Framework for AI Fairness

Evaluation Plan

Matthew Lowe, Elizabeth Kahle, Can Deniz Balkaya, Yifei Li

Boston College

Table of Contents

Section I. Evaluation Overview

Overview

The goal of Framework for AI fairness is to reduce bias in AI-powered hiring platforms. The purpose of this proposal is to test how LinkedIn's framework for AI fairness, a toolkit of design principles meant to reduce bias and increase social equity (Quiñonero-Candela, 2023), works within the context of other companies. Now that generative AI is taking hold across the globe and increasing the types and extent of work that can be done across all fields, it is of utmost importance that AI can strike a balance on providing equal/equitable opportunities for all. If bias remains in AI usage for job screening processes, we could see the bias and social stigmas present in society continue to be exacerbated. As a result, evaluation must be done for this program to fight for fairness in a process that already has a great deal of bias.

The main participants in the evaluation throughout the proposal will be a panel of companies utilizing LinkedIn's Artificial Intelligence fairness framework and the evaluation team. Necessary expertise and resources include hiring-related datasets; diversity, equity, and inclusion experts to identify potential bias; data analysts with machine learning expertise; and collaboration with programmers and statisticians. The primary audience for this evaluation is companies that have adopted AI in their hiring platforms. Additionally, the conclusions drawn from this evaluation will provide recommendations for reducing issues related to the fair use of AI frameworks and adopting a less biased approach, thus contributing to a fairer and more equitable hiring process. The Evaluation Team is responsible for evaluating the AI Fairness Framework by collecting and analyzing data to assess its effectiveness and impact. The evaluation team will use the results to provide actionable insights and recommendations for improvements to the framework and communicate the results to other stakeholders.

Timeline

The evaluation will last about six months, to allow sufficient time to evaluate the effectiveness of LinkedIn's AI fairness framework in the context of other companies. This is necessary in order to evaluate the fairness of the AI fairness framework in recruitment.

Section II. Needs Analysis and Stakeholders

Needs Analysis

Generative AI tools used for hiring processes are marketed for their efficiency and objectivity, however AI has its limitations. There is an imperative need for evaluation teams like ours to help companies identify where limitations in their AI systems exist and how to mitigate the possible biases they are perpetuating.

Growing need for artificial intelligence in the hiring process. As AI technology becomes more prevalent in the hiring process, companies across industries, including giants like Google, Meta, and Microsoft, are beginning to use AI-powered hiring processes as a way to improve hiring efficiency and streamline candidate selection, making the fairness of these systems important. Development is still early on and many are hesitant to give full reigns to AI in determining who gets hired so the biggest impact AI has made is its usage in the screening process.

Reducing bias and guaranteeing fairness. The importance of fairness has been increasingly recognized in modern society, and there is a growing call for social justice. Having a fair chance when recruiting is one of those things. However, the hiring process can be subject to a variety of biases[1] that result in unequal opportunities for candidates from different backgrounds. One instance of bias being flagged in company AI models is in 2018 when Amazon had to scrap an AI recruiting tool that was systematically ranking the resumes of female candidates as lower than those submitted by males who had equal qualifications: “Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry” (Dastin 2018). Failure to ensure fairness in AI recruitment processes risks perpetuating these biases. Therefore, in order to provide equal and fair opportunities to all job seekers, we should ensure that AI follows the principle of fairness in the hiring process.

Implementing transparency and accountability. Due to concerns from some stakeholders about the fairness and ethical use of the AI recruitment framework, implementing transparency and accountability is necessary to build stakeholder trust in this hiring framework.

---

[1] In this study, the bias focused on will be the difference between the diversity rate observed in recruitment processes due to any reason and the diversity rate observed in the real world.

Establishing strong documentation practices, including transparent disclosure of training data sources, algorithms, and a clear audit system, can help enable stakeholders to have a comprehensive understanding of the framework's operations and thus trust in the framework.

Improving data quality. Data is one of the foundations on which the AI hiring framework operates; training and improving the quality of the data is an important way to improve the bias situation. The data should cover a wide range, be representative enough of different user groups, and not contain any discriminatory information.

Stakeholders

Companies using LinkedIn's Artificial Intelligence Fair Hiring Framework. These companies are direct users of the framework, and their main goal is to optimize their hiring methods and increase efficiency while ensuring fair hiring and attracting diverse talent. These companies will use the results of this evaluation to refine and adapt the framework to meet their needs and create a more inclusive and fair hiring process.

Job Seekers. Job seekers are the main beneficiaries of the AI fairness framework, and their main claim is to get a fair hiring process that is not biased because of any of their personal backgrounds.

