Investigating Gender Equity, Diversity, and Inclusion at Black Saber Software

An analysis of gender bias in hiring, promotions, and salary

Report prepared for Black Saber Software by Hamiltonian Path Consulting Co.

2021-04-28

Contents

Executive summary **3**
Background & Aim . . . 3
Key findings . . . 3
Limitations . . . 4

Technical report **5**
Introduction . . . 5
Are the company's hiring decisions biased on the basis of gender? . . . 6
Are promotions processes fair, and based on talent and value-add to the company? . . . 20
Do employee salaries differ on the basis of gender? . . . 25
Discussion . . . 32

Future Work **34**

Consultant information **35**
Consultant profiles . . . 35
Code of ethical conduct . . . 35

Executive summary

Background & Aim

Systematic gender discrimination, specifically against women, remains one area in which many companies still struggle with. This discrimination can manifest in many ways, including in the form of biases in hiring, promotion tracks, or salary. Nonetheless, gender bias is antithetical to the values of equity, diversity, and inclusivity that Black Saber Software has dedicated itself to uphold. In order to ensure that the company's internal processes are fair and meritocratic, we examine Black Saber's current employee and hiring data to find evidence of gender bias within the various hiring, promotion, and payment pipelines at Black Saber Software. We take a holistic and statistically rigorous approach in order to make conclusions about any biases within the company's practices.

This report seeks to resolve the questions of whether gender bias against women exists in the various stages of Black Saber Software's hiring pipeline, whether gender is a significant factor in decisions to promote employees internally, and whether gender is relevant to salary decisions.

Key findings

We find no conclusive evidence of any particular gender bias. However our statistical tests show evidence of women being scored lower on speaking and leadership phase 2 assessments, and therefore raise potential concerns over algorithmic bias. All genders perform otherwise equally on all other phase assessments, and initial phase 1 applicants are all very comparable across genders. Women applicants from phase 3 are given offers at a lower rate than men ($\approx$ 29% versus $\approx$ 53%). We therefore recommend more investigation over time to collect more data points, which would allow for a more complete picture of possible gender bias.

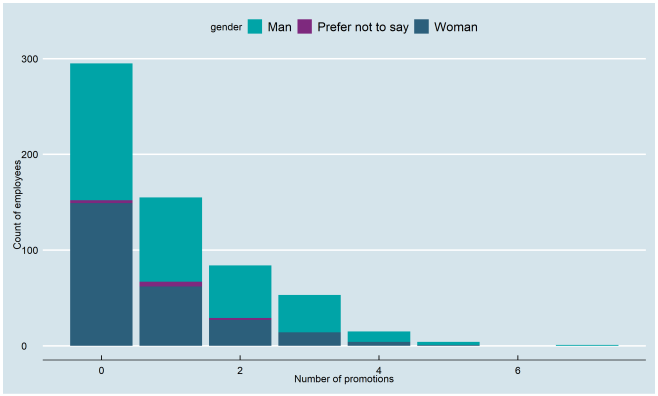

Figure 1: Women's salaries appear to get fewer promotions than their male colleagues.

Promotions appear overall to be well-correlated with meritocratic factors such as leadership, experience, and productivity. An employee's expected number of promotions increases by 4.7% for each time they receive an adequate rating for their leadership evaluation and 7% increase for an exceeding expectation evaluation. However, we also find that the odds of a woman getting a promotion are 37% lower than a man given the same level of performance. This also translates into women being expected to earn 37% of the total promotions than their male counterparts and can be seen in Figure 1 above.

In terms of salary, statistically significant bias against women in general was also found. Women are expected to make $1024 less than their male counterparts and additional bias in the salaries for positions below the managerial level. This discrepancy ranges from between $357 to $1606 and worsens with higher seniority. An overview of this can be seen in Figure 2 below. Worryingly, productivity is a negative factor with each unit of productivity associated with a loss of $14 when determining women's salaries.

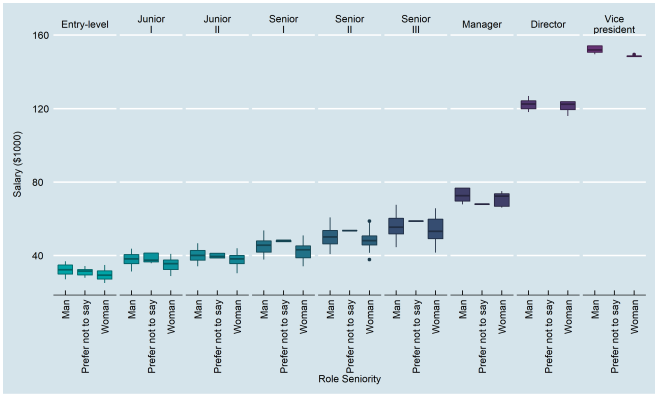

Figure 2: Women's salaries appear consistently lower than their male colleagues for roles of the same seniority.

Limitations

Some limitations of this analysis include that the hiring pipeline is limited by algorithmic bias that can be inherent in what data the AI was trained on. This could lead to biased ratings of candidates that affects the company's ability to make truly merit-based decisions Productivity and leadership metrics are also somewhat subjective measures and therefore subject to implicit bias against women. Selection bias is also present dataset only contains current employees of Black Saber Software and does not include any past employees. This report also mostly excludes non-binary individuals from the analysis as they are not explicitly covered under the "prefer not to say" gender designation

Technical report

Introduction

Supporting equity, diversity, and inclusion initiatives is of paramount importance for companies such as Black Saber that hope to promote healthy and supportive workplaces. Workplaces should also strive to be as meritocratic as possible, making key decisions based on an individual's productivity, leadership, skills, and overall value for their companies. In order to ensure that both EDI initiatives can succeed and merit driven decisions are made requires companies to be retrospective about their pipelines for areas such as their hiring, promotion, and salary. Without careful consideration of these core values, gender, racial, and other inequalities on the basis of protected classes can easily manifest in a company's hiring, promotion, and salary decisions.

Our report covers our analysis of Black Saber Software's hiring, promotion, and salary processes as requested by the company board. In our review of these three processes, we will analyze data across many factors, such as applicant and employee gender, productivity, applicant interview scores, and we will determine if the company's hiring, promotion, and salary processes are merit based. That is, do people get hired based on their talent and potential value to the company, promoted based on the value of work done for the company, and are employee salaries reflective of meritocratic factors like company seniority and the value of their work?

At the same time, we will analyze any potential biases stemming from potentially discriminatory practices based on applicant or employee identity. As we are only given access to the gender, we evaluate any potential biases in an applicant or employee's hireability, promotions, or salary on the basis of their gender. Examining the conclusions of this report holistically will provide a better understanding of Black Saber's position when it comes to supporting EDI. It will also provide insight into which key areas the company should seek to focus its efforts in order to best make Black Saber a more welcoming and meritocratic environment.

Research questions

This report will concern these three main questions regarding Black Saber's hiring, promotion, and payroll pipelines:

- *Are the company's hiring decisions biased on the basis of gender?*
- *Are promotion decisions influenced on the basis of gender?*
- *Do employee salaries differ on the basis of gender?*

Are the company's hiring decisions biased on the basis of gender?

As stated in the introduction, we are looking to review Black Saber's hiring practices to examine if they are meritocratic or potentially sexist (biased against one or more genders). Our approach will be to examine merit based factors for hiring, i.e., factors like GPA, work experience, and interview performance. We will conclude that there is a strong case of gender biased hiring practices in the hiring process if we find that despite similar performances and merit across sexes, one or more particular sex is hired or moves along the hiring process at a far lower rate.

Analysis of phase 1 hiring

First, we will begin our analysis with phase 1 of the hiring process. We will examine several measures of merit, broken down by gender. If these measures of merit are not significantly different between genders, we would expect to find a similar proportional breakdown of gender in phase 2 applicants as phase 1 ($\approx 47\%$ men, $\approx 51\%$ women, $\approx 2\%$ "prefer not to say").

The observed gender within applicants is approximately evenly distributed, with a small proportion of applications responding "prefer not to say" as shown by Figure 3 below:

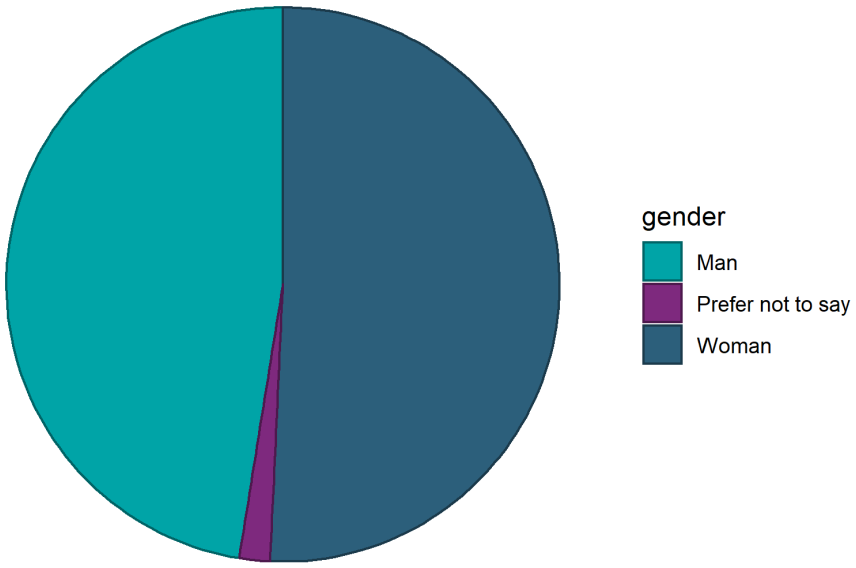

Figure 3: Phase 1 Applicant Genders.

We will then examine the histograms of our meritocratic factors: GPA, extracurricular score,

and work experience in Figure 4.

