reliably assess whether disparities are due to true fit differences or due to algorithmic suppression.

Counterfactual reasoning is important because observed outcomes reflect both algorithmic choices and human behavior. For example, if a protected group has lower interview rates, is it because the platform showed them less, because employers contacted them less, because candidates declined interviews more, or because the jobs they were shown were less compatible? A system evaluation seeks to decompose disparities into components attributable to platform allocation and components attributable to participant decisions. This decomposition is not purely statistical; it requires assumptions about what would have happened under alternative allocations. Theoretical rigor comes from making those assumptions explicit and testing sensitivity. A platform can conduct controlled experiments that alter one mechanism at a time, such as changing ranking while keeping interface constant, or changing information revelation while keeping ranking constant, to identify which mechanisms drive changes.

Online experimentation raises ethical concerns because it can affect real job opportunities. A fairness evaluation framework therefore includes constraints on experimentation, such as limiting the magnitude of exposure changes, ensuring that no group is systematically harmed during tests, and providing opt-out where feasible. In a marketplace, experiments can have interference: changing exposure for some candidates changes the competition faced by others. This makes naive A/B testing unreliable because the treatment effect is not isolated. Evaluation must therefore consider network effects and congestion. One approach is to randomize at the employer level or at the job-posting level, treating each employer's candidate pool as a unit, while monitoring spillovers (29). Another is to use phased rollouts across regions or job categories, though this can also create fairness concerns if rollout order correlates with demographics. The theoretical contribution is to recognize interference as a fundamental property of two-sided markets and to design evaluation accordingly.

Longitudinal monitoring is essential because fairness can drift. Model updates, seasonal hiring changes, and macroeconomic shifts can change applicant pools and employer behavior. A platform can maintain dashboards that track disparities over time, including distribution shifts in features and in outcomes. Monitoring should include alerting for sudden changes and for slow trends. It should also include checks for strategic adaptation, such as increased use of particular filters or changes in job description language that might indicate attempts to circumvent constraints. The theoretical idea is that fairness is not a one-time property that can be certified; it is an ongoing governance commitment that requires measurement and response.

Evaluation should also incorporate qualitative signals. Candidates may report experiences of bias, and employers may report confusion about platform behavior. While qualitative reports are not definitive, they can reveal issues that metrics miss, such as stigmatizing messaging or interface cues that trigger stereotype activation. A system framework treats these reports as part of the evidence base, complementing quantitative metrics. It also emphasizes that fairness interventions can change perceptions even when outcomes change little, affecting participation. Participation is itself a fairness issue because disadvantaged groups may be less likely to engage if they perceive the platform as biased or opaque.

Robustness evaluation examines whether fairness holds under plausible deviations (30). For instance, if an intervention assumes that employer response rates are stable, what happens if employers increase filtering? If an intervention assumes that candidates do not change application volume, what happens if application behavior shifts? Robustness can be tested through stress scenarios and simulations that model adaptation. Simulations in hiring are necessarily simplified, but they can still be useful for identifying qualitative failure modes, such as exposure constraints that lead to severe congestion for certain employers, or anonymization that leads employers to rely on more biased proxies. The theoretical stance is cautious: simulations are not proof, but they provide a structured way to anticipate second-order effects that offline metrics miss.

A further evaluation dimension is procedural fairness and contestability. Platforms can measure how often candidates correct profile errors, how quickly corrections propagate to ranking, and whether correction rates differ across groups. They can measure whether explanations are accessed and whether they correlate with improved outcomes. They can also evaluate whether appeals processes are used disproportionately by certain groups, which might indicate unequal burden. These procedural metrics connect fairness to user experience and to institutional legitimacy.

Finally, evaluation must consider the risk of metric gaming. If a platform publicly commits to a particular fairness metric, actors may adapt to improve the metric without improving substantive fairness. Employers might adjust behavior in ways that satisfy monitored parity while maintaining discrimination in unmonitored dimensions. Candidates might adapt profiles to fit monitored categories. A robust evaluation framework therefore uses multiple metrics, monitors for inconsistencies, and periodically revises metrics. The theoretical point is that fairness metrics are part of the mechanism: they influence behavior when they become targets. Therefore, fairness governance includes the ability to evolve measurement while maintaining accountability (31).