HR employees and employees involved in the hiring process. HR employees and employees involved in hiring are the main implementers of the AI fairness framework. Their goal is to recruit employees efficiently and correctly and avoid false negative and false positive hiring outcomes. These employees will use the framework in their daily work, so the results of the evaluation will help them in their work and promote a fair and equitable recruitment process.

Programmers and data analysts. Programmers and data analysts are the primary builders and maintainers of AI fairness frameworks and are interested in creating and optimizing algorithms that reduce bias in hiring decisions. They will take into account the results of the evaluation and make adjustments to the technical aspects of the framework to improve its fairness and accuracy.

Section III. Intervention Description and Visualization

The intervention being both implemented and evaluated is a case study of how well Linkedin’s Equal AI Treatment principles can be utilized in building a framework for ethical AI

use by other companies. The framework of the logic model has three columns that each signify one phase of the developmental evaluation process and three rows that each represent one of the three principles. These principles will act as the intervention's inputs and give a lens to stakeholders to view the rationale behind each activity's inclusion in the evaluation as well the many aspects behind AI fairness.

To set the stage for the explanation of the intervention, these principles should be stated and can be found in Quiñonero-Candela et al (2023).

1.) "We will measure and work to mitigate algorithmic bias so that our AI systems treat everyone equally."
2.) "We will not consider equal AI treatment the end of our work but will treat it as the foundation of a broader fairness and equity strategy."
3.) "We will validate our approach externally and lead with transparency in this developing field."

Additionally, it should be understood that, in general, all activities conducted within the same phase will occur simultaneously.

The first phase of this intervention cycle revolves around exploring and developing bias mitigation strategies as well as collecting quantitative data on AI fairness metrics. The activities conducted in this phase will be preparatory and establish a foundation for the rest of evaluation to stand on. Activities will include conducting a root-cause analysis on the AI fairness measure of predictive parity[2], which ensures that the proportion of positive predictions that are correct is the same for all protected groups (Microsoft 2022), developing a framework for the specific company that incorporates fairness and equity into strategy, and conducting audits on in-use technologies via statistical analysis.

The second phase will focus on inclusion by developing strategies regarding bias mitigation, broader equity, and explore/exploit methods for algorithm testing. This algorithm testing in particular is useful for the company to assess how their own technologies have affected their employee demographics and ultimately the efficacy and success of their organization. Activities for this phase will build off of the outputs of the previous phase. This will include

[2] Synonymous with demographic parity

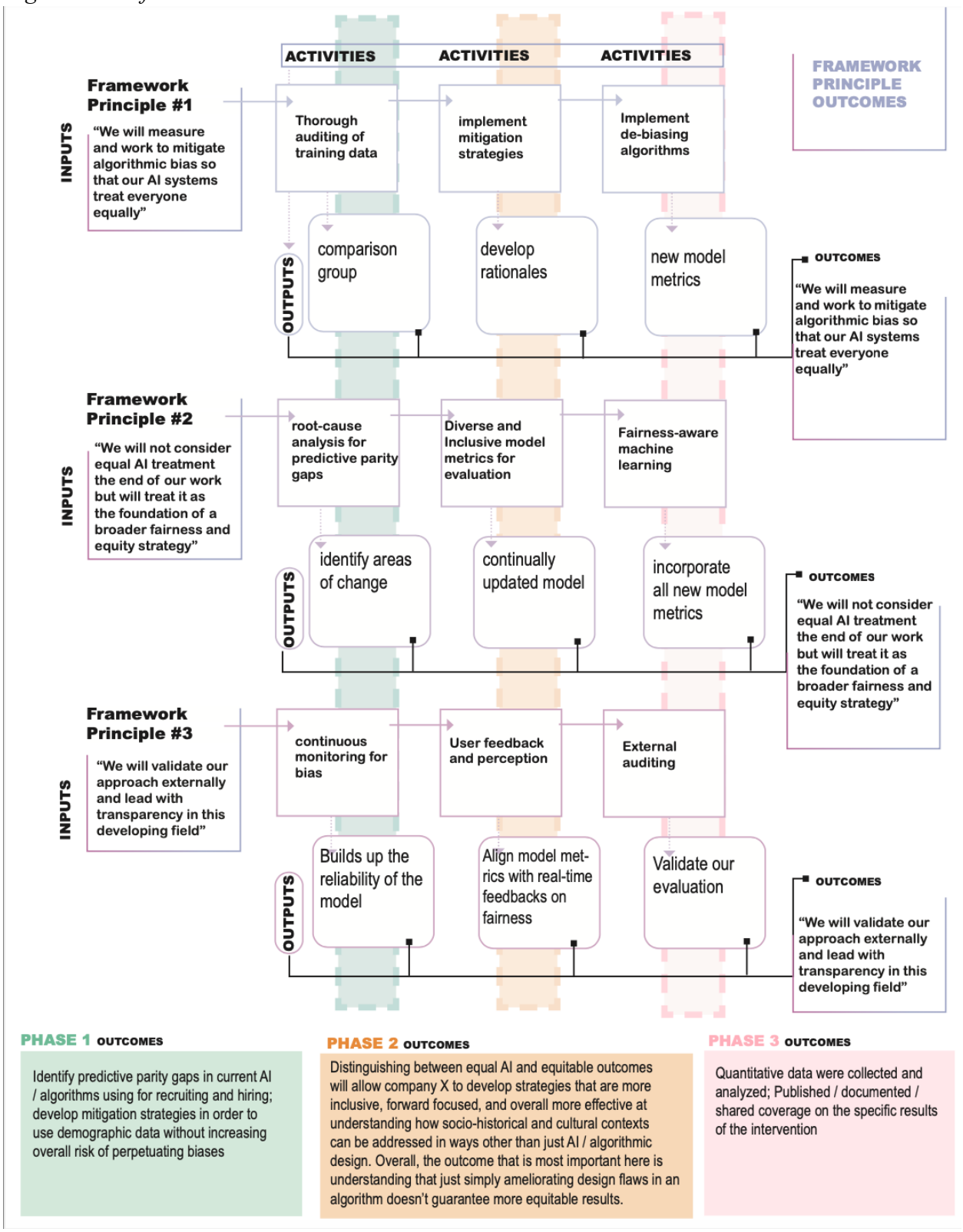
developing mitigation strategies for existing gaps present in the use of predictive parity, exploring broader equity strategies to address equitable outcomes, and the aforementioned explore/exploit strategy. The idea behind it is to explore for candidates that may be qualified for a certain position but not ensured by the algorithm. These individuals then have their data processed and translated to the algorithm as a qualified candidate which helps the algorithm grow.