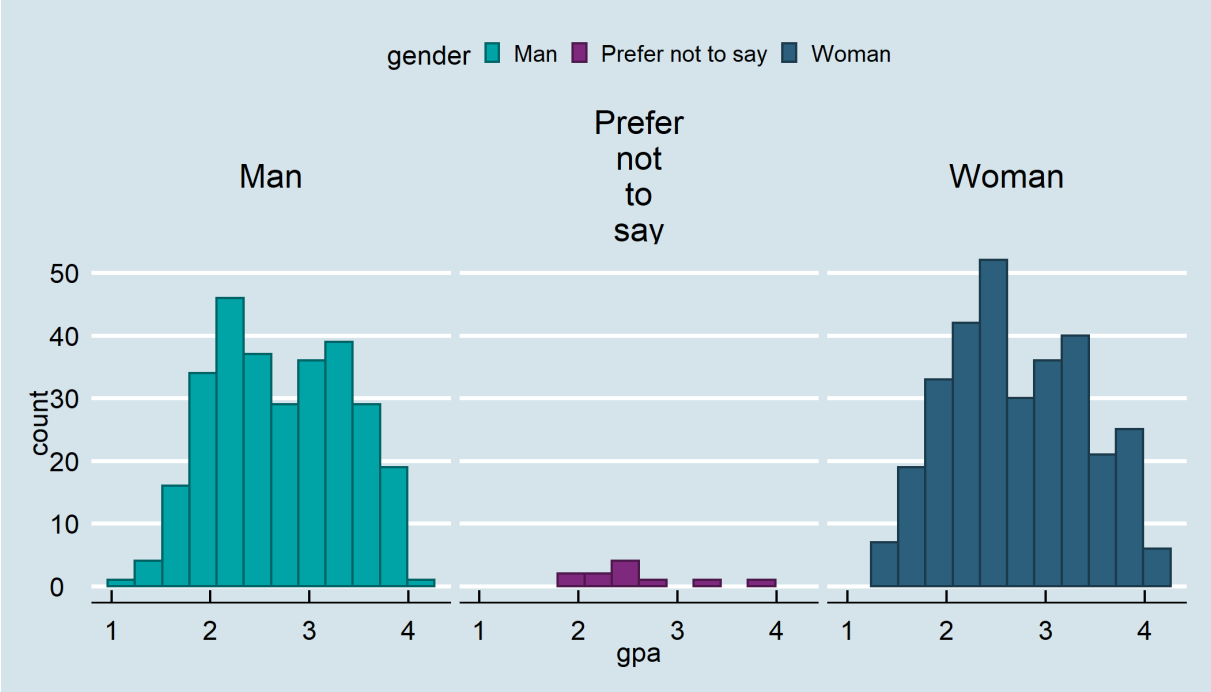

Figure 4: Phase 1 Applicant GPAs by Gender.

Clearly, from the histogram of GPAs split by gender, we find that generally all 3 distributions have most of their densities concentrated around 2 to 3.5. As such there were no signs that any group has a radically different distribution compared to the others. We also conduct an ANOVA test for the means of the three gender groups, and there's no evidence that the means differ at the 0.05 significance level ($\alpha = 0.05$), where the p-value was $0.569 > 0.05$. This is also reflected below in extracurricular scores:

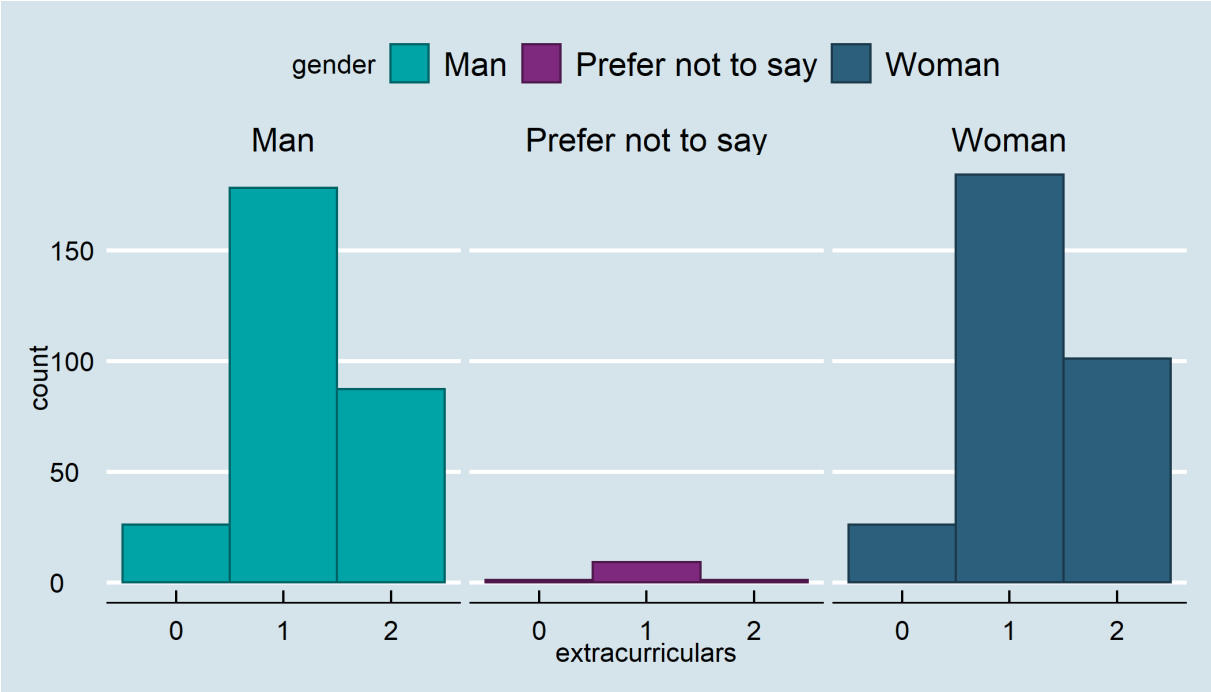

Figure 5: Phase 1 Applicant Extracurricular Scores by Gender.

We find that the distributions of applicant extracurricular scores by gender in Figure 5, look very similar with most applicants having a score of 1, i.e., some relevant/skill building extracurriculars. From our ANOVA test for the means of the extracurricular scores of all three gender groups, we found no evidence at the 0.05 significance level that the means were different, where the p-value was $0.364 > 0.05$.

Similarly by work experience:

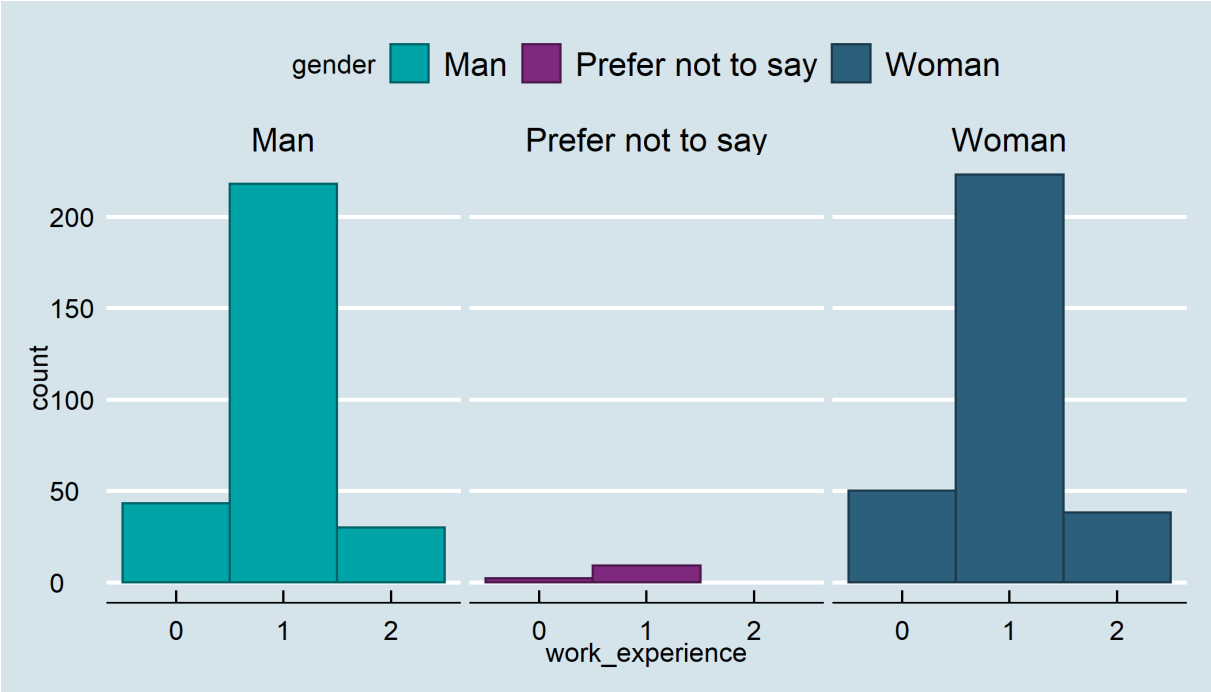

Figure 6: Phase 1 Applicant Work Experience Scores by Gender

Figure 6 above shows that the histogram for all three genders are shaped very similarly, with the most applicants having a work experience score of 1. And as with the previous two metrics, an ANOVA test concludes that no group had a statistically significantly different mean at the 0.05 significance level, where the p-value was $0.663 > 0.05$.

So clearly, we see that applicants, regardless of gender had similar distributions for GPAs and extracurricular and work experience scores. Further, we examine the proportions of applicants that submitted cover letters and CVs by gender. We conduct a test for equal proportions of the three gender groups for both the proportion of cover letters and CVs; that is, we conduct the following test for both the proportion of cover letters and CVs (separately):

$$H_0 : p_{\text{male}} = p_{\text{female}} = p_{\text{Prefer not to say}}$$
$$H_a : \text{the proportion is different in at least one of the gender groups}$$

We find that there was a statistically significant difference for our test of the proportions of CVs submissions, i.e., at least one group had a significant difference. Thus, we further conduct a pairwise test for proportions using the Bonferroni correction for multiple comparisons, which

ultimately showed no significant difference between any two groups at the 0.05 significance level. However, the closest p value to our significance level is the difference between male and "prefer not to say", at 0.084, which is close to 0.05, so possibly suggestive of a small true difference. We found no evidence of a significant difference in proportions between genders groups for the proportion of cover letter submissions.

Holistically, we find that the profiles of applicants across all three gender groups were comparable, with no direct evidence that any gender cateogry is clearly different than any other on any measure of merit in phase 1. It follows that we should find the gender breakdown in phase 2 to be similar to that of phase 1, which was indeed the case as indicated by our pie chart in the next subsection.

