Exploring how to Mitigate Algorithmic Bias in AI-driven Automated Hiring Systems

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Ilyas Jaghoori

Summer, 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Bryn E. Seabrook, Department of Engineering and Society

Introduction to Algorithmic Bias in Current Hiring Systems

Artificial intelligence (AI) has emerged as a transformative force in modern society, driven by advancements in datasets, algorithms, and hardware that have ushered in a new "AI Spring.". In the workforce, specifically the recruitment and hiring process, the usage of “AI implementation can potentially provide a competitive advantage by enabling a better understanding of talent compared to competitors, thereby enhancing the company’s competitiveness” (Chen, 2023). Tools used in recruiting includes AI powered systems that scan through resumes to identify best fit candidates based upon predefined criteria, and machine learning algorithms used to predict the likelihood of a candidate succeeding at the company. However, AI’s black and white approach to decision making in hiring processes is extremely vulnerable to bias. For example, a common problem that exists in AI CV/Resume screenings is the issue of measurement bias. Measurement bias is a form of bias in which the “training data for AI algorithms inadequately represents the intended construct it seeks to measure” (Albaroudi et al., 2024). This bias has the potential to exclude qualified candidates whose qualifications may not perfectly align with what the trained algorithm is looking for, creating an unfair playing field.

In order to approach a potential resolution to this problem, the question must be asked: what is currently lacking that makes algorithm bias so prevalent? To find an appropriate answer to this question, it is necessary to incorporate the Actor Network Theory (ANT) to analyze the necessary interplay between the human and non-human actors in this system. ANT emphasizes the importance of understanding the complex relationships between the different actors in a system. Instead of viewing actors as separate entities, ANT looks to understand how they come to form networks of relationships and interactions. In the case of this research, instead of a strictly AI or recruiter approach to application screening, there must be an effective combination of the two that integrates human oversight with AI tools to ensure a balanced and fair evaluation process that mitigates biases and leverages the strengths of both actors.

Data Collection and Analytical Approaches

This study uses both quantitative and qualitative data from publicly available sources on hiring statistics from companies. The specific methods consist of analyzing available company research containing information about AI algorithmic performances and reports, as well as examining case studies with a general sociotechnical analysis that emphasizes the importance of ANT. The key words guiding the research are: Algorithmic bias, AI hiring systems, Actor Network Theory (ANT), human supervision, HR professionals, and AI ethics.

The remaining portion of the paper consists of three major sections. The first is a detailed methodology explaining the results and discussion of the research. The results and discussion focus on two major topics: the integration of human supervision with AI tools and the effectiveness of the ANT framework in the context of AI in hiring. The second major section provides information about the limitations of the study and potential future research, and the last section involves a conclusion of the study.

Exploring the Prevelance and Impact of AI Bias

In order to understand the consequences of algorithmic biases in AI resume screenings, it is vital to consider the context in which they occur. Classic screenings in which a human is able to make reasonable decisions based on a multitude of factors such as soft skills, culture fit, and body language are phasing out rapidly. It is noted that “Eighty-eight percent of companies globally already use AI in some way for HR, with 100 percent of Chinese firms and 83 percent of U.S. employers relying on some form of technology" (Brin, 2024). This reliance on AI significantly reduces the amount of resources companies have to spend on recruitment processes.

However, on top of the biases that these algorithms tend to pick up on mentioned earlier, they also lack some fundamental aspects of recruiting that a human would excel at. This is due to the fact that AI algorithms do not have the capability to fully understand an applicant just through detecting key words or numbers in files. What these algorithms do is search for patterns in which it was trained to make decisions. This approach fails to account for the nuanced qualities of a candidate which are often discernible through human interaction and judgment. The reliance on AI can lead to the exclusion of well qualified candidates who might not fit the harsh patterns established by the training data.

The unfair patterns extend deeper. Due to the lengthy history of racial and gender prejudice, either intentional or unintentional, AI algorithms pick up on these trends. When these biases exist in datasets, "AI may replicate these prejudices in its decision making" (Chen, 2023). For example, if data has shown a preference for certain demographic groups over others, the AI system will eventually learn to favor these groups. As a result, this can severely put less-favored candidates in disadvantageous positions leading to a less diverse workforce.