The third phase of the cycle centers around analyzing data on the current context of the workplace and educating stakeholders on new innovative tools and techniques to be more progressive and accurate in their AI design and implementation. The activities for this phase include continuing to identify gaps in the algorithm's use of predictive parity and developing bias mitigation strategies that allow demographic data to be used without increasing the risk of perpetuating biases. Additionally, qualitative data specifically will be generated through written records of results and findings that will be shared with stakeholders both within and outside of the company. This should help reinforce the idea that we cannot simply intervene where the algorithm makes mistakes, but should develop strategies that are more inclusive and effective at understanding contexts. With the completion of these activities, the intervention can begin again at phase one by conducting the same activities on the new data and strategies generated from the previous iteration.

The outcomes that come from the activities related to each of the three principles framed within the intervention's three phases reveals the outputs generated. From these outputs come some of the intended impacts of the intervention which include reducing systemic bias due to AI in the job screening process, setting industry standards for fairness in AI hiring, and promoting long term social equity. There are, however, some concerns of unintended outcomes arising out of this intervention. For one, AI systems may struggle to correctly identify and address fairness issues which may lead to an increase of false results. False negatives could lead to oversight on different biases in the algorithm while false positives can lead to unnecessary adjustments to the algorithm that can affect the quality of the hiring process. There is also the possibility of not using the correct/most effective AI fairness metric which may compound bias. Predictive parity was chosen as a general standard established by Microsoft's Responsible AI Standard (Microsoft 2022) but, as we have discussed, context is key and there is no catch-all solution to determining what is fair.

Figure 1

Logic Model of AI Fairness Intervention



Section IV. Evaluation Questions & Criteria

Evaluation Questions

- Is the use of the AI framework helping to reduce bias in the hiring process?
    - Do AI frameworks contribute to similar representation of marginalized groups in the hiring process as compared to the real world?
    - How effective is LinkedIn's AI Fairness Framework for reducing algorithmic biases present in the historical database of Company X?

Criteria

The fairness of the AI hiring framework will be determined based on the following criteria:

Equality: The framework should use equality indicators to ensure that no group is unreasonably discriminated against in the hiring process using this system.

Transparency: Establishing strong documentation practices helps build trust in the framework by providing stakeholders with a comprehensive understanding of how it works.

Replicability: This framework should be adaptable and usable by different companies. That means the system needs to be replicable.

Sustainability: We expect this framework to make an ongoing contribution to hiring equality long after the evaluation is complete.

Outcome: The evaluation should meet our short and long-term goals of identifying and addressing specific biases in the hiring process and reducing them, improving effectiveness, helping some companies adapt the framework, and advocating for broader industry improvements.