Analysis of phase 2 hiring

The gender demographics of applicants who made it to phase 2 were similar to phase 1, with male and female applicants mostly evenly split and a small component of "prefer not to say" as indicated by the pie chart ($\approx 47\%$ men, $\approx 51\%$ women, $\approx 2\%$ "prefer not to say"):

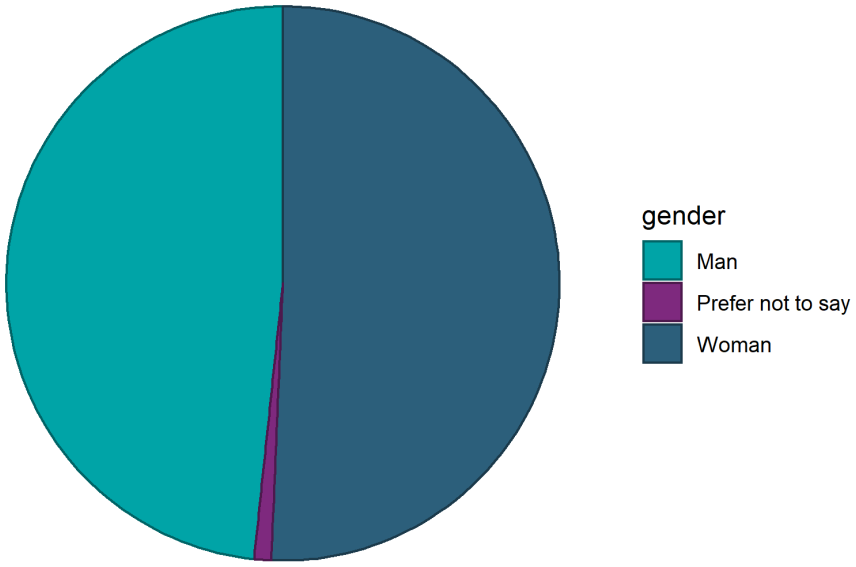

Figure 7: Phase 2 Applicant Genders.

In phase 2, the hiring processes considered AI autograded tasks on technical coding skills, writing skills, speaking skills, and leadership presence. Like the previous sections, we examined

the histograms of these factors:

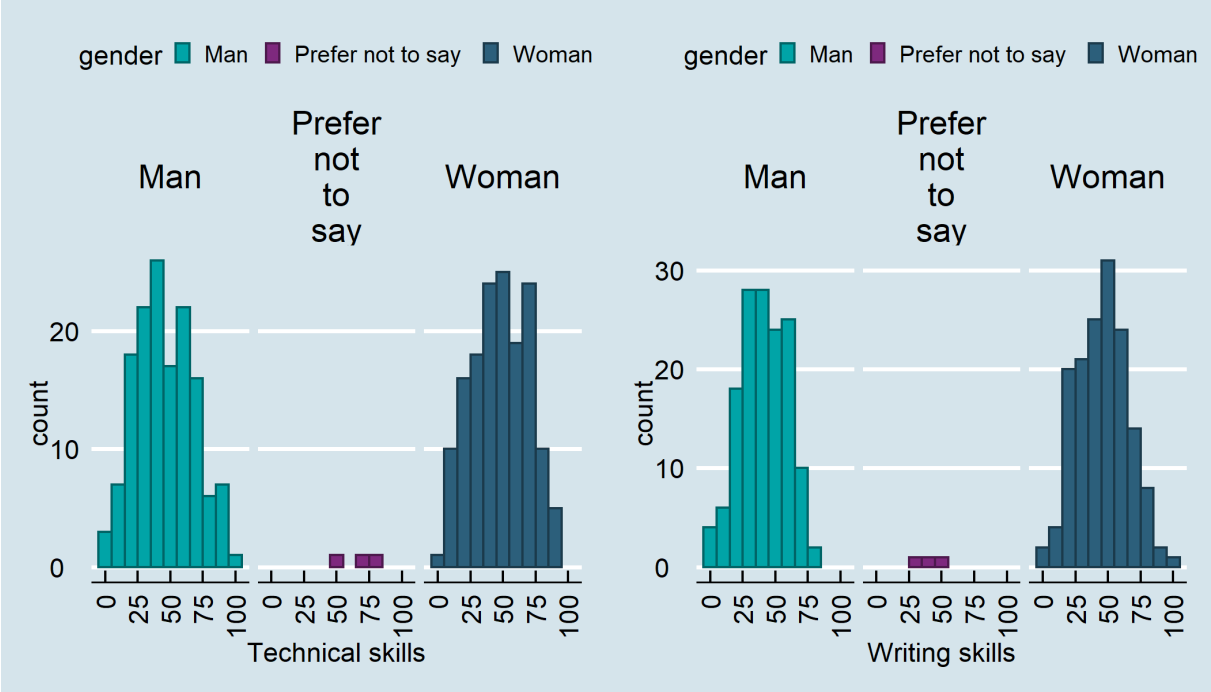

Figure 8: Round 2 applicant technical and writing skills by gender.

In phase 2, the technical and writing abilities across genders were distributed very similarly, showing no indication that any particular gender was allowed without merit into phase 2 in terms of these two skills. To further confirm, we performed an ANOVA test for both metrics, finding no statistical evidence that one gender outperformed the others.

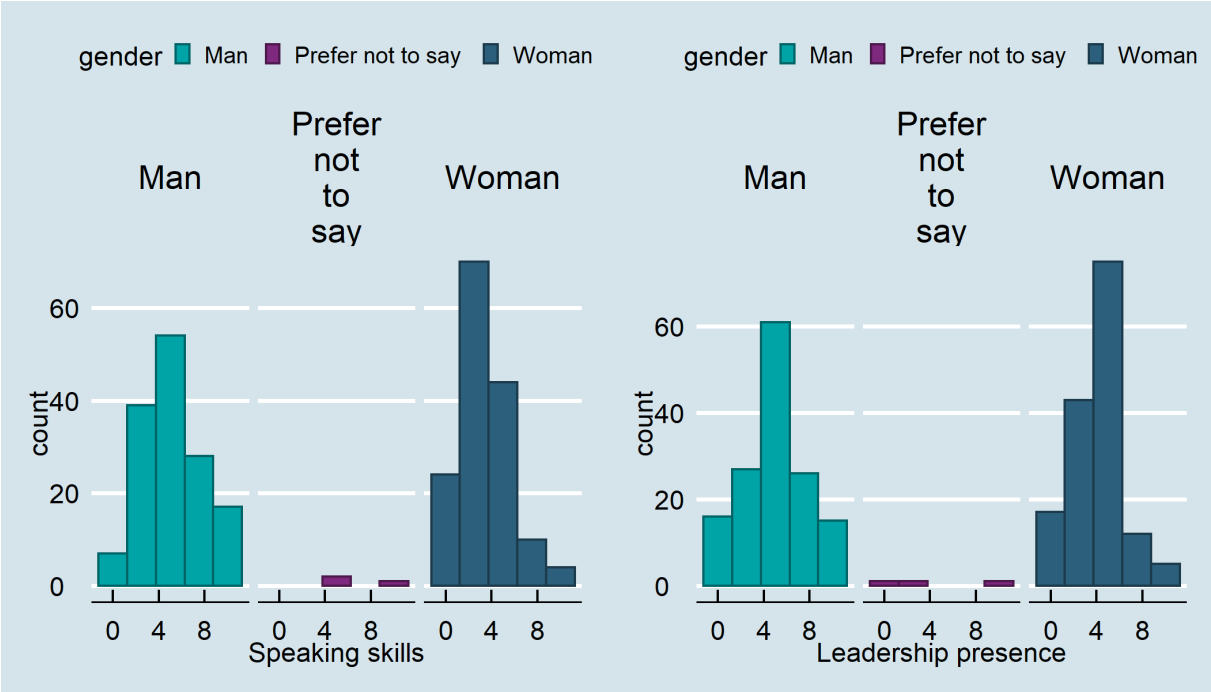

Figure 9: Round 2 applicant speaking and leadership skills by gender.

From Figures 7 to 9, we see that much like phase 1, the distributions across sexes were quite similar, but women were given somewhat lower leadership presence and speaking scores. In particular, in both of these metrics, we find the distribution for women to have a very noticeable right skew, especially in comparison to men. Given the nature of AI based technologies, certain non-meritocratic features of datasets may be used by AI algorithms such as deep neural networks that act as a proxy for meritocratic features like leadership presence. For example, training an AI model to detect leadership presence on male subjects may harm generalization performance on female applicant data as the model may associate male subjects with leadership. Since women were generally given lower scores for leadership presence and speaking scores, we want to evaluate if any potential gender bias was present due to a biased AI system, especially because given that all other measures of merit were similarly distribution across genders.

To investigate our suspicion of algorithmic bias, we further conducted an ANOVA test finding statistical evidence at the 0.05 significance level that there was evidence that all the speaking skills score means across genders were equal ($p = 6.86 \cdot 10^{-11} < 0.05$). We then conduct an additional pairwise t test using the Bonferroni correction for multiple testing, which concluded statistically significant evidence at the 0.05 significance level of the means of the speaking skills score of women was less than that of men, where $p = 6.32 \cdot 10^{-11} < 0.05$. We found no statistical evidence of the a difference between men and "prefer not to say" ($p = 1 > 0.05$), but note that

the sample size was 3 in this case, meaning it is unlikely for us to get an accurate assessment of the true difference if there is one.

We then repeat the process with leadership presence is similarly suspect for women as speaking skills. Using an ANOVA test, we find that the means, like speaking skills, shows statistical evidence that they were not equal across genders with a p-value of 0.007 at the 0.05 significance level. Then using, an additional pairwise t test with the Bonferroni correction for multiple testing, we find additional evidence that women specifically had a lower mean leadership presence score than that of men with a p-value of 0.002 at the 0.05 significance level. Again, while "prefer not to say" had no evidence of a lower mean leadership presence score than that of men, we can't get the full picture with only a sample size of 3. Clearly, we have substantial evidence that women have been assigned lower scores with respect to speaking skills and leadership presence than men. This compels us to investigate further.

To do this, we constructed a generalized linear mixed effects model (GLMM) with the potentially suspect assessments (leadership presence and speaking scores) given an additional random slope of gender, and with a random intercept for gender. Further, our response variable was a dummy variable with levels 0 and 1, indicating if the applicant moved onto phase 3 or not, and we also included the other phase 2 assessments (the writing task and technical coding assessment) as fixed effects. We specifically wanted to model the log odds (probabilities) of an applicant making it to phase 3, while taking into account all assessments performed in phase 2, so naturally, we used a logit link. To emphasize, our key intuition for this model is that we want to investigate the effect of gender on the odds of an applicant from phase 2 moving on to phase 3, specifically overall (the intercept) and how the suspect assessments are "weighted" possibly differently for certain genders (the predictor coefficients with random slopes). Mathematically, that is:

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit}(p_{ij}) = \beta_0 &+ \beta_1 \cdot \text{technical skills}_i + \beta_2 \cdot \text{writing skills}_i + \beta_3 \cdot \text{technical skills}_{ij} \\
&+(\beta_4 + U_{1j}) \cdot \text{speaking skills}_{ij} + (\beta_5 + U_{2j}) \cdot \text{leadership presence}_{ij} + U_{3j} + \epsilon_{ij} \\
U_1 &\sim N(0, \sigma_1^2) \\
U_2 &\sim N(0, \sigma_2^2) \\
U_3 &\sim N(0, \sigma_3^2)
\end{aligned}
$$

Note that above we are modelling the i^{th} individual with gender j. It follows that U_1, U_2, U_3 are the respective distributions from which U_{1j}, U_{2j}, U_{3j}, the random slope for speaking skills corresponding to gender j, the random slope for leadership presence corresponding to gender j, and the random intercept for gender j are drawn from. Also note that Y_{ij} is therefore the dummy variable measuring if the the i^{th} individual with gender j moved on from phase 2 to

phase 3.