This bias doesn't only exist within datasets. In virtual environments, where candidates are tasked with recording and answering prompted questions, AI models find a way to unfairly discriminate against candidates based on voice detection. The AI model uses concepts such as "the idea that the intonation of our voice can predict how successful we will be in a job" (Corbyn, 2024). This can lead to discriminatory practices based on factors such as accents and speech patterns. Candidates with non-traditional forms of speech may seem to deviate from the model's training data and be unfairly judged as less competent for the role. Even though a candidate may be the most qualified for the job, they will be seen as inadequate by the algorithm if they do not fit within the scope of acceptable forms of speech.

Lastly, concerns about using AI algorithms for recruiting are reflected by recruiters. According to Wardini, 35% of current recruiters believe that using AI for these processes can and will eventually destroy the HR industry (Wardini, 2023). As these algorithms can simply automate the process of screening candidates and make decisions much quicker than their human counterparts, the need for human HR workers will eventually become obsolete. Another study finds that "while 96% of recruiters believe AI can help them in their current jobs, 60% are afraid it will eventually kick them out of work" (K, 2024). This increasing reliance on the use of AI can really shift the landscape of the HR industry and put millions of workers out of jobs.

Actor-Network Theory in Understanding the Interplay in Hiring Systems

To understand how to mitigate algorithmic bias, it is essential to find what STS framework should be used to successfully analyze the problem. Actor Network Theory (ANT) is the framework that provides the most comprehensive approach to examining the complex interplay between human and non-human actors in the issues of automated hiring systems. When using ANT to understand this complex engineering feat, the approach is not to emphasize solely the technical factors, but rather, the importance of the relationships and interactions among the actors that are involved. ANT is essential when considering large-scaled engineering projects as it supports James Trevelyan's emphasis on the significance of non-technical factors, such as social interactions and relationships, that go into delivering results in accordance with real-world expectations. Bruno Latour's *Aramis, or the Love of Technology* does the best in presenting the ideas of ANT with a clear and detailed story. In discussing the entire lifespan of the rise and fall of a complex public transportation project, Latour shows the mix of social, technical, and political entities that affect engineering work making ANT more understandable. This comprehensive approach of ANT helps to expose the many challenges faced in algorithmic biases.

In the initial scanning of the preface, epilogue, and prologue of *Aramis, or the Love of Technology*, it seemed to be a text that hyper fixated on a failed transportation system without any clear reason. However, Latour's in-depth analysis of the creative process and eventual fall of the Aramis project reveal to the readers the great depth of interactions that defined it. For example, the chronological ordering of the steps in the project was given by Latour in the prologue: "1974, February: Final report on Phase 0; creation of the Aramis development committee…...1976: Final report on Phase 1; Aramis simplified for economic reasons" (*Aramis, or the Love of Technology,* 13). This demonstrates that the Aramis project was not a simple technical project; but rather, it was a multifaceted one that took the effort of different actor networks to progress the project. However, the failure in managing the networks was why the Aramis project came to an untimely end. Latour expresses that the lack of effective coordination with the sheer complexity of actors made it difficult to rally around the Aramis project. Latour states: "Now, this variation in the relative size, in the representativeness of the actors, is not limited to Mr. Petit; it characterizes all members of a technological project. Mr. Lagardère supports the project, to be sure, but who can say whether his stockholders will follow? " (*Aramis, or the Love of Technology*, 45). In showing this example of how this complex network lacked the fundamental characteristics of deploying a successful project such as coordination and leadership, Latour provides a blueprint for all future practicing engineers to follow. Instead of following in the footsteps of this failed project, Latour preaches the Actor-network theory in a way that ensures that an engineering project is negotiated appropriately, loved thoroughly, and researched extensively with its many actors, making it easily understood and implementable.

"Diving in Magma" by Tommaso Venturini takes a different approach in discussing the Actor-network theory compared to Latour, but it falls short in conveying its full meaning in the practice of real-world engineering challenges like automated hiring systems. Unlike Latour's more direct approach, Venturini stirs the argument that it is useful to deal with controversies when examining the Actor-network theory. Venturini states that the "cartography of controversies is the exercise of crafting devices to observe and describe social debate especially, but not exclusively, around technoscientific issues." ("Diving in Magma", 258). He states that when dealing with controversies, individuals need to do more than to "just observe", but instead be more open and curious to assess different viewpoints. In theory, this way of approaching Actor-network theory is valuable in the sense that it allows engineers to think more deeply about important issues in society such as climate change. However, what Venturini fails to do is convey the importance of dissecting controversies in the practice of real-life engineering projects. Venturini even states it himself: "Since its introduction, the cartography of controversies has someway served as an educational version of Actor-network theory." ("Diving in Magma", 258). The essay does not go in depth as to how practicing engineers in today's age can use the information given to them and directly apply it to the line of work that they're in. Venturini's failure in applying the theoretical knowledge to the practice of real-world engineering makes it difficult for engineers to effectively use ANT in addressing current problems such as those found in automated hiring systems.