Section V. Evaluation Design

Our evaluation project is a hypothetical intervention that follows a traditional explanatory case study design. The evaluation team chose to implement our intervention in one single case because of A) the novel framework we are evaluating, and B) unique phenomenon we are attempting to understand and delineate solutions for. In general, case studies are more holistic in their design approach, so this format enables more flexibility for the evaluation team to

incorporate social and cultural contexts in the model. With that said, our evaluation project is designed as a case study so that we can alter the intervention's activities based on real-time, ongoing feedback (i.e. outputs).

Purposive sampling grants us the ability to implement the intervention in the "case" (i.e. company) of our choosing, thus maximizing the results' potential for providing insight into the underlying mechanisms we are trying to understand through the intervention. By taking a sensemaking approach in our case study, we will alter algorithms to mitigate biases as they emerge, and we hope that these changes are perceived as collaborative by stakeholders in order to promote ongoing learning and reflection.

The evaluation team's primary goal in utilizing this design approach is to evaluate whether or not LinkedIn's AI Fairness framework – the three principles and their corresponding activities (i.e. data collection and analysis methods) – is effective at reducing algorithmic biases that may exist in company databases used for talent acquisition. An explanatory approach to the case study is significant to the design because it allows us to conduct a more in-depth analysis of all possible factors that may be contributing to patterns and relationships that may prevail in the data.

In terms of the data sources for our case study, we will be using the company's historical data for both the "treatment" and "control" in order to assess how effective LinkedIn's AI Fairness Framework is at reducing existing biases coveted in the company's database and AI algorithms. While explanatory case studies are not traditional experiments and thus do not have "treatment" and "control" groups, we instead consider the "control" as being the "comparison." The comparison dataset is yielded by the initial historical data audit. Likewise, the "treatment" group can instead be understood as the new outputs / results that are produced by altering the algorithms. The proceeding section will detail the data collection and analysis methods developed by our team using the three LinkedIn framework principles.

Section VI. Data Collection & Analysis

Almost all of the activities in our evaluation project are intended to produce quantitative data that we will compare to the company's existing dataset (i.e. historical data) used for talent acquisition. Overall, the data collection and data analysis methods used in explanatory case

studies are more in-depth as researchers are ultimately looking at every possible contributing factor to the observed outcomes in the results.

The activities outlined in our logic model are categorized into either A) the framework principle they apply to, or B) the phases (which are separated chronologically) in which they must be implemented throughout the duration of the intervention. The purpose of the phases is to highlight how the outputs impact which specific methods/strategies will be used within the next activity. For instance, as displayed in the logic model for our developmental intervention, the outputs of certain activities inevitably affect what strategies / methods of data collection are implemented in the next activity. Here is a step-by-step example:

1. The outputs of the root-cause analysis will be which predictive parity gaps exist
2. Depending on the context of these gaps this will determine which debiasing algorithm(s) is/are implemented in order to strategically mitigate bias.
3. Mitigation strategy outputs are thus generated via the implementation of various desbiasing algorithms; these outputs become the metrics used in the subsequent activity, which is implementing strategies for fairness-aware machine learning algorithms.

The data collection and analysis methods used will be delineated in accordance with the aforementioned classification system of our activities: by principle and by phase. All of these activities aim to answer the central question in our evaluation project: *How effective is LinkedIn's AI Fairness Framework for reducing algorithmic biases present in the historical database of Company X?*

Framework Principle #1: "We will measure and work to mitigate algorithmic bias so that our AI systems treat everyone equally"

Framework Principle #2: "We will not consider equal AI treatment the end of our work but will treat it as the foundation of a broader fairness and equity strategy"

Framework Principle #3: "We will validate our approach externally and lead with transparency in this developing field"

Table 1

Phase 1 Data Collection and Analysis Methods

Method/Strategy	Instruments (measurement tools used to collect and analyze data)	Measures & Indicators	Data Source	Important Notes / Comments
Audit Training Data **Review data sources** **Understand context of the data (such as time period it was collected)**	Programming language (such as Java or C++) Statistical softwares (such as r or python)	identify key variables. As this data is used for hiring, non-demographic key variables could include education level, work experience, etc. features in the data and analyze their distributions (z score)	Historical data (company database)	The findings of the audit should be recorded; any biases or imbalances clearly identified so that the steps can be taken in phases 2 and 3 to address all of them
Root-Cause Analysis for Predictive Parity Gaps	Programming language (such as Java or C++); Statistical softwares (such as r or python) Descriptive statistics can be used for this activity, such as mean, median, mode	Conduct a demographic analysis of existing data and calculate the distribution of key variables (from the previous step) within each demographic group. This identifies disparities in representation of different groups	Historical data (company database)	