The model fit indicates that there was no difference in the coefficients across genders; that is, all genders had the same exact coefficients to three decimal places:

Table 1: Phase 2 model summary.

	(Intercept)	technical	writing	speaking	leadership
Man	-20.673	0.078	0.088	0.7	0.928
Prefer not to say	-20.673	0.078	0.088	0.7	0.928
Woman	-20.673	0.078	0.088	0.7	0.928

So we find that there's no indication provided by our model that any gender had their phase 2 assessments (e.g., the timed technical coding challenge) scores "weighted" higher or lower, and we have no indication that any gender has overall lower or higher log-odds (and thus probabilities) of moving on to phase 3 as our intercepts were the same across genders.

Provided that the distributions of leadership presence and speaking skills across genders are visually different and have statistically different means, we investigate further with another GLM that asks a similar but different question: does gender have an effect on the probability of moving on to phase 3 while controlling for phase 2 assessments?

To answer this question, we use a logistic model with the response variable again being the dummy variable for if an applicant in phase 2 moved on to phase 3. The covariates of the model are the phase 2 assessment scores and gender. Mathematically, that is:

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + \beta_1 \cdot \text{technical skills}_i + \beta_2 \cdot \text{writing skills}_i + \beta_3 \cdot \text{technical skills}_i \\ + \beta_4 \cdot \text{speaking skills}_i + \beta_5 \cdot \text{leadership presence}_i + \epsilon_i\end{aligned}$$

Note that above we are modelling the i^{th} individual, so it follows that Y_i is the dummy variable for if the i^{th} individual moved on from phase 2 to phase 3.

From our model fit, we find that the coefficient for gender was statistically insignificant for "prefer not to say" and women with all p-values above 0.05, meaning that we do not have evidence from our model that gender affected the probabilities of someone who identified as "prefer not to say" or a woman from moving on to phase 3.

So overall from our models, we find no evidence of gender bias. Specifically, we find no evidence that a particular gender has a lower probability of moving on to phase 3 overall, no evidence that a particular phase 2 assessment was "weighted" higher or lower for a particular gender, and no evidence of gender having an effect on the probabilities of moving on to phase 3 particularly for applicants who identified as women or "prefer not to say".

While we find that there's no evidence from our models of gender bias, we caution that our model results are not conclusive of no gender bias in phase 2. The graphical analysis we performed showed a pretty noticeable right skew in the distributions of leadership presence and speaking skills of women in comparison to that of men. Our graphical analysis was further backed up by statistical tests that concluded different mean scores. While it may well be the case that women do truly perform worse in speaking and leadership, conventional logic would question how applicants that otherwise perform very similarly differ so much on just these two assessments? Further, as stated earlier, scores in phase 2 were given by an AI process, of which may be trained on gender imbalanced data. Such a process leads to proxies for gender and ultimately poor generalization performance on genders other than men, who historically are more represented in corporate structures and by extension have more data training models measuring leadership presence. It would be sensical to review the AI processes that generate these assessment scores. Finally, we caution that phase 2 may be subject to gender biases because considering what we've already stated, the demographics of phase 3 are suspect. In that phase 3 comprises of $\approx 32\%$ women and $\approx 68\%$ men, and while such a demographic may simply be due to variation of factors other than gender, especially given our model results, we find that women going from representing 50% of applicants in phase 2 to only $\approx 32\%$ in phase 3 to be definitely strange as applicants irrespective of gender perform similarly. We will analyze phase 3 in conjunction with final hires in the next section.

Analysis of phase 3 hiring and hiring results

As stated in the previous paragraph, our demographics in phase 3 look quite different from the previous two phases ($\approx 32\%$ women and $\approx 68\%$ men). It's also represented graphically as:

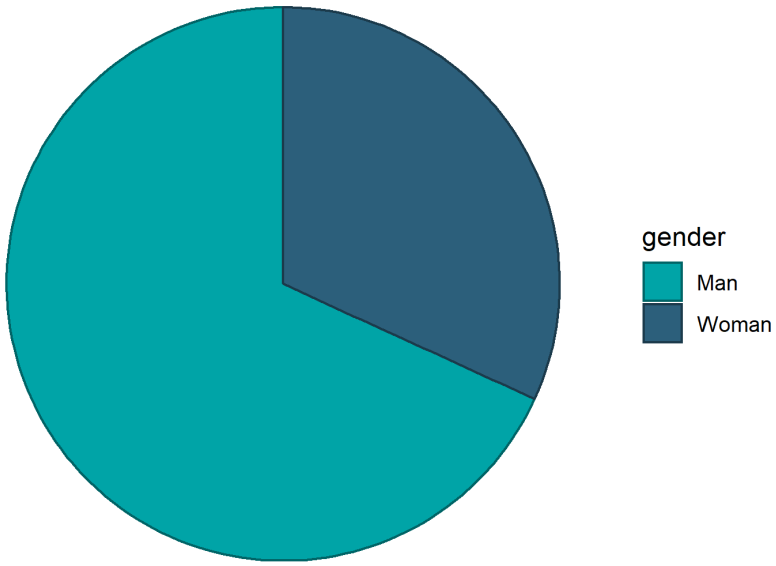

Figure 10: The proportion of genders that moved on to round 3.

As with the previous hiring phases, we begin our analysis by examining the distributions of meritocratic factors in this phase, broken down by gender. We see the following:

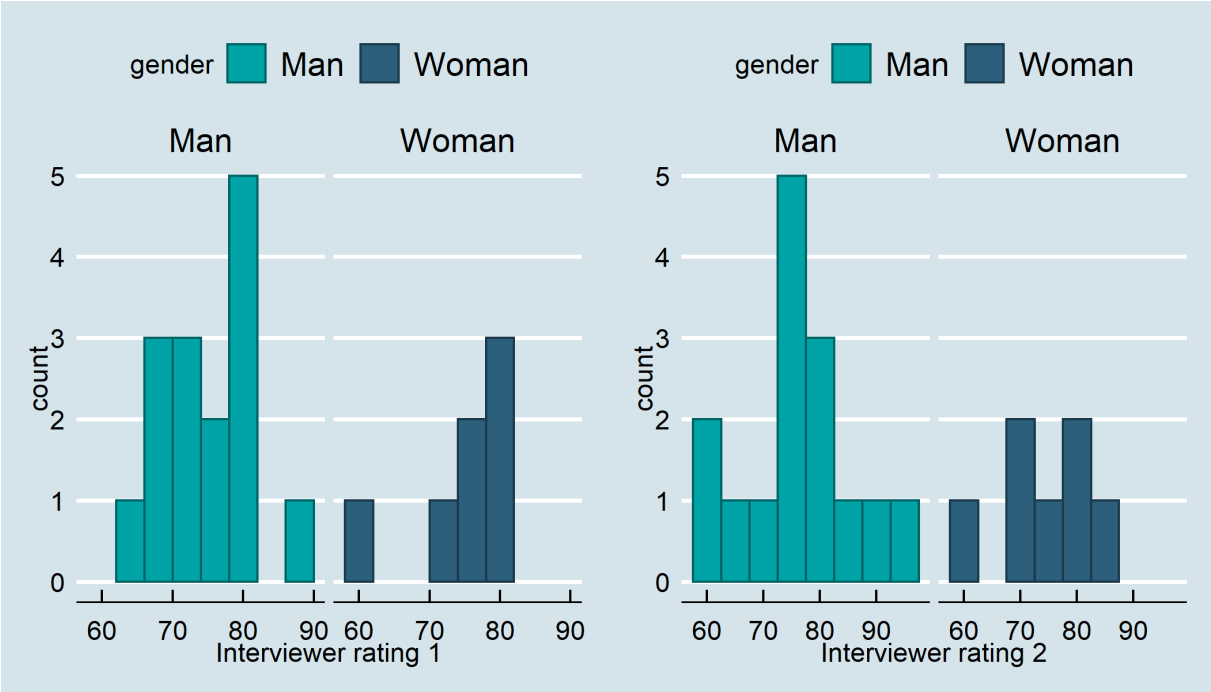

Figure 11: Round 3 interviewer 1 and 2 ratings by gender for all candidates.

We see in Figure 11 that the distributions of the two interviewer scores across are fairly similar in that the range of scores for each gender given an interviewer is consistent for the most part. In particular, most scores given by the first interviewer for both genders were around 80, but men did get some higher scores, i.e., one male applicant got a score above 80. Additionally, most scores given out by the second interviewer were concentrated around 70 to around 90. So on face, it seems that regardless of gender applicants performed similarly in phase 3 interviews.

To further explore this, we wanted to test the mean scores of the two distributions for both interviewers. To do this we performed a two sample t-test with the null hypothesis being that the means are equal and the alternative being that the means are not equal, and we used the 0.05 significance level. We found no evidence of a difference in the means of the distributions of the interviewer 1 scores between men and women as $p = 0.73 > 0.05$, and we also found no evidence of a difference in the means of the distributions of interviewer 2 scores between men and women because $p = 0.93 > 0.05$. So we have evidence that on average men and women performed similarly, providing credence to our visual observations.