Using this prior research on ANT in the context of automated hiring systems, the framework points out the importance of examining not only AI algorithms, but its relationships with other actors such as training data, HR professionals, and job applicants. For example, when examining the major issues of algorithmic biases, it may be easy to simply classify it as just a technical problem; however, there is much more that goes into the issue once analyzed through the scope of ANT. Algorithmic biases stem from historical prejudice that continues to impact our society. Historical biases such as men predominately working in the engineering workforce and much of the HR field being women is used as data that the algorithms work with. Though unfair and unrepresentative of how the workforce performs nowadays, these algorithms make decisions based solely on this historical data.

Furthermore, HR professionals play as a pivotal actor that algorithms use to evaluate job candidates. They collaborate with developers to define what makes a candidate qualified that serves as initial input. This input directly influences the behavior of the AI system and if it happens to reflect biases or outdated standards, it is perpetuated by the algorithm and continue to foster unfair advantages and disadvantages. Additionally, HR professionals can sometimes help in interpretating decisions that algorithms make. Though most systems use AI algorithms as initial screenings that make immediate decisions without human consultation, HR professionals can sometimes view if the system made the right decision. As an intermediary, they have the power to help make these systems fairer and through this constant feedback loop, they can make or break the future of automated hiring.

Though it may not seem like it, job applicants are not exempt in this unfair system and must also be analyzed as actors. As developers create these algorithms based on a pre-defined criteria that their companies emphasize to be most important, applicants can cheat the system by tailoring their resumes to match what the algorithm is looking for. By doing this, they may present an inaccurate version of themselves, but to the eyes of the algorithm, they are a great fit. Like the prejudices mentioned earlier, this gives these applicants an unfair advantage to the system. This behavior creates a loop that further enforces biases to hiring and makes it so that diversity is reduced again within firms.

Using ANT to analyze automated hiring systems exposes the critical roles played by its various actors. This approach reveals that addressing algorithmic bias requires more than just technical fixes, but also a comprehensive understanding of the interconnected network of influencers and interactions. By leveraging ANT, organizations can have a better understanding of automated hiring issues and learn to develop more equitable AI-driven hiring practices.

Mitigating Bias through Human-AI Collaboration

This study finds that integrating human supervision with AI tools, creating an environment of feedback loops rather than relying on a fully autonomous AI system, effectively reduces algorithmic bias in hiring processes. Actor Network Theory (ANT) highlights how the interactions between human HR professionals and AI systems are crucial for this integration. By incorporating human expertise to review and adjust AI decisions, biases that AI alone may overlook are identified and corrected. ANT demonstrates that a collaborative approach, where human evaluators and AI systems continuously interact, leads to more balanced and fair hiring decisions. This method leverages the strengths of both actors: the efficiency and data processing capabilities of AI, and the contextual understanding and ethical considerations of human professionals. By fostering a dynamic interaction between these actors, the hiring process not only mitigates biases but also enhances the overall quality of hiring decisions. To mitigate bias in

AI hiring systems, it is essential to implement a structured feedback loop where human HR professionals regularly review AI decisions, provide corrections, and refine the algorithms. This continuous interaction ensures that AI tools evolve and improve, aligning more closely with fair and equitable hiring practices.