Implement Debiasing Algorithms ***	Programming language (such as Java or C++) Statistical softwares (such as r or python)	Establish a baseline for model performance to compare the effectiveness of debiasing techniques Identify the current metrics in place for assessing bias in the model's predictions (see step below) Debiasing algorithms will be integrated into the model training pipeline Regularly evaluate the impact of these algorithms: *Are these debiasing algorithms improving model performance and fairness metrics?*	Historical data Literature review: explore existing debiasing algorithms and techniques proposed in relevant literature (this will become more clear and directed once root-cause analysis for predictive parity is done)	This step and the step below should ideally be done in tandem… ("implementing debiasing algorithms and establishing diverse and inclusive model metrics should be performed together or compared alongside one-another and adjusted accordingly") in order to assess how different metrics in the model lead to different results of predictive parity

Note. The method/strategies listed in red are designed in accordance with Framework Principle #1; The method/strategies listed in green are designed in accordance with Framework Principle #2; The method/strategies listed in purple are designed in accordance with Framework Principle #3.
Note. See Appendix Sec. III ***

Table 2

Phase 2 Data Collection and Analysis Methods

Method/Strategy	**Instruments (measurement tools used to collect and analyze data)**	**Measures & Indicators**	**Data Source**	**Important Notes / Comments**
Implement Mitigation Strategies* **	Programming language (such as Java or C++) Statistical softwares (such as r or python)	Mitigating existing predictive parity gaps ***	Historical data ; If new data needs to be added to the model in order to test whether a mitigation strategy is effective, then longitudinal data from the U.S. Census Datasets should be used.	***
Continuous Monitoring for Bias: Continue to quantify disparate impact by calculating adverse impact ratios to	Programming language (such as Java or C++); Statistical softwares (such as r or python)	Establish a feedback loop where insights gained from model performance in the intervention are used to refine the training data further	Continuous monitoring of historical data; incorporate changes to algorithms that occur throughout the intervention	This should be done routinely throughout the intervention, such as twice a week. This is why it is in phase 1.

assess the likelihood of different groups being selected or rejected				
Diverse and Inclusive model metrics for evaluation	Programming language (such as Java or C++) Statistical softwares (such as r or python)	Explore established fairness metrics Conduct demographic subgroup analysis to evaluate the model's performance across diverse groups Intersectionality analysis: explore how the model's impact varies for individuals with intersecting identities	New model that has been developed through adding / subtracting data from historical dataset based on debiasing algorithms and mitigation strategies	
User Feedback and Perceptions				*** note about this activity below in future directions

Note. User feedback and perceptions is a qualitative method and therefore is included only in our future directions. No actual qualitative data will be collected for analysis. The method/strategies listed in red are designed in accordance with Framework Principle #1; The method/strategies listed in green are designed in accordance with Framework Principle #2; The method/strategies listed in purple are designed in accordance with Framework Principle #3.

Note. See Appendix Sec. II ***

Note. See Appendix Sec. III ***

Table 3

Phase 3 Data Collection and Analysis Methods

Method/Strategy	**Instruments (measurement tools used to collect and analyze data)**	**Measures & Indicators**	**Data Source**	**Important Notes / Comments**
Implement Mitigation Strategies	Programming language (such as Java or C++) Statistical softwares (such as r or python)	Mitigating existing predictive parity gaps	Historical data ; If new data needs to be added to the model in order to test whether a mitigation strategy is effective, then longitudinal data from the U.S. Census Datasets should be used.	
Fairness-Aware Machine Learning This step is really	Programming language (such as Java or C++) Statistical	Incorporate fairness constraints by integrating fairness	Historical data Investigate fairness-aware machine learning frameworks and tools	In this strategy, researchers would normally collect qualitative data on how fairness is

important for understanding how the LinkedIn Framework's values and ethics interplay and compare to the culture at other companies	softwares (such as r or python)	constraints into the model optimization process to explicitly account for fairness in decision making	available in the field. During this activity, we could have stakeholders perform their own research of other frameworks or open the floor for discussion on frameworks they have been exposed to in the past; this will not be data that is collected but it could be a useful strategy for encouraging a more self-aware and collaborative work environment	defined by stakeholders and this would be quantified into a metric in the model; however, we are going to use the definitions of fairness and equity provided by the LinkedIn Framework. Future directions should incorporate this
External Auditing: engage external auditors with expertise in AI ethics and fairness - such as a stakeholder at another company who works in this department.	Programming language (such as Java or C++) Statistical softwares (such as r or python)	This is how we as evaluators can check the **reliability** of our model strategies and processes	Historical data **and** any new data that has been generated and added to the model; **all of this should be externally audited:** These findings should be made available to other companies, both the successful and not, significant and not;	The amount of outputs that will generate from this activity are numerous; transparency with data and our mitigation strategies for closing predictive parity gaps are the first steps for establishing **validity** of our intervention and its processes

Note. The method/strategies listed in red are designed in accordance with Framework Principle #1; The method/strategies listed in green are designed in accordance with Framework Principle #2; The method/strategies listed in purple are designed in accordance with Framework Principle #3.