Thus, it would follow that we would expect to see a similar gender demographics of those hired. We see that the gender demographics of those hired are 80% men and 20% women, i.e., 8 men hired and 2 women hired. The proportion of men versus women are quite different from those in phase 3, but we do only have a sample size of 10 in this phase in comparison to 22 applicants

in phase 3, so analysis here may not provide the full picture due small sample sizes and natural variability. Nonetheless, since hiring decisions are based off of phase 3 interview scores, we examine the distribution of scores of those hired by gender:

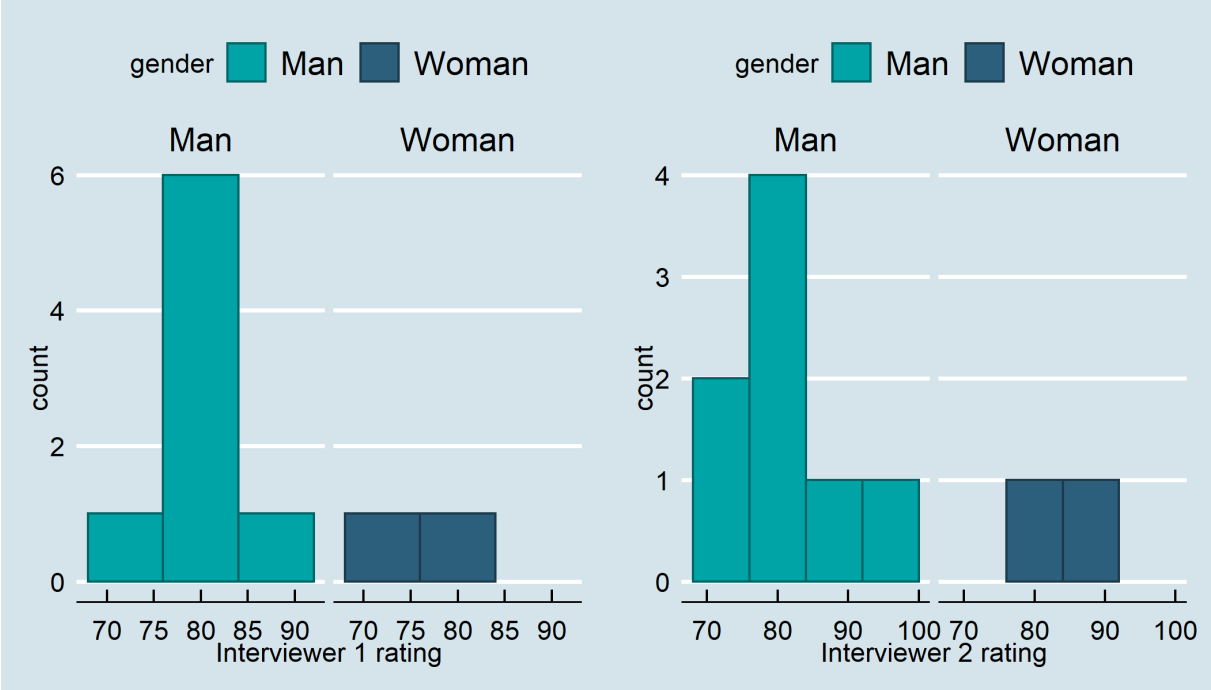

Figure 12: Round 3 interviewer 1 and 2 ratings by gender given that the candidate was hired.

From looking at the distributions of those hired in Figure 12, we see that all hired applicants, irrespective of gender, were those who scored well (above 70) on the phase 3 interviews. Since only 10 people were hired and only 2 were women, we did not perform any statistical tests over concerns of the power of tests and imbalanced sample sizes. In particular, any sample statistics we calculate may drastically change simply due to the natural variability of the data and our small sample size. Instead, we examine the success rates between phase 3 applicants and those hired by gender. We find that the success rate of men is approximately 53% and the success rate of women is approximately 29%; that is, approximately 53% of men in phase 3 were hired while only approximately 29% of women in phase 3 were hired. While such a discrepancy may be due to small sample sizes and natural variation in our data, this is quite a discrepancy as both genders had similar phase 3 scores but were hired at very different rates. It follow then, do we see such a discrepancy possibly due to women scores being "weighted" lower? To answer this question, we use a GLMM with a logit link to model the log-odds of an applicant from phase 3 getting hired (a dummy variable given by either 0 or 1). Our predictors were the two interview scores from phase 3, i.e., the meritocratic factors measured during phase 3. Also we

added gender as a random slope to the two predictors and intercept to examine if women had a lower overall odds of being hired, and if women's scores were being "weighted" less than those of men. Mathematically, we have:

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit}(p_{ij}) = \beta_0 + (\beta_1 + U_{1j}) \cdot \text{Interviewer Rating 1}_{ij} + (\beta_2 + U_{2j}) \cdot \text{Interviewer Rating 2}_{ij} + U_{3j} + \epsilon_{ij} \\
U_1 &\sim N(0, \sigma_1^2) \\
U_2 &\sim N(0, \sigma_2^2) \\
U_3 &\sim N(0, \sigma_3^2)
\end{aligned}
$$

Note that above we are modelling the i^{th} individual with gender j. It follows that U_1, U_2, U_3 are the respective distributions from which U_{1j}, U_{2j}, U_{3j}, the random slope for interviewer rating 1 corresponding to gender j, the random slope for interviewer rating 2 corresponding to gender j, and the random intercept for gender j are drawn from. Also note that Y_{ij} is therefore the dummy variable measuring if the the i^{th} individual with gender j got hired after phase 3.

From the fit model, we find that the coefficients were exactly the same between men and women, meaning we have no evidence from our model that the odds of being hired are different between men and women. So similar to our phase 2 analysis, we have no evidence from our model that there was any gender bias in using the phase 3 scores to determine final offers, but graphically and from the demographics of those hired, it seems potentially suspect that despite having similar interview performances only 2 women were hired while 8 men were hired. Certainly, this may just be due to the nature of small samples sizes for example, an increase in 1 woman hired and a decrease of 1 man hired would net a female hire rate of approximately 43% and a male hire rate of 47%, which seems less imbalanced.

Overall hiring conclusions

We find no conclusive evidence of a particular gender bias for either of the three categories. In particular, phase 1 does not show any worrying signs or evidence of gender bias, and as such the demographics of phase 2 applicants closely resemble those of phase 1. Although phase 2 and 3 show potential concerning signs of gender bias. In phase 2, the averages of women's speaking skills and leadership presence scores were significantly lower than those of men, despite all other scores being similar, which seems somewhat concerning as it would be strange that women are systematically worse than men at speaking or having a "leadership" presence despite performing otherwise similarly. In phase 3, women performed similar to men, but were hired at a concerningly lower rate. So while the models showed no evidence of gender bias, there are

certain aspects of the hiring process that would warrant more investigation, namely if AI systems in phase 2 generalize well to women rather men, which these models are likely to be train on, and if these hiring differences between men and women were truly due to natural variation.

Are promotions processes fair, and based on talent and value-add to the company?

We consider the dataset provided by Black Saber Software's data team that contains data on all current employees for the entire duration of their employment, where each row represents salary, demonstrated leadership and productivity for an employee in a given financial quarter.

We conduct exploratory data analysis where we find that each of the 607 employees who remain employed by Black Saber Software have worked within the same team for the duration of their employment. We note that for each observation in the data, an employee's salary, demonstrated leadership, and productivity are reported for a given quarter. We assume that in a given quarter, the leadership evaluation that an employee receives is from their immediate supervisor which differs per employee depending on their current role and team.

We modify the variables in Black Saber's current employee data that represent role seniority and demonstrated leadership to represent the ordinal structure of the data. We add a binary variable that represents whether an employee is promoted to a more senior role in the following quarter. We add the number of financial quarters that an employee has been employed to date, their total number of promotions, and the number of quarters since their last promotion (quarters in role).

We create a new dataset that aggregates information from Black Saber's current employee data, where each employee is associated with a single observation and contains the employee's gender, team, the total number of promotions they have received throughout the duration of their employment, the total duration of their employment (in financial quarters), their starting role and the financial quarter in which they began employment.

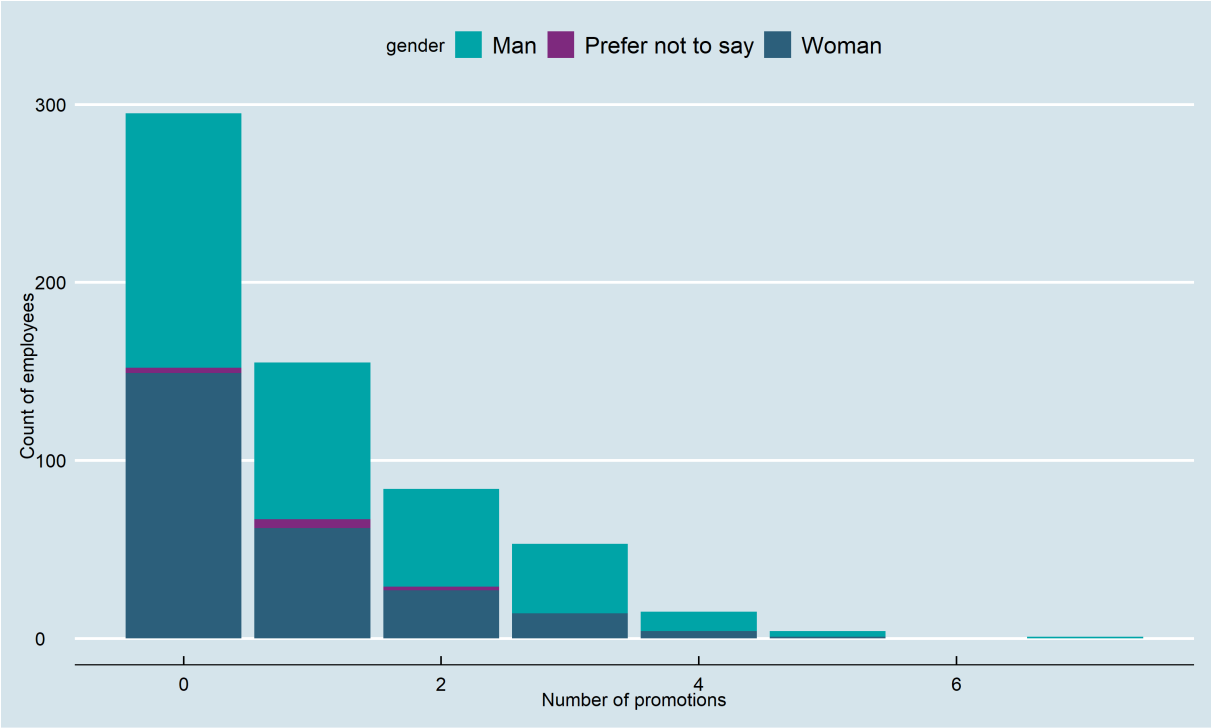

Figure 13: The distribution of promotions by gender.

Figure 13 reveals a fair amount of variability in the number of promotions with a range from 0 to 7. It also evident that the number of promotions is not normally distributed.

