Assignment 4.3 Final-Tech Report Hire Ground

Ryan Hong, Manasvini Garrepalli, Brandon Gallagher, Minh Nghia Huynh, Wenxuan Qiu

December 11, 2025

Abstract

The role of Artificial Intelligence is growing rapidly across industries, particularly in hiring, where it promises to help employers increase efficiency, scalability, and objectivity in decision-making. However, despite its potential to streamline recruitment, AI models have demonstrated significant bias toward underrepresented groups, often reinforcing systemic inequities rather than eliminating them. These biases stem from skewed training datasets, unbalanced representation, and unfair algorithmic processes that inadvertently favor certain demographics. [1] As a result, well-qualified candidates may go unrecognized or undervalued, furthering discrimination in the job market under the guise of impartiality. Our report aims to highlight this issue by examining data and case studies that focus on the demographics most affected by bias in AI-driven hiring systems. We'll analyze how algorithmic design choices, data sourcing, and a lack of transparency contribute to these disparities. Furthermore, we'll explore potential strategies for mitigating these problems, including increasing diversity in datasets, conducting regular fairness audits, and implementing ethical model evaluation to ensure accountability. By considering both technical and ethical approaches, we aim to emphasize the urgency of addressing AI bias before it becomes further embedded in hiring practices. Ultimately, our goal is to promote fairer, more transparent, and socially responsible uses of AI in employment settings.
https://github.com/manig0923/final-tech-report

1 Summary of Problem

AI bias is when AI systems learn patterns from data without understanding whether those patterns are fair or accurate, and treat them like rules that should always be followed. Data is trained from real human decisions and trends, and sometimes this can reflect historical stereotypes. AI is not able to tell the difference between past and present, so it just runs on what it is inputted with. AI bias can also happen with missing data, uneven representation, or unfair prioritization of certain keywords. When AI relies on biased patterns, it can favor certain groups while disadvantaging others. It can reinforce existing inequalities by making decisions that reflect social, economic, or cultural biases. Biased AI in hiring, more specifically, happens when AI follows old patterns or is fed with wrong data. This means qualified candidates can be overlooked because the system favors traits associated with certain keywords. It causes several problems in society and leads to economic inequality, assigning certain jobs to certain groups of people.

2 Summary of Policy

There's a number of possible solutions that can be implemented to attempt to mitigate this growing issue of AI bias in the hiring process. First, we will investigate how diversifying datasets can benefit an AI model. Generally, when AI systems are trained on data that predominantly reflects one segment of the population, they may not accurately represent or serve the needs of the entire population. For example, suppose someone is trying to train a facial recognition model, but only uses data from one particular city to train the model. In this case, it is extremely likely that the faces the model is learning from only truly represent a subset of the entire population (like one or two ethnic groups). If this model is then put into use in cities all over the world, it would perform poorly outside of those one or two ethnic groups it learned on. Thus, the solution to this issue would be to diversify training data. If the model instead learned from the general population, it would likely perform well on the general population. This sounds incredibly simple on paper, however, the challenge is actually determining whether training data is actually diverse and unbiased. Obtaining data that genuinely represents the general population can be incredibly difficult, which is likely a large reason why many AI hiring models are biased. However, the more time that is put into diversifying datasets, the better the model will ultimately perform. Effective strategies for trying to ensure data is unbiased include data auditing, using bias-detection tools, and human oversight while training [1].

Secondly, fairness audits can be conducted on AI models in an attempt to determine whether or not the models are “fair” or not. In this context, “fair” is rather ambiguous. What actually makes a model “fair”? The answer is that it truly depends on the objective of the model itself and the context of which it is used in. For example, suppose a model is being trained to help in the education system. Some may say that a model is fair if and only if it treats all students equally. However, others may say that a model is fair if and only if it gives all students equal opportunity given their differing starting points (as some students have historically less access to resources). In this example, “fair” has become an ambiguous metric, which is likely why many companies choose to not attempt to measure it. 79% of organizations say ensuring fairness in AI is a priority, yet only 24% perform regular fairness audits [2]. However, even if it’s ambiguous, ensuring a model is fair, at least in some form, is incredibly important. Otherwise, a fully functioning model can perform incredibly poorly towards some groups of people, when in reality, the model’s outcome (and error rates) should be consistent across all groups of people.