Section Ⅶ. Evaluation Reporting & Use

The primary audiences for the evaluation include internal stakeholders (e.g., companies using the framework, HR professionals, and programming/software analysts) and external stakeholders (including job seekers and oversight organizations). To keep these audiences informed, we will provide regular updates throughout the evaluation, approximately every two weeks; a comprehensive interim report summarizing progress, challenges, and preliminary findings is scheduled to be released in about 3 months; and a final report detailing the full evaluation, findings, and recommendations will be released approximately 6 months after the evaluation begins. The written portion is in the form of an email report that succinctly communicates the purpose, methodology, and key findings and utilizes visualization tools such

as box-and-line diagrams to illustrate equity trends. In addition, an oral report during the final presentation will detail the approach, methodology, key findings, biases found, and actionable recommendations. This format ensures holistic understanding by internal and external stakeholders.

This evaluation identifies biases, discrepancies, or inequalities in the hiring process. Through continuous updating and comprehensive reporting, the evaluation results can guide internal stakeholders in making informed decisions to enhance fairness in hiring practices. Detailed evaluation results will be disseminated through reports and case studies. The evaluation results are designed not only to share insights, methodologies, and successful strategies with other companies but also to advocate for the adoption of fair hiring practices across the industry.

Section Ⅷ. Strengths & Limitations of Plan

This section will outline some of the strengths and limitations of this evaluation plan. The comprehensiveness of the program, contributions to social justice and the ability to track changes instantly are strong points of this plan, while the unknown bias-free nature of the dataset for comparison and the excessive time required to monitor changes can be considered weaknesses of the program.

Firstly, the comprehensive nature of the program can be considered its strongest aspect. The multitude of stakeholders implies coverage of many different areas of the artificial intelligence recruitment process, from software aspects to human aspects. Additionally, the steps of the assessment plan are highly suitable for providing comprehensive feedback. For example, the time between steps allows all relevant stakeholders to comment on developments, complete any deficiencies, or correct mistakes. Secondly, the evaluation method used requires re-monitoring of results with each change. This ensures continuity within the program, allowing stakeholders or relevant groups to conduct a transparent information-gathering process about the program. The last, but perhaps the most valuable, strength of this intervention plan is its contribution to social justice. Undoubtedly, bias prevents various cultures from fully realizing their potentials, which in turn disrupts social balance, justice, and equality. Consequently, the importance of this plan in reducing bias for future artificial intelligence systems and, consequently, enhancing social justice is critically significant.

On the other hand, the program has weaknesses as well as strengths. The first is the unknown bias-free nature of the comparison datasets. As known, artificial intelligence models learn from the databases presented to them. At this point, whether unbiased data required to train the model exists, and if so, to what extent it is affected by bias, is unclear. Additionally, the uncertainty of how much the dataset used to measure the accuracy of the evaluation plan is affected by bias remains. A second disadvantage can be the cost. Running and testing trained artificial intelligence models and the prolonged collaboration of all stakeholders are among some of the significant financial expenses in the program. Finally, the time required to monitor every change and interpret its results is much greater than conducting a classic pretest-posttest evaluation program. Therefore, this program requires the definitive participation of stakeholders for at least a year. Additionally, the time cost of restarting the program after making a correction or completing a deficiency is very high.

In conclusion, while this evaluation program proposed to reduce biases in the recruitment process of artificial intelligence systems appears strong in its comprehensiveness, improving the social justice and ability to monitor instantaneous changes, aspects such as time, cost, and the uncontrolled bias of training and comparison data are weaknesses that need to be considered. Lastly, it should be noted that this program was developed specifically for LinkedIn's AI Hiring Framework so this intervention program, if used at another company, must be modified in order to fit the specific context and culture of that company.