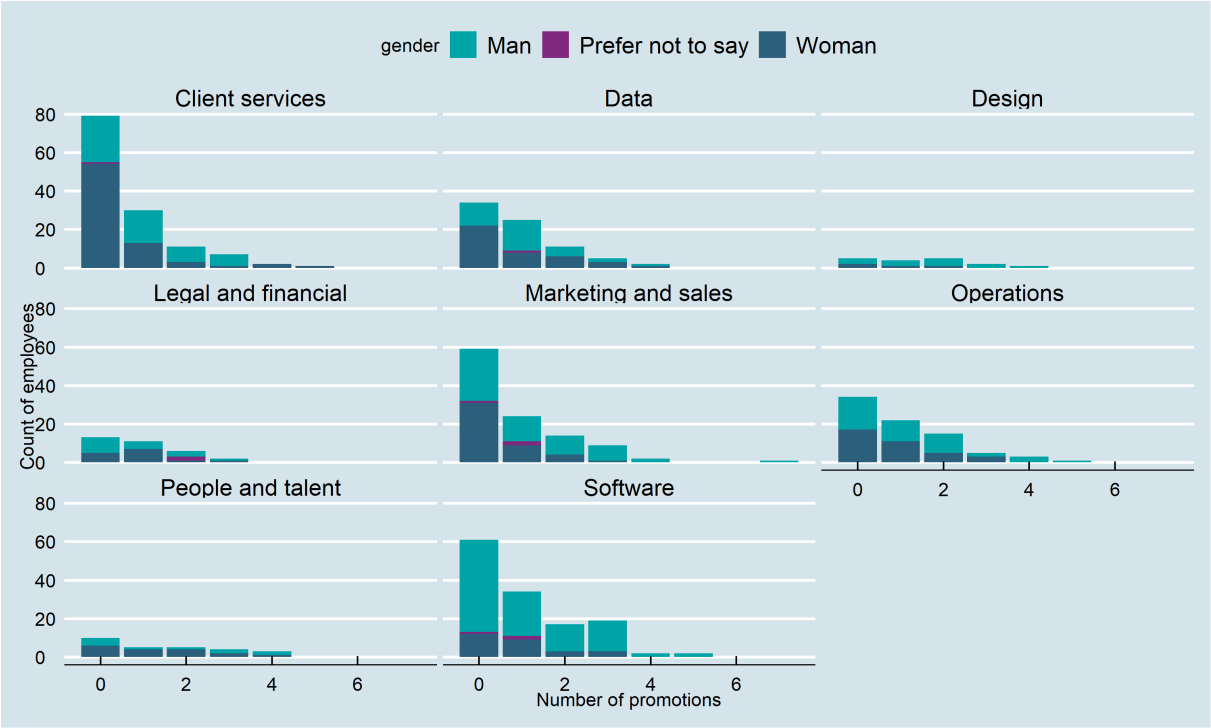

Figure 14: The distribution of company-wide promotions by team and gender.

When analyzing between teams as in Figure 14, we can further conclude that number of promotions can be reasonably modeled with a Poisson distribution for each team.

While a Poisson regression model is a good choice because the responses are counts, it is important to note that the counts are not directly comparable because each individual has been with the company for a different number of fiscal quarters - we expect individuals who have been with Black Saber Software longer to have been awarded more promotions. We can take the difference in number of fiscal quarters into account by including an offset in our model. We also note that the majority of employees have not been promoted. This suggests that a Zero Inflated Poisson model is appropriate to model the latent variable consisting of the group of employees who are unlikely to ever seek promotions as opposed to the group that have a desire to be promoted within the company.

We iteratively construct the model by comparing likelihood ratios to get an appropriately complex and well-fitting model.

Table 2: Zero-inflated poisson model summary, where the coefficient estimates have been exponentiated.

	Estimate	P-value
count: base line odds	0.0328213	0.0000000
count: average productivity	1.0051898	0.2596814
count: appropriate for level	1.0470683	0.0000000
count: exceeds expectations	1.0698693	0.0177248
count: prefer not say	0.9058502	0.7709626
count: woman	0.7305371	0.0024282

We find that when considering up to a 95% level of confidence, that the number of times an individual's leadership is rated appropriate for level or exceeding expectations, as well as being a woman are statistically significant predictors of how many promotions an individual employee is expected to have over their tenure at Black Saber Software. Their average productivity over their time at the company is found to be statistically insignificant.

Each time that an individual is given a leadership ranking that is appropriate for level, we see a corresponding 4.7% increase in the odds of an individual being promoted, and a 7% increase for each exceeding expectation evaluation.

However, we also see that individuals who are open to being promoted during their time at Black Saber Software, the average number of promotions that a woman will be 37% less than a man who has the same track record at the company.

How is an individual to get any given promotion?

In this section, our response variable y_i of interest is (the presence or absence of a promotion in the following financial quarter) whether an employee will be promoted in the following quarter based on talent and value-added to the company. That is, we seek to investigate the relationship between a promotion and an employee's productivity and demonstrated leadership. Our response is our newly created binary response variable that returns 1 if an employee will be promoted, and 0 otherwise. Since our response variable is binomial (i.e. *not* Gaussian) and we have a clear violation of the independence assumption due to the presence of repeated measures of employees, we fit a generalized linear mixed model.

We iteratively add terms and compare the goodness of fits using the linear. In the end, we settle on the following logistic model for the probability that an employee will be promoted for any given quarter. We fit random intercepts for role seniority, and financial quarter to take into account that the rate of promotion. This represents the fact that promotions may be contingent on overall company success in a financial quarter for which we should take into account. The likelihood of promotion is also capped by role seniority as we expect that lower level jobs have more opportunities to climb the ladder as opposed to higher level positions. We also use a random intercept to account for the repeated measures with respect to employees. This leaves leadership for level, productivity, quarters in role (as a measure of experience), and gender as fixed effects for the model.

$$Y_{ijk} \sim \text{Bernoulli}(p_{ijk})$$

$$\begin{aligned}\text{logit}(p_{ijk}) = \beta_0 + \beta_1 \cdot \text{productivity}_{ik} + \beta_2 \cdot \text{gender}_i + \beta_3 \cdot \text{leadership for level}_{ijk} \\ + \beta_4 \cdot \text{quarters in role}_{ijk} + U_{1i} + U_{2j} + U_{3k} + \epsilon_{ijk}\end{aligned}$$

$$\begin{aligned}U_1 &\sim N(0, \sigma_1^2) \\ U_2 &\sim N(0, \sigma_2^2) \\ U_3 &\sim N(0, \sigma_3^2)\end{aligned}$$

Note that above we are modelling the i^{th} individual with role j during financial quarter k. So it follows that U_1, U_2, U_3 are the respective distributions from which U_{1i}, U_{2j}, U_{3k}, the random intercept for employee id of the i^{th} individual, the random intercept for role j, and the random intercept for financial quarter k are drawn from. Also note that Y_{ijk} is the dummy response variable for if this individual received a promotion that quarter.

Table 3: Logistic regression model summary and 95% confidence intervals, where the coefficients and CIs have been exponentiated.

	Estimate	2.5%	97.5%
(Intercept)	0.00	0.0000	0.0000
productivity	1.16	1.1433	1.1710
gender: prefer not say	1.17	0.2682	5.0743
gender: woman	0.64	0.4242	0.9711
evaluation: appropriate for level	2.18	0.4920	9.6216
evaluation: exceeds expectations	1.27	0.5444	2.9715

	Estimate	2.5%	97.5%
quarters in role	1.23	1.1831	1.2805

We find that when considering up to a 95% level of confidence, that productivity, the number of quarters an individual has had in their current role, and being a woman are all statistically significant predictors of whether an individual will get a promotion. Leadership evaluations are found to not be significant in an individual's promotion rate.

We find that the baseline odds of an individual odds of being promoted are close to 0. This suggests that the odds of being promoted are fully based on the other predictors in the model. In particular, an individual's experience and productivity in their role are strong indications of whether or not they will be promoted. For every unit increase in an individual's productivity score, we expect a corresponding 16% increase in their odds of being promoted. Furthermore, we have the number of quarters an individual has spent in their previous role has a corresponding 23% increase in their odds of being promoted.

One worrying finding is that the odds of getting any given promotion as a woman are 36% lower than that of a man with the same level of experience and productivity. In summary, we have reason to believe that the promotion process takes productivity and experience into account, but remains biased against women and fails to take into account an individual's leadership skills. This also manifests in a lower number of promotions for women overall as compared to their colleagues.

Do employee salaries differ on the basis of gender?

For the purposes of this analysis, we will primarily focus on the salary differences self-identified men and women. Non-binary individuals and those who selected to prefer not to answer constitute a small minority of the overall employees of Black Saber and it is difficult to conclude anything due to the reduced sample size and unclear dilineation between those who are non-binary as opposed to those who did not wish to identify. However, where relevant we will draw attention to this group.

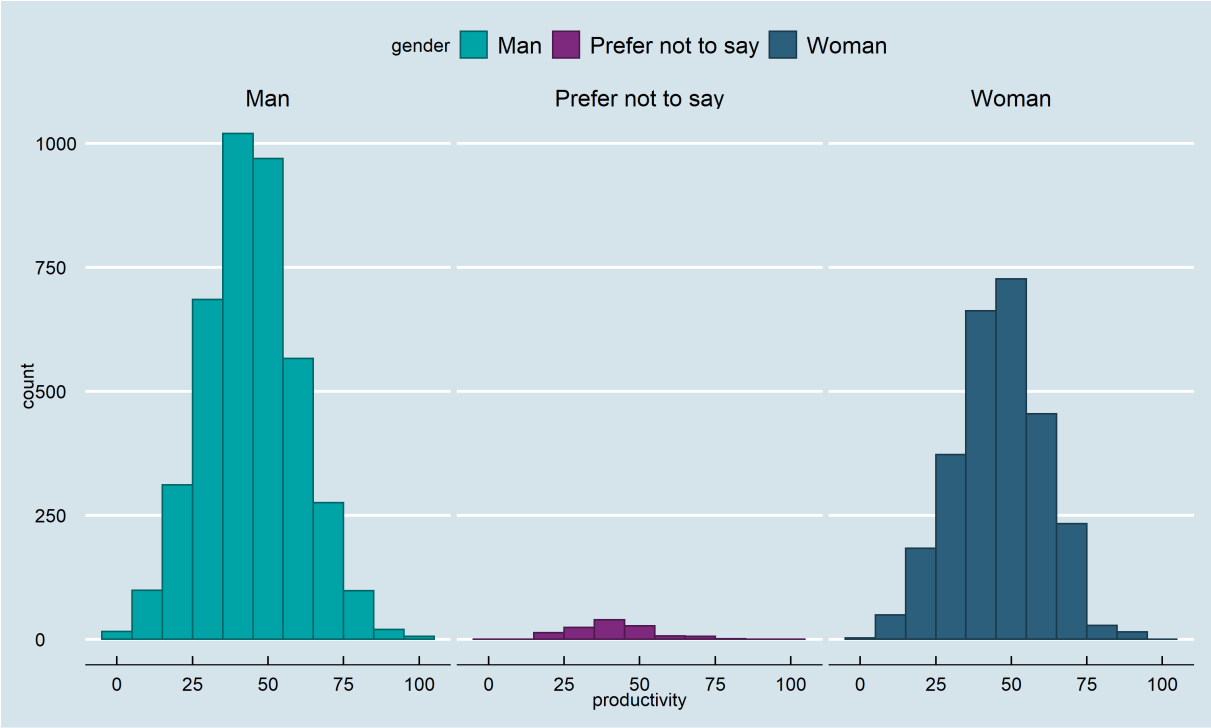

Figure 15: Employee Productivity by Gender.