Integration of Human Supervision on AI Hiring Systems

Though susceptible to biases, the usage of Artificial Intelligence is only growing in the field of decision making, and the growth is bound to continue. This means that in the scope of hiring systems, AI is not going anywhere. In fact, there are numerous reports and studies that show the effectiveness of AI in the hiring process as the amount of time needed to evaluate potential candidates is significantly reduced. For example, Hilton's use of AI for talent acquisition showed that "the software evaluated the candidates and their information, increasing hiring rates by forty percent and reducing vacant position fill-up time by 90%" (Wirick, 2023). Another study shows that a global leader in consumer goods, Unilever, was able to significantly reduce time spent on the recruitment process using AI. According to a HireVue assessment of the company, "Not only has the process been significantly improved for candidates, saving over 50,000 hours in candidate time, the Unilever team has seen over 1 million pounds in savings in just one year" (HireVue, 2023). These studies confirm that AI will continue to be used in this field. However, as mentioned earlier, the existence of bias in these systems can not be pushed under the rug, which introduces the counterargument of *only* using humans for hiring systems.

Traditional versions of hiring systems in which a human was the sole recipient and reviewer of applications have proved to be effective throughout time. Companies have known to invest heavily towards their HR programs because at the end of the day, these are the individuals who keep the company populated with talent. Talent acquisition will continue to be crucial for a company's success, but the dominance of traditional HR practices may diminish as AI and automation become more prevalent in the sector. According to a report by Mckinsey & Company, "by 2030, activities that account for up to 30 percent of hours currently worked across the US economy could be automated – a trend accelerated by generative AI" (Ellingrud et al., 2023). This could include major HR operations such as resume screening, initial candidate assessments, and potentially aspects of onboarding and training. However, what gets suppressed in this surge of automated excitement is the loss of personable guidance on critical HR tasks. To take away the human aspect of HR means to get rid of vital skills needed to understand applicants on a deeper level. Machines can only understand applicants to a certain extent, making the continued involvement of humans in the process essential.