References

2022. Microsoft Responsible AI Standard. (2022). https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl

Bakalar et al. (2021, March 24). *Fairness On The Ground: Applying Algorithmic Fairness Approaches To Production Systems*. arXiv.org. https://arxiv.org/pdf/2103.06172.pdf

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. The MIT Press. October 19, 2023, https://fairmlbook.org/pdf/fairmlbook.pdf

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Beutel et al. (2019, Jan 14). *Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements*. arXiv.org. https://arxiv.org/pdf/1901.04562.pdf

Dastin, J. (2018, October 10). *Amazon scraps secret AI retruitig tool that showed bias against women*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Genpact. (n.d.). *Creativity and constraints: A framework for responsible generative AI*. LinkedIn. https://www.linkedin.com/pulse/creativity-constraints-framework-responsible-generative-ai-genpact?trk=organization_guest_main-feed-card_feed-article-content

Quiñonero-Candela, J., Wu, Y., Hsu, B., Jain, S., Ramos, J., Adams, J., Hallman, R., & Basu, K. (2023, May 30). *Disentangling and operationalizing Ai fairness at linkedin*. arXiv.org. https://arxiv.org/abs/2306.00025

Appendix A.

I.) Tabular version of Logic Model for AI Fairness Intervention.

	Input 1: Principle #1: "We will measure and work to mitigate algorithmic bias so that our AI systems treat everyone equally"	**Input 2: Principle #2:** "We will not consider equal AI treatment the end of our work but will treat it as the foundation of a broader fairness and equity strategy"	**Input 3: Principle #3:** "We will validate our approach externally and lead with transparency in this developing field"	**Output**	**Output**
Phase 1	**Activity 1** **Algorithmic auditing of historical data:** evaluation team will perform audits of the dataset the company currently uses so that all data can be evaluated in terms of the algorithms and AI systems.	**Activity 1** a **root-cause analysis of predictive parity:** mathematically assess the predictive parity of two equally qualified candidates using current technologies; control / do not control for demographic data; factor in demographic data / use it as a moderator.	**Activity 1** **Implement debiasing algorithms:** This will generate new data that we will use to determine if predictive parity gaps are widening or getting smaller	**Phase 1 output:** Collection of quantitative data at the beginning stages through predictive parity assessment of current software, as well as algorithmic auditing (which will be performed by the evaluation team's experts)	**Phase 1 output:** Develop justifiable frameworks for when and when not to include demographic information as variables in AI hiring algorithms

Phase 2	Activity 2 **Develop mitigation strategies using the metrics developed through the implementation of debiasing algorithms** for reducing / closing gaps in predictive parity. **Explore/Exploit Strategy is one mitigation strategy** to be initiated by evaluators in order for stakeholders to see long-term effects retrospectively. This strategy includes purposefully dedicating a prescribed "budget" for exploration of candidates who may be qualified but (on our end) lack the available data to know for sure; the identified qualified candidates are automatically processed into the algorithm so that their information is translated as an equally qualified candidate.	Activity 2 **Inclusive and Diverse Model Metrics: Explore broader equity strategies** in order to address equitable outcomes. Equitable outcomes pertain to contexts well past just the scope of algorithmic fairness and design;for example, a specific demographic group's under-representation in recruiter search results or connection recommendation. We first aim to understand if under-representation stems from a real-world structural inequality.	Activity 2 **Continuous monitoring of bias:** Learn and adopt several new innovative tools for addressing privacy and security for applicant data containing demographic information. Learn about three new innovative tools being developed by Linkedin teams on the initiative: 1) Privacy-preserving machine-learning, 2) homomorphic encryption, and 3) A/B testing under differential privacy	**Phase 2 output:** Three new strategies are developed: 1) Mitigation, 2) Explore / exploit, 3) Broader equity All three strategies have the common goal of inclusion. Broader equity strategies allow stakeholders to work bi-directionally in achieving AI equality.	**Phase 2 output:** Exploring and developing mitigation strategies to understand where, when, why, and to what extent using demographic data as variables can help mitigate unfairness in AI algorithms..
Phase 3	Activity 3 **Transparency: Publish / Share/ Document overall findings;** this will look different for each company, but it is important that findings on areas of AI unfairness are described in written and oral manner, as well as how these predictive parity gaps were mitigated using specific strategies.	Activity 3 **Incorporate new results into fairness-aware machine learning practices:** continuous monitoring for bias includes the stakeholders incorporate new algorithms into the model that reduce predictive parity	Activity 3 **External auditing** – the purpose of publishing results from this project will also be for external stakeholders to audit them (test their significance and use). Transparency in the field of AI is crucial so that common methods and practices are implemented at a large scale.	**Phase 3 output #1:** Qualitative data on the current context of the workplace, such as how fairness and equity are manifested through company goals, organization, morals, etc.	**Phase 3 output #2:** It's important that we, as evaluators, are educating these stakeholders along the way; teaching them new, innovative tools being created in tandem will allow them to be more progressive and accurate in their AI design and implementation