Our analysis of salary practices at Black Saber begins with a cursory look at the overall employee population and their productivity metrics. When examining Figure 15 we see no evidence that the distribution of productivity between men and women is significantly different. With respect to individuals that preferred not to answer, the distribution is unclear due to smaller sample size. In any case, the three groups have roughly similar but statistically significantly different means with women appearing to have marginally average productivity scores as opposed to men.

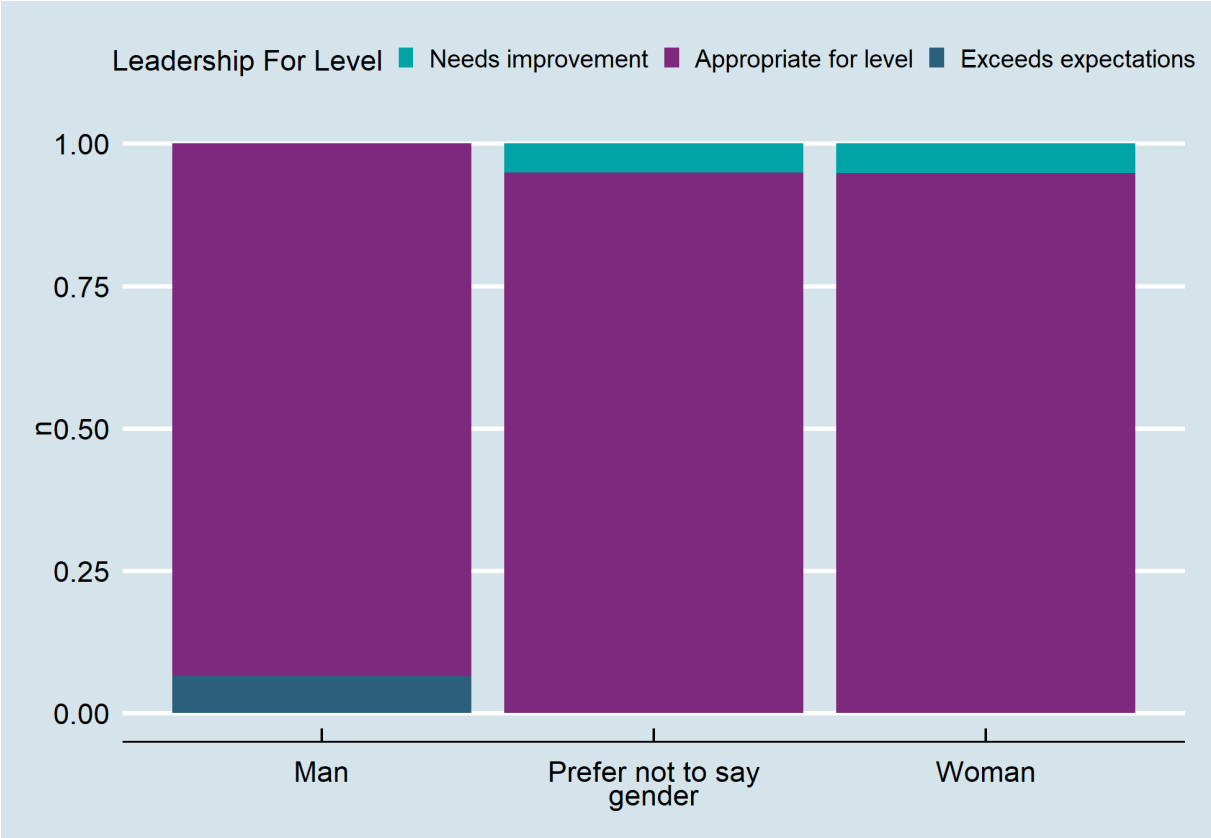

Figure 16: Leadership for Level by Gender.

Looking to Figure 16, we examine the distribution between different levels of leadership. We find that the vast majority of individuals across all gender categories display a level of leadership appropriate for their roles. However, what is concerning is that only men have ever been evaluated to demonstrate exceeding leadership for their level and have never been rated to need improvement. Conversely, women have never been rated to demonstrated exceeding leadership. This gap is suspicious and warrants further analysis.

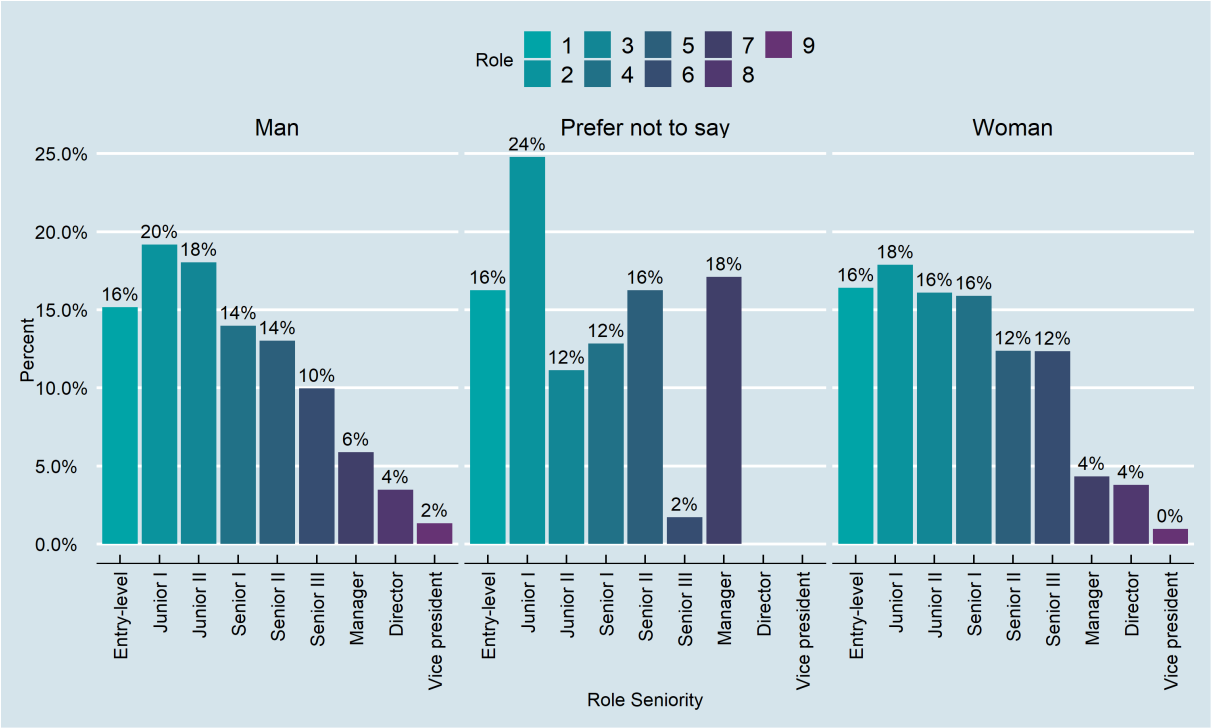

Figure 17: Gender seniority.

When we analyze Figure 17, we find that the distribution of role seniority across men and women is very similar. This suggests that variance in salary between gender groups should not be based on any inherent bias between men and women's roles within a company. The distribution of role seniority for individuals who preferred not to answer is quite skewed from those of self-identified men and women but is also likely a result of the smaller sample size.

Taking these into account, Figures 18 and 19 which plot the average salaries of men and women across teams and role seniority. A few worrying trends are immediately apparent as it seems that the average salaries of women are consistently lower than those of men across all levels of role seniority and also across many teams. We find that the average woman's salary is consistently under their male's counterparts in many teams such as Client Services and Data. However this is not true of all teams, and we find that some teams such as Software, Operations, and Legal & Financial seem to have an equitable distribution.

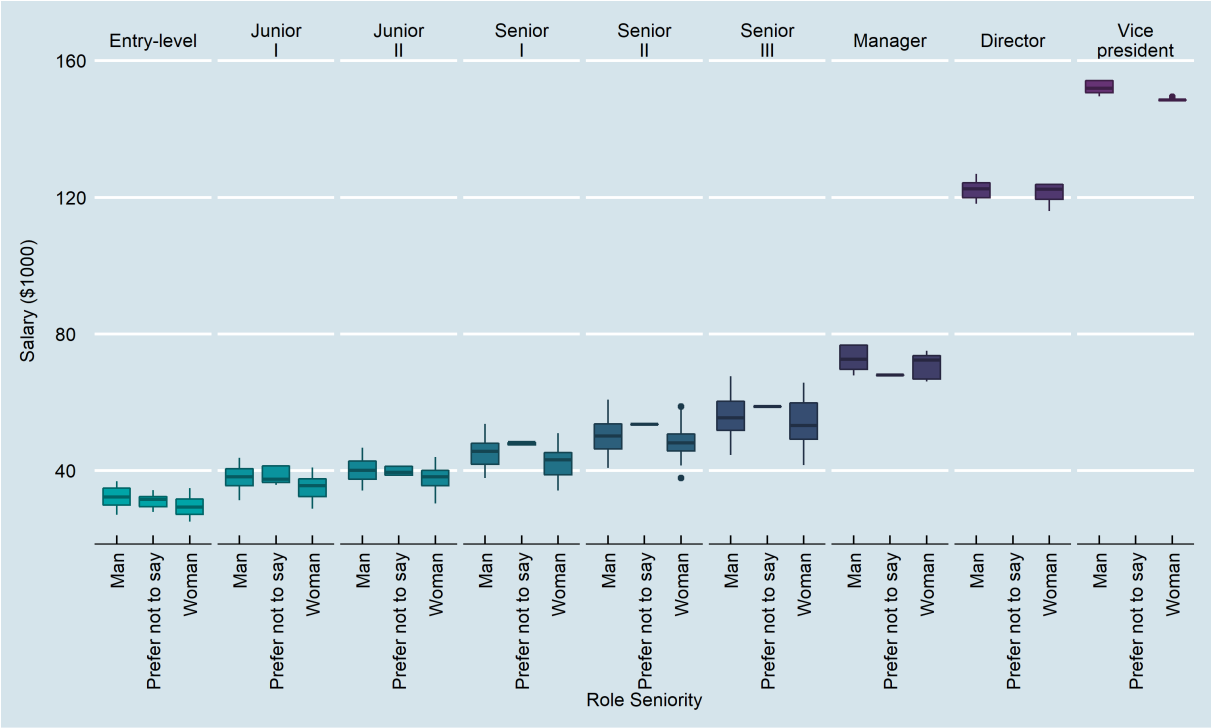

Figure 18: Salary by Seniority and Gender.