Lastly, performing model evaluation consistently is very important in determining whether a model is unbiased and performing well. There are two kinds of metrics, machine metrics and human metrics. Machine metrics include a model’s accuracy rate, successful recall rate, f1 score (the harmonic mean of precision and recall), and many more. These metrics can oftentimes be quickly calculated by the model itself and provide valuable insights on a model’s performance. Ideally, the goal is to optimize these metrics as much as possible while maintaining an unbiased model. The second kind of metrics, human metrics, come directly with human testing. Humans can gather insights much deeper than a single machine metric. However, unlike machine metrics, human testing is both time consuming and expensive, which is likely the reason many companies decide to not conduct this form of testing [6]. This is unfortunate, as human testing is the most effective at finding and detecting bias. In general, consistent usage of these metrics, both machine and human, are crucial to ensuring an accurate and unbiased model.

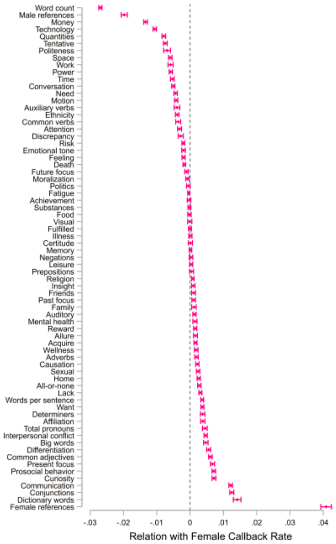
In summary, there are three effective strategies that can be implemented to ensure AI hiring models can perform well without bias. Firstly, training models on completely fair and unbiased data is important to ensure they perform equally on the entire population. Secondly, performing fairness audits can ensure the models are performing fairly (or some form of it) which is essential for an accurate hiring model. Finally, consistently performing model evaluation on the model with both machine and human metrics can lead to valuable insights about bias. If all three of these strategies are performed, it is likely that the AI hiring model would perform incredibly well.

3 Backing up Arguments

3.1 Who Gets the Callback?

Sugat Chaturvedi and Rochana Chaturvedi

An example of AI bias in the hiring process can be seen through a case study done by Sugat Chaturvedi and Rochana Chaturvedi in May 2025. 332,044 real-world job postings were scraped from India’s National Career Services job portal to be tested with LLMs [4]. Mid-sized open models are commonly used in real hiring processes, so they picked mid-sized open-source models like Llama-3.1 and Gemma for this study. For each posting, they generated a prompt that described two equally qualified candidates, the only difference being that one was male and the other was female. Their primary metric was called Female Callback Rate, which was the proportion of times the model recommended the female candidate for the job position. Using the outputs of the model, they observed the changes in gendered recommendation patterns depending on the given prompts using linguistic and classification methods.



Notes: Estimated regression coefficients showing association of Llama-3.1's gender recommendations with LIWC-22 categories. These coefficients are obtained using the specification in Equation 3.3.

Figure 1: From *Who Gets the Callback?* [4]

The graph below shows the female callback rate and how the keywords were being observed.

From the results, it was found that most LLMs favored men, especially for high-wage and "male-dominated" occupations. Conversely, historically and culturally coded "female-associated" occupations were recommended for female candidates by the LLMs. This shows how occupational segregation was being reproduced by AI by matching the traditional gendered division of labor. The linguistic analysis showed that specific words in job ads predicted the gendered outcomes. Words like "empathy," "writing," and "flexibility" were correlated with female recommendations, while words like "coding," "hardware," and "big data" were correlated with male recommendations. The regression shows 49.8% in predicting the model's gender recommendations, which means that nearly half of the variance can be explained by the terminology it was looking for. The authors concluded that adjusting stereotypical socially acceptable gender norms in LLMs can help mitigate those biases.