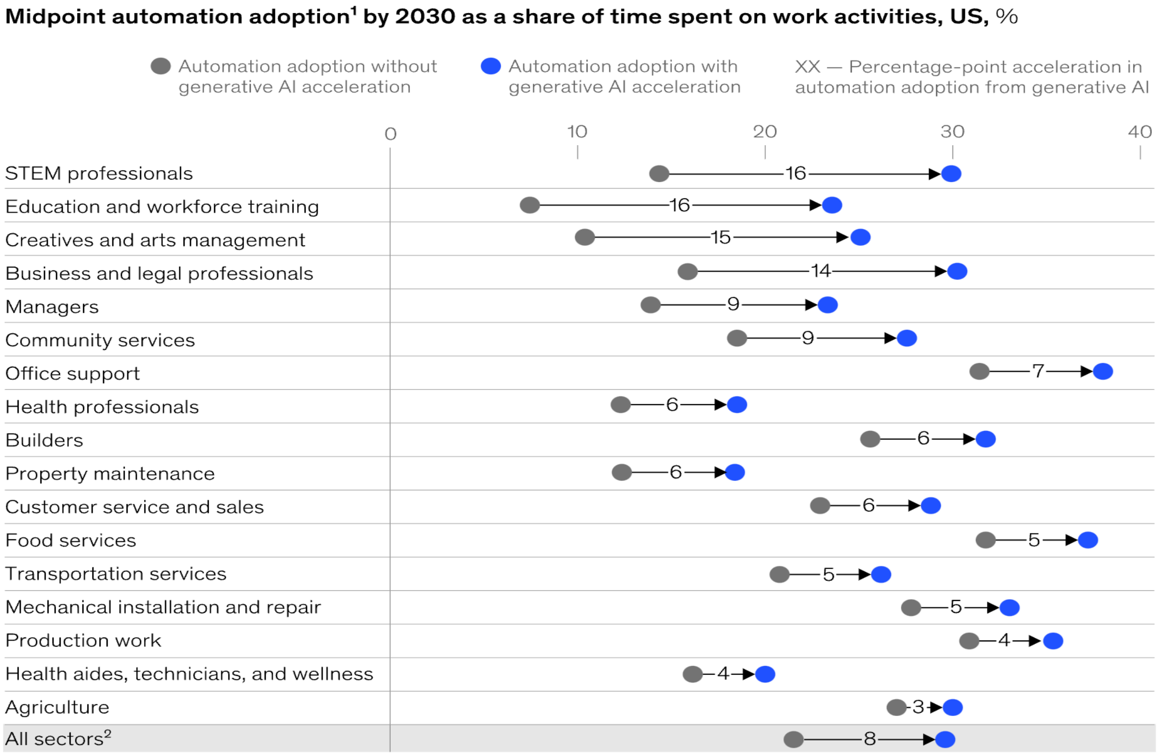
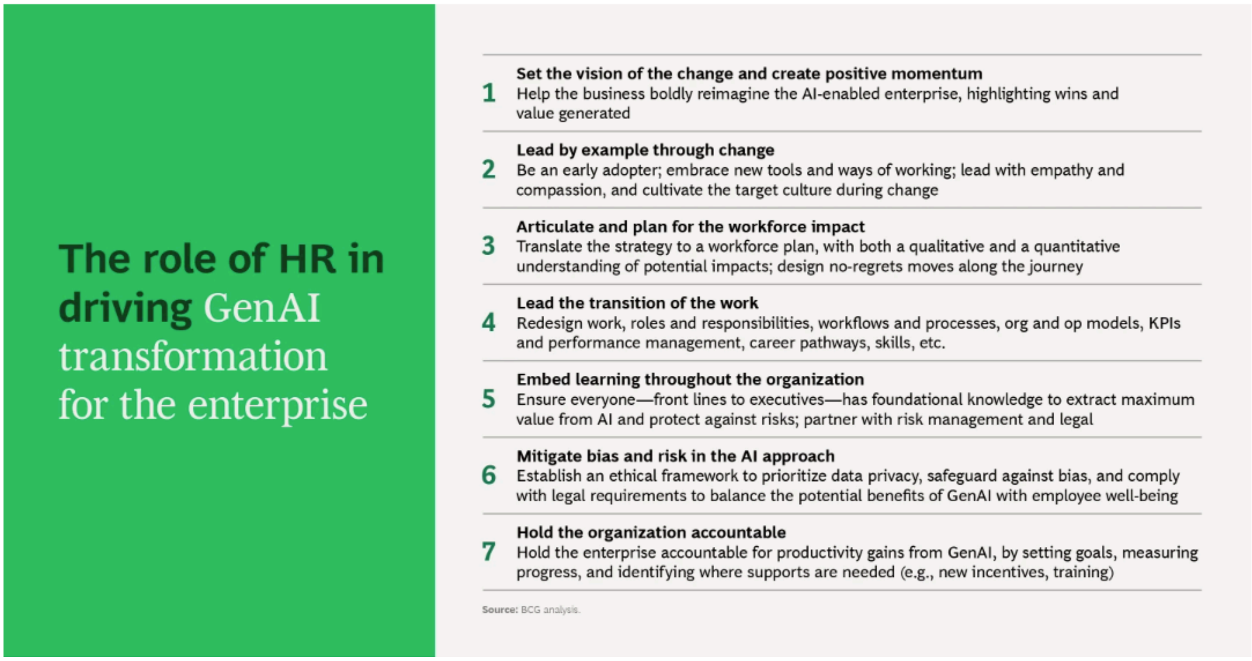


Figure 1. How specific roles can be automated by the year 2030 (Ellingrud, 2023).

Given the emergence of AI automation in HR and the biases inherent in these systems, the most effective solution would be to integrate human supervision into AI processes. According to an article published by the U.S. Department of Labor, they emphasize the fact that "organizations should have clear governance systems, procedures, human oversight, and evaluation processes for AI systems for use in the workplace" (Artificial Intelligence and Worker Well-Being: Principles and Best Practices for Developers and Employers, n.d.). The Department of Labor is clearly emphasizing the fact that AI must be double checked. By having HR professionals who are skilled in selecting best candidates supervise the decisions that the AI algorithms make, organizations can ensure that decisions made are ethical, fair, and aligned with their values. Additionally, the integration of human supervision in this system is essential for continuous improvement of AI algorithms. AI models are trained on historical data as mentioned previously that are not up to date with market trends or future needs. HR professionals are useful in providing ongoing feedback and updates to these models to make them as effective as possible. By constantly looping the AI output to the human HR professional, the role of the professional verifies that the system refrains from making decisions that are based on bias that may have been overlooked in the initial training phase. According to a SHRM article that references a Gartner study by Eser Rizaoglu, a senior director and analyst in Gartner's HR practice, "Having a dedicated role within HR, potentially sitting within the HR IT subfunction, is becoming more important, as it can help coordinate and plan GenAI efforts for the HR function and help build an HR AI strategy while also working with key cross-functional stakeholders and HR tech vendor partners to mitigate risks and implementation challenges" (Maurer, 2024). Another article written by The Boston Consulting Group (BCG) emphasizes that human interaction is still crucial in the implementation of AI in hiring systems. As detailed in Figure 2 below, the group emphasizes the important roles of HR in driving GenAI transformation. This new role for HR professionals leverages their strengths as well as the power of AI to effectively revolutionize the nature of their work as well as the future of hiring systems.