8. Conclusion

Algorithmic fairness in two-sided hiring marketplaces differs from fairness in isolated prediction tasks because the platform mediates interaction between strategic actors under attention constraints and feedback. Discrimination can arise from measurement disparities in how skills are represented, from employer preferences and biases that are amplified by engagement-optimized objectives, from process design choices that shape who enters and progresses through the funnel, and from dynamic feedback that reinforces historical patterns. In this setting, fairness cannot be fully captured by static parity metrics computed on observed outcomes, because observed outcomes are shaped by selection and by the algorithm itself. A theoretical foundation therefore benefits from treating fairness as a system property defined over exposure, engagement transitions, match outcomes, and welfare proxies, while remaining explicit about baselines and normative commitments.

Interventions in two-sided hiring platforms operate across multiple layers. Representation interventions aim to reduce differential measurement error and to manage identity cues. Ranking and exposure interventions allocate scarce attention under constraints, with the risk of shifting disparities if downstream employer behavior remains biased. Interaction design interventions modify human decision environments to reduce reliance on stereotypes and to support structured evaluation. Incentive and governance interventions aim to make discriminatory behavior more costly, to improve accountability, and to ensure that fairness objectives persist through model updates and market shifts. In each case, theoretical analysis highlights potential failure modes that are specific to marketplaces, including displacement of discrimination across stages, congestion and interference effects, participation changes, and metric gaming.

Evaluation must align with this system view. Credible assessment requires selection-aware measurement, data collection strategies that reduce confounding, online experimentation designed for interference, and longitudinal monitoring that detects drift and adaptation. Because fairness is intertwined with privacy, legal constraints, and institutional legitimacy, governance mechanisms such as documentation, contestability, and stakeholder oversight are not optional add-ons but part of the intervention space. A neutral position recognizes that no single metric or mechanism resolves all normative disagreements, and that trade-offs are inherent. The practical implication is that fairness in hiring marketplaces is best approached as an ongoing design and governance problem that integrates technical constraints with careful attention to incentives, information flows, and the long-run dynamics of opportunity (32).

References

[1] B. Le, D. Spina, F. Scholer, and H. Chia, *A Crowdsourcing Methodology to Measure Algorithmic Bias in Black-Box Systems: A Case Study with COVID-Related Searches*, pp. 43–55. Germany: Springer International Publishing, 6 2022.

[2] N. Rzepka, L. Fernsel, H.-G. Müller, K. Simbeck, and N. Pinkwart, "Unbias me! mitigating algorithmic bias for less-studied demographic groups in the context of language learning technology," *Zenodo (CERN European Organization for Nuclear Research)*, 6 2023.

[3] A. Muralidharan, J. Savulescu, and G. O. Schaefer, "Public justification and normatively meaningful bias: Against imposing egalitarian accounts of algorithmic bias.," *Bioethics*, vol. 40, pp. 73–84, 11 2025.

[4] N. V. Patel, "An end-to-end, stage-wise life-cycle framework for mitigating algorithmic bias in two-sided marketplaces," *Power System Technology*, vol. 49, no. 3, pp. 434–457, 2025.

[5] A. Vickers, "Do not treat bill gates for prostate cancer! algorithmic bias and causality in medical prediction.," *BJU international*, vol. 131, pp. 263–264, 1 2023.

[6] K. D. Gray, H. L. Subramaniam, and E. S. Huang, "Effects of racial bias in pulse oximetry on children and how to address algorithmic bias in clinical medicine.," *JAMA pediatrics*, vol. 177, pp. 459–459, 5 2023.

[7] D. Spennemann, "The layered injection model of algorithmic bias as a conceptual framework to understand biases impacting the output of text-to-image models," 1 2025.

[8] M. Miron, S. Tolan, E. Gómez, and C. Castillo, "Evaluating causes of algorithmic bias in juvenile criminal recidivism," *Artificial Intelligence and Law*, vol. 29, pp. 111–147, 6 2020.

[9] P. Nanda, V. Kumar, and K. U. Pulmones, "Navigating algorithmic bias and data ethics in the gig economy: Balancing transparency and privacy," *EDPACS*, pp. 1–12, 10 2025.

[10] R. Isley, "Algorithmic bias and its implications: How to maintain ethics through ai governance," *N.Y.U. American Public Policy Review*, vol. 2, 10 2022.