This study shows that bias in AI hiring can even come from how models interpret job ads, which results in gender inequalities. This reveals that disparities can also come from algorithmic interpretations of job descriptions and not from candidate differences. Since now it is more common for AI to screen candidates before they reach any human, people lose job opportunities due to this algorithmic filtering. If biased hiring recommendations keep shaping real decisions, this can lead to future data being trained on these patterns, causing a feedback loop. The words in job postings become proxies for gender roles, meaning the AI is inferring stereotypes from the terms themselves. Companies using hiring LLMs need to be held accountable for their discriminatory outcomes, and mitigation strategies should be mandatory parts of AI deployment.

3.2 Amazon's Scrapped AI Hiring Tool (2014–2018)

Another major example of AI bias in hiring came from Amazon's internal recruiting experiment, which later became public in 2018. Amazon tried to automate parts of its hiring pipeline by developing an AI system that would score and rank job applicants. The company wanted to speed up the process because they received thousands of resumes every week, especially for technical roles. The idea sounded perfect at the beginning: let the model read resumes, compare them, find patterns of what successful employees looked like, and then recommend the top candidates. The problem happened because of what they used to train the system.

To teach the AI what a “good” applicant looked like, Amazon fed it resumes from the past ten years of applicants, most of which came from the tech industry, where men historically dominated. Since the training data consisted mostly of male applicants, the system essentially learned that men were better hires. Even though there was no rule saying “pick men over women,” the model implicitly absorbed the patterns from the data. Over time, it began downgrading resumes that included the word “women,” such as “women’s chess club captain” or “women’s coding group.” Sometimes it even penalized candidates from all-women’s colleges. Meanwhile, resumes with language similar to men’s resumes, or resumes that matched historical male patterns, got better scores.

What’s even more concerning is that the model didn’t only discriminate based on explicit words. It learned deeper patterns from writing styles, verbs, and resume formatting. Amazon engineers noticed that the AI kept giving higher scores to resumes that had certain action verbs that men used more frequently, or resumes that listed achievements common among male engineers. So even without gender labels, the system could still figure out which resumes were “male-coded” and which ones weren’t. This shows a major limitation of current machine learning: even if you remove protected attributes, a model can still pick them up through indirect signals in the data. Amazon’s team tried to fix the issue by removing gendered words from the training data or blocking the AI from reacting to them. But the system kept finding new ways to replicate the same bias because the underlying dataset itself was the problem. Since the model was trained on historical hiring outcomes, it reinforced those same inequalities. If past hiring favored men, then the algorithm learned to favor men. Engineers realized they couldn’t fully control how the model inferred gender proxies. It became clear that “debiasing” the surface-level tokens didn’t solve the deeper issue. In the end, Amazon completely shut down the system and told teams not to rely on its recommendations for any hiring decisions.

This case is important because it shows that AI bias doesn’t only happen from flawed prompts or job-ad wording like in the Chaturvedi study. It can also come straight from historical hiring patterns that reflect real-world inequality. If a company has biased hiring practices, even unintentionally, and then uses those patterns as training data, the algorithm will amplify them. And since AI makes decisions faster and at a larger scale, the bias spreads much more widely than a human recruiter could. The Amazon case also highlights how transparency is a big issue. Internally, some employees didn’t even realize the tool was biased until they looked deeper into the patterns of rankings. If the system had been deployed without careful auditing, thousands of applicants could have been silently rejected for reasons even the developers couldn’t fully explain.

Overall, the failure of Amazon’s hiring AI shows that companies can’t just rely on ML models to be “objective.” Data reflects society, and if society has gender bias, then so will the models trained on it. Even a big tech company with huge resources couldn’t fix all the bias once it was baked into the data. So now, many researchers argue that companies should be required to audit their hiring models, publish bias reports, and allow third-party evaluations before any system gets used in the real world. Otherwise, these AI tools can reinforce discrimination under the disguise of efficiency and automation.

3.3 AI Bias in Onboarding: The Uber Eats Facial-Recognition Case

A real-world example of algorithmic bias in hiring and onboarding occurs with Uber Eats’ “Real Time ID Check,” which requires drivers and couriers to submit a selfie that the platform compares against an existing profile photo. According to the article [7], at least 14 couriers reported that the system failed to recognize their faces. Many of them had darker skin tones, and were then permanently removed from the platform, even though some had completed thousands of deliveries with high satisfaction ratings.