The role of HR in driving GenAI transformation for the enterprise

1 **Set the vision of the change and create positive momentum**
Help the business boldly reimagine the AI-enabled enterprise, highlighting wins and value generated

2 **Lead by example through change**
Be an early adopter; embrace new tools and ways of working; lead with empathy and compassion, and cultivate the target culture during change

3 **Articulate and plan for the workforce impact**
Translate the strategy to a workforce plan, with both a qualitative and a quantitative understanding of potential impacts; design no-regrets moves along the journey

4 **Lead the transition of the work**
Redesign work, roles and responsibilities, workflows and processes, org and op models, KPIs and performance management, career pathways, skills, etc.

5 **Embed learning throughout the organization**
Ensure everyone—front lines to executives—has foundational knowledge to extract maximum value from AI and protect against risks; partner with risk management and legal

6 **Mitigate bias and risk in the AI approach**
Establish an ethical framework to prioritize data privacy, safeguard against bias, and comply with legal requirements to balance the potential benefits of GenAI with employee well-being

7 **Hold the organization accountable**
Hold the enterprise accountable for productivity gains from GenAI, by setting goals, measuring progress, and identifying where supports are needed (e.g., new incentives, training)

Source: BCG analysis.

Figure 2. The Role of HR in Driving GenAI Transformation for the Enterprise (Bedard, 2023).

Effectiveness of the ANT Framework in the AI/HR Hiring System

This new system, where HR professionals actively interact with and are influenced by AI algorithms in the hiring process, necessitates its understanding through the lens of Actor-Network Theory (ANT). As noted earlier in the complex engineering project of Aramis in Latour's analysis, the system seemed to be unconnected with its many actors. Projects that tend to have a vast number of actors are more susceptible to this flaw, which often leads to its demise. In the context of this hiring system, it is shown that the separation of these major actors has led to inefficiencies and biases discussed earlier. By applying ANT, the interconnectedness of human supervision and AI is better managed. This connection reduces the possibility of mistakes from both parties and allows for a more holistic review of applicants. By leveraging AI's ability to scan through resumes and CV's quickly as well as the ability of HR professionals to more accurately find alignments with applicants, this connected system enhances both efficiency and fairness of the hiring process.

The study "Fairness and Abstraction in Sociotechnical Systems" by Selbst et al. (2019) explores various traps that researchers might encounter when developing fair machine learning/Artificial Intelligence systems. These concepts are extremely relevant to understanding the connections between AI and human actors in the hiring process. The **ripple effect trap** discusses the unintended consequences that can arise when technology is introduced into social systems that can potentially alter behaviors and decisions. In the context of this research paper, the introduction of AI tools in the system of hiring talent can inadvertently reinforce biases if not managed properly. ANT provides a crucial framework to understand and manage these interactions so that both AI and human judgment work together to mitigate such effects. The **formalism trap** involves the oversimplification of social concepts that fail to capture their full meaning. In the context of only using AI technology for hiring, this means that simples metrics are insufficient, and a broader sociotechnical perspective is necessary to address the ethical nuances of hiring decisions. The **solutionism trap** warns against assuming technical solutions are always the best approach. Again, in the context of only using AI technology for hiring, it highlights the importance of integrating human supervision with AI and not letting the technology do all the work.

By introducing these concepts and applying ANT, this research can better manage the interconnectedness of AI and human HR actors. This approach can potentially lead to improved hiring outcomes and a more balanced evaluation of candidates using AI's efficiency and human judgement.

Addressing the Gaps and Future Directions

Limitations inevitably exist when considering the scope of this research. To begin, this analysis looks over a new modern style of hiring practice in which an applicant sends a resume/CV to be processed by AI algorithms. This initial screening exists in most of the company hiring processes but not all. For those in which a different style of screening exists, this research does not reach that scope. Secondly, statistics found for this research were taken from larger companies in which data is publicly available. For smaller to mid-tier companies that make up most industries, data is not publicly available to be analyzed. This also limits the scope that the research seeks to touch upon.