Outcomes / short-term goal	**Outcome** (from input 1's 3x activities) Identify predictive parity gaps in current AI / algorithms using for recruiting and hiring; develop mitigation strategies in order to use demographic data without increasing overall risk of perpetuating biases	**Outcome** (from input 2's 3x activities) Distinguishing between equal AI and equitable outcomes will allow company X to develop strategies that are more inclusive, forward focused, and overall more effective at understanding how socio-historical and cultural contexts can be addressed in ways other than just AI / algorithmic design. Overall, the outcome that is most important here is understanding that just simply ameliorating design flaws in an algorithm doesn't guarantee more equitable results.	**Outcome** (from input 3's x 3 activities) Quantitative and qualitative data were collected and analyzed; Published / documented / shared coverage on the specific results of the intervention .		

II.) List of mitigation strategies

Mitigation Strategy Name	**Purpose / process**	**Connects with ___ activity listed above**
Re-evaluate and Retrain the model	Augment the data to ensure better representation of all demographic groups	–Continuous monitoring for bias
Feature Engineering and Fair Representations	Removing or adding demographic data as variables in the model: when do demographic factors make a difference?	Root-cause analysis of predictive parity gaps
Adjust Model Complexity	Simplifying the model architecture or introducing new regularization techniques to prevent overfitting or underfitting	Debiasing algorithms, specifically
Fairness-aware regularization	Include fairness constraints into the model training process to explicitly penalize disparate impacts on different demographic groups	Fairness aware machine learning
Consideration of domain-specific factors	Understand the specific factors within the application domain that contribute to predictive parity gaps; adjust the model accordingly	Diverse and Inclusive Model Metrics

Transparency and Interpretability	Enhance the transparency of the model by incorporating explainability techniques; this helps stakeholders understand how the model makes decisions which can be crucial for addressing and mitigating biases	- External auditing
Stakeholder Involvement	Involve diverse stakeholders including members from underrepresented groups in the model development and evaluation processes. Solicit feedback and insights to improve fairness	- User feedback and perception

III.) **List of Debiasing Algorithms:** several methods/approaches for this

Debiasing Algorithms: Method name	**Purpose / function**
Predictive parity / demographic parity	Adjusting model parameters to achieve parity in predicted outcomes across different groups
Adversarial Training	Teaches the algorithm to be more robust and not rely on biased patterns
Reweighting Loss Function	Implement post-processing techniques to re-rank or re-weight predictions to reduce biases and enhance fairness
Equalized odds	Focus on equalizing true positive rates across demographic groups especially in contexts where equal error rates are desirable
Individual fairness	What is one similarity they can be compared across?
Calibration	Adjust model predicts to align with the true distribution of outcomes; understand that finding an objective means to an end is not rewarding; understand that the end is also of equal importance

IV.) **Future Directions for Data Collection and Analysis**

*** The purpose of this case study is to specifically evaluate LinkedIn's AI Fairness Framework, therefore we did not want to incorporate qualitative measures, such as operationalizing employees' definitions of fairness or the overall ethical considerations and frameworks currently in place at company X. Nevertheless, us choosing not to incorporate them

does not mean these are not important and valuable sources of data when discussing bias in terms of how it affects demographic groups. Therefore, we have provided a future directions section with ideas for qualitative data collection methods that could be incorporated in future research.

Two possible questions to answer in future research utilizing this framework / intervention model

1. Question: Company X's AI framework compared to Linkedin AI Fairness Framework: Do their current ethical guidelines and standards for AI in hiring align / deviate from the Linkedin framework?

2. Question: Investigate how people at company X define "fairness" and "equity": How do we operationalize the stakeholders' definitions of fairness and equity?

Qualitative methods for data collection: An example

Method	Instruments	Measures / Indicators	Data source:
Exploratory qualitative approaches - Focus groups and semi-structured interviews	A new "instrument" could be developed at the preliminary time point using EFA	Thematic analysis; the items that become codes that become themes that become variables that become scales;	Focus groups and semi-structured interviews Recorded / transcribed audio data

1. In other words, at pre-intervention we could conduct a small "pilot study" analysis in order to understand how fairness and equity are understood company-wide. We pick a random sample of employees in order to identify what common themes appear when asked to speak on perceptions, conceptualizations, and real experiences of fairness and equity
2. Pretest items with another small sample using exploratory factor analysis; the point of this is to develop some sort of small questionnaire with two subscales (perceptions of equity and fairness at company X)
3. Administer large sample and conduct confirmatory factor analysis