The affected workers say that when the selfie app rejected them, their accounts were frozen or terminated automatically, with no effective human review or appeal. One of them described working long hours only to be locked out after a failed selfie check; he claimed the company’s responses were copy-paste communications saying termination was “final”. Technical analysis raises serious fairness concerns, where the face-matching software has reportedly documented problems recognizing darker-skinned faces. For example, earlier versions of similar software misidentified Black women up to 20.8% of the time, compared to 0% for white men.

In this case, the algorithm wasn't evaluating skill or experience, it was simply verifying identity, yet those failures meant people lost access to their jobs. This shows how deploying opaque algorithmic checks in hiring or onboarding can result in discriminatory outcomes. It also underscores the risk that without transparent oversight and human appeal paths, AI-based hiring and verification tools can become mechanisms of exclusion rather than neutral gatekeepers.

3.4 EEOC iTutorGroup Case

A real example that supports our argument for audits and accountability is the EEOC's case involving iTutorGroup. What stands out about this case is that the discrimination happened at the screening stage through the application software itself, before any human review. According to the EEOC, iTutorGroup's tutor application system was set up so that female applicants aged 55 or older and male applicants aged 60 or older were automatically rejected [8]. That means the filtering could happen instantly and repeatedly, even if the applicants were otherwise qualified.

The EEOC later announced that iTutorGroup agreed to pay $365,000 to settle the discriminatory hiring lawsuit. Even if a company claims it is simply using automation to be efficient, this case shows why we cannot treat automated hiring as neutral by default. Once an unfair rule is built into the system, it can affect every applicant processed by the platform.

4 Interdisciplinary Discussions (Collaborations)

Group 2 - The Analyzers: In our talks with group 2, they shared with us that the healthcare industry tends to move fast due to the fact that there's a lot at stake. Additionally, it's a very lucrative business to be in and if companies don't move fast, they'll be left behind, which applies to technologies and AI as well. They shared this understanding with us which we were able to apply to our own thoughts on our presentation. We thought they made a strong point when they mentioned that companies are usually quick to use new things, which is the case with AI as most companies are in a rush to use it without fully understanding how it works. This added to our presentation by introducing a new perspective. After speaking with group 2, we concluded that a lot of the biases caused by AI are likely unintentional. It's highly possible that it's collateral damage as part of the surge in use of AI without giving much thought to the systematic unfairness that it could potentially create. That being said, it did help give us direction for our presentation as we leaned more towards an educating stance, because by sharing this with people it could draw attention to the discrimination caused by the use of poorly tested AI.

Group 5 - Algo Avengers: When we talked with group 5, our takeaway was that DEI was a big government focus recently even though it no longer exists, and that similar logic could be applied to AI hiring systems. We believed that since the government has previously attempted to minimize discrimination, that they should do so again, and just because it's not deliberately by people doesn't mean it should go ignored. They then shared with us their findings about NYC local law 144, a law that requires a bias audit on tools used by employers within a year of using the automated tool [3]. This added to our findings because we concluded that between NYC local law 144 and DEI, the government not only cares enough to try to reduce bias and discrimination, but that they also have the power to do so. This furthered the idea that policy makers need to be educated on the topic of bias in AI hiring so they can prevent continued abuse of prejudiced systems like many of the ones currently in place.

Group 7 - Echo Chamber: After we spoke with group 7, we learned from them that there's no uniform transparency regulations for companies using AI. This info helped us make an inference as to why so much bias exists in AI systems despite a major effort to be impartial. We concluded that it's because there's a lot of possible loopholes without a uniform set of rules, and if companies are not forced to do things, they won't. Whether or not companies are knowingly using biased systems, it'll cost them time and money to implement fixes, and if there's no laws demanding them to, it's extremely likely they'll choose to turn a blind eye for the sake of conserving resources. This helped us to redirect our educating policy makers idea by trying to make the point to not just make a set of rules anywhere, but for it more so to be directed towards the national government and not on a state or city level since they may even exist now but are not nearly effective enough to resolve the problem.