[11] V. R. Lee, V. Delaney, and P. Sarin, "Eliciting high school students' conceptions and intuitions about algorithmic bias," in *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2*, pp. 35–36, ACM, 8 2022.

[12] S. Qureshi and B. Oladokun, "Human freedom from algorithmic bias: What is the role of accountability in addressing health disparities?," *Medical Research Archives*, vol. 12, no. 8, 2024.

[13] T. Bär, *Algorithmic Bias: Verzerrungen durch Algorithmen verstehen und verhindern*. Springer Berlin Heidelberg, 2022.

[14] H. Habib and R. Nithyanand, "Youtube recommendations reinforce negative emotions: Auditing algorithmic bias with emotionally-agentic sock puppets," 1 2025.

[15] Y. J. Park, "Tprc 2024: Algorithmic bias, marketplaces, and diversity regulation," 1 2024.

[16] D. Aobdia, H. Ma, and S. Zhang, "Auditing effects on employment hiring: Evidence from the new york city algorithmic bias audit law," *SSRN Electronic Journal*, 2025.

[17] M. Kasy, "Algorithmic bias and racial inequality: a critical review," *Oxford Review of Economic Policy*, vol. 40, pp. 530–546, 11 2024.

[18] R. Garg, R. A. Kalra, and N. Rangaswamy, "Unveiling algorithmic bias: Gender disparities in gaming and dating platforms in india," *Indian Journal of Gender Studies*, vol. 32, pp. 407–432, 10 2025.

[19] A. Bellogin, L. Boratto, S. Kleanthous, E. Lex, F. M. Malloci, and M. Marras, "Report on the 5th international workshop on algorithmic bias in search and recommendation (bias 2024) at sigir 2024," *ACM SIGIR Forum*, vol. 58, no. 2, pp. 1–6, 2024.

[20] J. C. Flowers, "Rethinking algorithmic bias through phenomenology and pragmatism."

[21] S. P. Sleeman and B. Rademan, "Freedom from social echo chambers: Policy implications of an algorithmic bias," *SSRN Electronic Journal*, 2017.

[22] J. Xiao, Z. Li, X. Xie, E. Getzen, C. Fang, Q. Long, and W. J. Su, "On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization," *Journal of the American Statistical Association*, vol. 120, pp. 2154–2164, 10 2025.

[23] A. F. Zambrano, J. Zhang, and R. S. Baker, "Investigating algorithmic bias on bayesian knowledge tracing and carelessness detectors," in *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pp. 349–359, ACM, 3 2024.

[24] D. A. Adler, C. A. Stamatis, J. Meyerhoff, D. C. Mohr, F. Wang, G. J. Aranovich, S. Sen, and T. Choudhury, "Measuring algorithmic bias to analyze the reliability of ai tools that predict depression risk using smartphone sensed-behavioral data.," 4 2024.

[25] R. Holderegger and L. F. de Almeida Duarte, "Artificial intelligence and socioeconomic inequalities: Algorithmic bias, labor market impacts, and governance agendas for emerging economies," *ARACÊ*, vol. 7, pp. e9346–, 10 2025.

[26] G. Adomavicius and M. Yang, "Integrating behavioral, economic, and technical insights to address algorithmic bias: Challenges and opportunities for is research," *SSRN Electronic Journal*, 2019.

[27] P. E. Bandeira, R. Limongi, and S. F. Rohden, "Advancing the academic discourse on algorithmic bias: Unpacking conceptual conflicts and mitigation," *Academy of Management Proceedings*, vol. 2025, no. 1, 2025.

[28] W. Sun, O. Nasraoui, and P. Shafto, "Kdir - iterated algorithmic bias in the interactive machine learning process of information filtering," in *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pp. 110–118, SCITEPRESS - Science and Technology Publications, 2018.

[29] Y. Gao, C. Lin, N. TA, H. Fu, and K. Li, "Trapped before clicking enter digital inequality and search engine autocomplete algorithmic bias," 1 2023.

[30] C. Chuan, R. Sun, S. Tian, and W. S. Tsai, "Explainable artificial intelligence (xai) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives," 1 2023.

[31] P. M. Heider, "Algorithmic bias in de-identification tools," in *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pp. 712–713, IEEE, 6 2023.

[32] E. Yalcin, *The Unfairness of Collaborative Filtering Algorithms' Bias Towards Blockbuster Items*, pp. 233–246. Springer International Publishing, 1 2023.