Future research conducted can possibly be influenced by the solutions provided in this paper. By leveraging the combination of both AI algorithms and human HR professional intervention in hiring processes, the data on hiring will begin to transform. This shift will begin to phase out past biases that existed and will foster a more equitable hiring landscape. Algorithms in the future will begin to make more fairer judgements after training on these datasets and will slowly begin to have a sense of its own bias detection and mitigation mechanisms.

Conclusion

In conclusion, this research shows that integrating human supervision with AI in hiring processes has the potential to reduce algorithmic bias and enhance both the efficiency and fairness of candidate evaluations. By applying Actor-Network Theory, the complex interaction between AI algorithms and human HR professionals can come to life and be shown to be most effective. By applying this integration, this approach addresses inherent biases in AI systems and shows the strengths from both actors. The broader significance of this research lies in its potential to reshape the hiring landscape, ensuring fairer and more transparent practices that benefit both employers and applicants. As organizations increasingly adopt AI technologies, this study provides valuable insights for developing ethical and effective hiring practices. The takeaway message is clear: a hybrid approach combining human oversight with AI tools not only mitigates bias but also fosters a more inclusive and just employment environment.

References

Albaroudi, E., Mansouri, T., & Alameer, A. (2024, February 7). A comprehensive review of AI techniques for addressing algorithmic bias in job hiring. MDPI. https://www.mdpi.com/2673-2688/5/1/19#:~:text=Algorithmic%20biases%20may%20perpetuate%20prejudices,inequitable%20access%20to%20job%20opportunities.

Brin, D. W. (2024, January 2). Employers embrace artificial intelligence for HR. Welcome to SHRM. https://www.shrm.org/topics-tools/news/employers-embrace-artificial-intelligence-hr

Chen, Z. (2023). Ethics and Discrimination in Artificial intelligence-enabled Recruitment Practices. *Humanities and Social Sciences Communications*, *10*(1), 1–12. https://doi.org/10.1057/s41599-023-02079-x

Corbyn, Z. (2024, February 3). *The AI tools that might stop you getting hired*. The Guardian. https://www.theguardian.com/technology/2024/feb/03/ai-artificial-intelligence-tools-hiring-jobs

Wardini, J. (2023, November 21). *12 must-know statistics on how many companies use AI in hiring*. ARTSMART AI. https://artsmart.ai/blog/how-many-companies-use-ai-in-hiring/#:~:text=Over%2088%25%20of%20companies%20worldwide,in%20HR%20processes%2C%20including%20recruitment.&text=AI%20technology%20is%20widely%20used,to%20both%20firms%20and%20recruiters.

K, D. (2024) *Ai recruiting: The Complete Guide*, *Recruiterflow Blog*. Available at: https://recruiterflow.com/blog/ai-recruiting/

Latour, B. (1996). Aramis, or the love of technology (C. Porter, Trans.). Harvard University Press. https://dss-edit.com/plu/Latour-B_Aramis-or-Love-of-Technology_1996.pdf

Venturini, T. (2009). Diving in magma: how to explore controversies with actor-network theory. Public Understanding of Science, 19(3), 258–273. https://doi.org/10.1177/0963662509102694

Artificial Intelligence and Worker Well-being: Principles and Best Practices for Developers and Employers. (n.d.). DOL. https://www.dol.gov/general/AI-Principles

Ellingrud, K., Sanghvi, S., Singh Dandona, G., Madgavkar, A., Chui, M., White, O., & Hasebe, P. (2023, July 26). Generative AI and the Future of Work in America. Www.mckinsey.com; McKinsey Global Institute. https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america

HireVue. (2023). Global Talent Acquisition Strategy | HireVue + Unilever. Hirevue.com. https://www.hirevue.com/case-studies/global-talent-acquisition-unilever-case-study

Maurer, R. (2024, April 13). HR Adopts AI. Www.shrm.org. https://www.shrm.org/topics-tools/news/all-things-work/ai-hr-challenges-opportunities

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, 59–68. https://doi.org/10.1145/3287560.3287598

Wirick, G. (2023, March 22). The Impact of Artificial Intelligence on Recruiting. RecruitingDaily. https://recruitingdaily.com/the-impact-of-artificial-intelligence-on-recruiting/