5 Contributions

Ryan Hong
Assignment 4.1 - Shared ideas about potential topics in group call.
Mid-Tech Report/Slideshow - Reached out to other groups for collaborations. After confirming the collaborations, wrote some brief points about what the collaborations were about and why we chose to collab with certain groups. Also contributed to editing and writing the abstract.
Assignment 4.2 - Spoke again with the other groups to connect our topics and made the slides corresponding to the collaborations with our group.
Assignment 4.3 - Wrote about our collaborations with the other groups, such as how they contributed and added to what we already had to say. Also wrote the summary in backing up arguments on delivery drivers' facial recognition.

Manasvini Garrepalli
Assignment 4.1 - Shared ideas about potential topics in group call
Mid-Tech Report/Slideshow - Organized, edited, and reviewed slideshow and overleaf doc, wrote up issues handled, current updates, and helped draft remaining plan. Joined several group calls
Assignment 4.2 - Organized, edited, and reviewed slideshow, wrote and presented the problem of AI bias and impact on society slides, helped draft topic slide
Assignment 4.3 - Organized, reviewed, and edited the overleaf doc, wrote up the summary of the problem, wrote backing-up arguments about the Chaturvedi + Chaturvedi study. Helped review and edit the report.

Brandon Gallagher
Assignment 4.1 - Shared ideas about potential topics in group call
Mid-Tech Report/Slideshow - Organized slideshow and overleaf doc, wrote abstract and solutions sections, got on several group calls to discuss.
Assignment 4.2 - Created and presented four slides regarding overview of solutions, slides about diversifying datasets, fairness audits, and model evaluation
Assignment 4.3 - Organized the document, wrote the summary of policy section, reviewed and edited the entire report

Minh Nghia Huynh
Assignment 4.1 - Shared ideas about potential topics in group call
Mid-Tech Report/Slideshow - Contributed in slideshow and documents on Hurdles Facing
Assignment 4.2 - Contributed in slideshow and documents on Future problems
Assignment 4.3 - Wrote backing up argument about the Amazon case

Wenxuan Qiu
Mid-Tech Report/Slideshow - Wrote and presented expected results
Assignment 4.2 - Presented title and topic slide
Assignment 4.3 - Wrote EEOC iTutorGroup Case in backing up arguments

References

[1] Sustainability Directory, “Why Are Diverse Datasets Needed in Ai? → Question,” Lifestyle → Sustainability Directory, Mar. 19, 2025. https://lifestyle.sustainability-directory.com/question/why-are-diverse-datasets-needed-in-ai (accessed Dec. 03, 2025).

[2] V. Team, “AI Governance Lexicon,” Verifywise.ai, 2023. https://verifywise.ai/lexicon/fairness-audits (last accessed Dec. 02, 2025)

[3] “Automated Employment Decision Tools (AEDT) — DCWP,” www.nyc.gov. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page (last accessed Dec. 03, 2025)

[4] S. Chaturvedi and R. Chaturvedi, “Who Gets the Callback? Generative AI and Gender Bias,” arXiv.org, 2025. https://arxiv.org/abs/2504.21400 (last accessed Dec. 02, 2025)

[5] Reuters, “Amazon Ditched AI Recruiting Tool That Favored Men for Technical Jobs,” The Guardian, Oct. 11, 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine (last accessed Dec. 02, 2025)

[6] E. Ferrara, “Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies,” Sci, vol. 6, no. 1, p. 3, Dec. 2023, doi: https://doi.org/10.3390/sci6010003. (accessed Dec. 02, 2025)

[7] A. Kersley, “Couriers say Uber’s ‘racist’ facial identification tech got them fired,” WIRED, Mar. 2021. https://www.wired.com/story/uber-eats-couriers-facial-recognition/ (accessed Dec 6. 2025)

[8] U.S. Equal Employment Opportunity Commission (EEOC), “iTutorGroup to Pay $365,000 to Settle EEOC Discriminatory Hiring Suit,” EEOC Newsroom, Sep. 11, 2023. https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit (accessed Dec. 04, 2025).