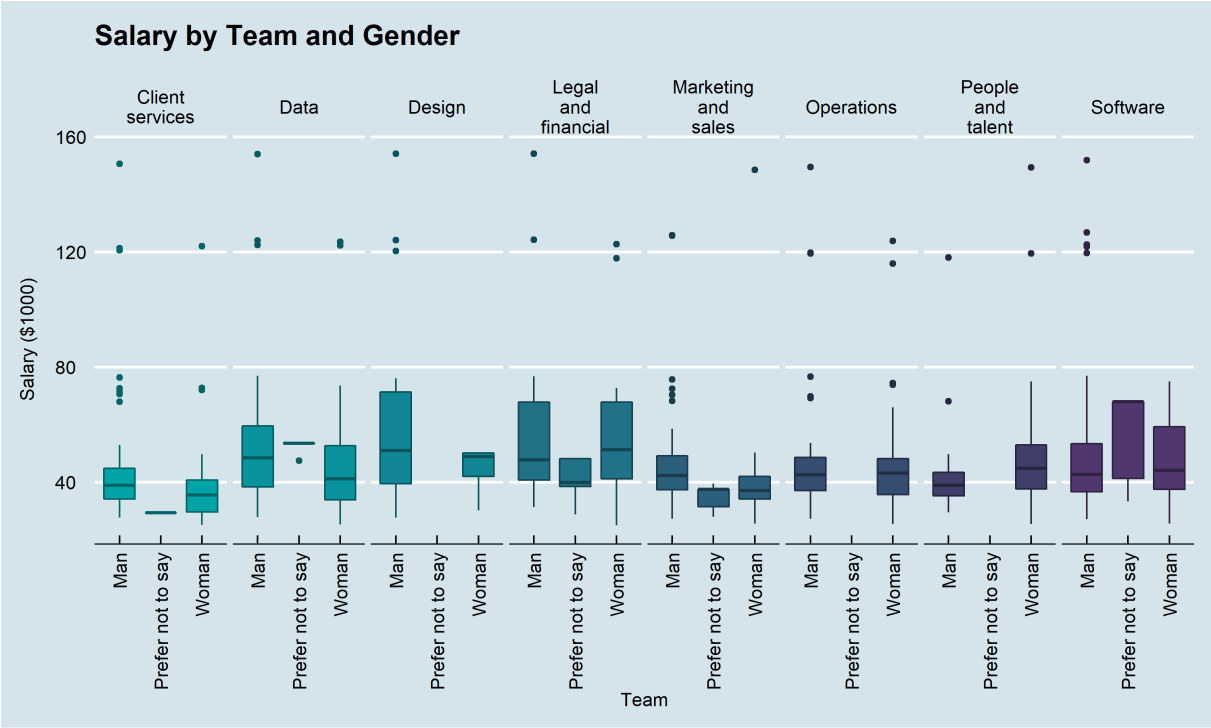

Figure 19: Salary by Team and Gender.

Taken in summary this suggests that Black Saber's remuneration decisions may be systematically biased against women, in particular that women's contributions at the same level seniority are undervalued compared to those of men. It also suggests that certain internal teams may specifically exacerbate this gap between men and women. However, this cursory analysis only provides a brief snapshot of the issue and warrants a more thorough analysis.

To this end, we build a linear mixed model that has salary as a dependent variable dependent on the various provided work productivity-based metrics. These include the productivity score, leadership evaluations, and role seniority. Due to repeated measures since the data contains many observations from the same individuals who are nested within a team, we fit a random intercept based on each employee ID as well as teams to capture this individual variance as well as the fact that different teams may inherently have higher salaries than others. We also include interactions terms between gender and each of the productivity based metrics to examine if gender has any effect on how these performance based metrics are valued in an individual. An interaction between gender and the random effect of team is also included to study any gender biases within teams.

$$\begin{aligned}\text{salary}_{ij} = \beta_0 &+ \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{productivity}_i + \beta_3 \cdot \text{role seniority}_i + \beta_4 \cdot \text{leadership for level}_i \\ &+ \beta_5 \cdot (\text{leadership for level}_i \times \text{gender}_i) + \ \beta_6 \cdot (\text{productivity}_i \times \text{gender}_i) \\ &+ \beta_7 \cdot (\text{role seniority}_i \times \text{gender}_i) + U_{1i} + U_{2j} + \epsilon_{ij}\end{aligned}$$

$$U_1 \sim N(0, \sigma_1^2)$$
$$U_2 \sim N(0, \sigma_2^2)$$

Note that above we are modelling the i^{th} individual on team j, so it follows that U_1 and U_2 are the respective distributions from which U_{1i}, U_{2i}, the random intercept for employee id of the i^{th} individual and the random intercept for team j are drawn from.

We find that an interaction between team and gender does not improve the model's ability to predict salary. Therefore, it does not appear that there is statistically significant bias towards or against any gender group between teams.

From this model, we make several key observations. We find that there only a few statistically significant predictors of salary to a 95% level of confidence. These include gender, role seniority, and their interaction terms, and the interaction term between gender and productivity.

We find that women on average make roughly $1024 less than their male counterparts with all other metrics remaining constant. Furthermore, we find that productivity is not a statistically significant but the interaction of productivity and gender in the case of women is statistically significant degree. In fact, evidence shows that women's productivity score actually correlates negatively with their salary. For each unit increase in a woman's assigned productivity, we see that their salary actually decreases by an average of $14.13 whereas an increase in productivity score. Considering the distribution of productivity with regards to women is not significantly different to men, we find this to be evidence of bias against women.

Another observation that we can make from this model is that role seniority is a significant factor in salary, but moreover, that the interaction of gender and role seniority is significant in the case of women where it is not for other genders. We find that women's salaries are lower than men for positions of the same level of seniority.

The results of this model therefore suggest that women are systematically underpaid and in particular, they are underpaid with to their colleagues of equal role seniority and their productivity is not taken into account in their salaries.

Discussion

First, from our analysis of the hiring phase, we find that there's no conclusive evidence of gender bias in the hiring phase since women do not seem to have a lower odds of moving on to subsequent phases due to gender and women's scores do not appear to be "weighted" lower. There are certain aspects of the process that are decidedly suspect. In all of the hiring phases applicants across genders performed very similarly, except in phase 2 where women appear to have lower speaking and leadership scores despite having otherwise very similar scores on all other assessments and meritocratic factors. Additionally, only approximately 29% of women from phase 3 were hired while approximately 53% of men from phase 3 were hired. Thus, we would recommend further investigation, especially looking at larger sample sizes, perhaps over a longer period of time with more data among those hired and phase 3 applicants so that these possible gender bias concerns can be better evaluated to either be truly natural variations in the data or as true gender biases.

In our investigation of the promotion rate for any possible gender bias, we consider both the total number of promotions that an individual can be expected to receive as well as the odds that an individual will be promoted at any given quarter. In terms of determining total lifetime promotions, it appears that the number of positive or adequate leadership evaluations, number of quarters since their last promotion, and gender are statistically significant predictors of the number of promotions. Promotion count also appears to not be statistically significantly dependent on an individual's average productivity.

We also found that since the odds of being promoted without being productive were effectively 0, and were determined to a significant degree solely by other factors such as leadership evaluations, time spent in their last role, and gender. The effect of gender however in both cases reduced a woman's expected number of lifetime promotions as well as their likelihood of being promoted by 37% as compared to men with the same track record at the company, such as having the same levels of productivity or experience. This suggests a clear bias against women in terms of hiring bias.

From our exploratory analysis and models focused on quantifying any salary gap between men and women, we find that women's salaries are on average lower than their male counterparts. This effect is significant in non-managerial roles but leads to women being paid significantly less than men for the same roles with the same levels of productivity. This discrepancy ranges between an average of $357 to $1606 depending on the role, with the gap. This discrepancy also grows with higher levels of role seniority. Therefore we conclude that women are undervalued when it comes to their seniority and their productivity. We also find that productivity not correlated to higher salaries only when the employee is a woman. In fact, we find that in the case of women, it appears that productivity has a negative association with their salary to the

scale of $14 less per point increase in productivity score. In general, this shows that there are severe deficiencies within Black Saber's salary decision processes that lead to discrimination against women and that the remuneration process is not fully based on an individual's merit.

Combining all of our findings, it is clear that not all merit-based metrics are being used to determine an individual's value to the company. We therefore suggest that Black Saber Software should re-evaluate its hiring, promotion, and salary decisions to ensure that they are being made holistically. That is including all aspects of an individual's work based metrics such as leadership evaluations, and to reduce bias against women in the interest of equity/diversity initiatives.

Limitations

One limitation of our analysis includes the way certain aspects of productivity were quantified. For example, leadership for level and productivity. These scores are only one way of quantifying an individual's productivity and the way that they are determined can be subjective and therefore subject to bias. For example, as leadership for level is determined by an evaluation by a supervisor, potential bias can leak into it, especially if women are considered to be less capable or leader-like as opposed to their male colleagues.

Our analysis is also limited in that it is predicated on data only from individuals who are current employees of Black Saber. This constitutes a form of selection bias that excludes previous employees of the company. It is reasonable to expect that many employees would leave a company where they were not being promoted or those who were consistently poor performers. This therefore may introduce some implicit bias in our dataset towards those remain at Black Saber.

This analysis is also limited by potential algorithmnic bias during data collection, in particular for hiring. In the case of hiring, many of the variables that are used to judge an individual, were automatically graded by an AI. As discussed previously, this injects an element of bias as the AI was likely trained on data that itself includes bias against women.

All the analyses also focus mainly on the difference between men and women in terms of equitability and cannot make any strong claims about non-binary individuals or those who chose not identify their gender. Although our analysis found no statistically significant differences, it is important to be cognizant that the relatively small sample size is a limiting factor in determining any bias.

Specifically in the analysis for hiring data, the small sample size might be a source of some inaccuracies, especially among later stages with less candidates.

Therefore all interpretations of the conclusions of this report should keep these limitations in mind in order to maintain a faithful and ethical representation of the actual situation at Black

Saber Software.

Future Work

In the future, it is recommended that Black Saber keep data on all employees, both past and present, as well as from multiple hiring periods so that an analysis be performed on a larger dataset can be performed. This would allow us to reduce the likelihood that any conclusions are merely due to random variation in the data. It would also eliminate the selection bias identified above that limits the scope of the conclusions that are found in this report.

We would also propose that in the future, Black Saber provides an option for individuals to self-identify as non-binary or another gender option as well as being given the option to 'prefer not to say'. This change would alleviate one of the limitations of the current analysis by allowing the explicit inclusion of non-binary individuals and removing the ambiguity in the 'prefer not to say' category.

We can also test the efficacy of these models in the future through splitting any data into a training and a testing set. This would allow for each model to be evaluated on its ability to generate the correct predictions on the company's data and could potentially be further evidence towards any trends discovered.

These strategies would all streamline Black Saber's data pipeline and give the company the ability to carry out more in-depth and consistent analysis in the future.

Consultant information

Consultant profiles

Brian Diep. Brian Diep is a senior consultant with Hamiltonian Path Consulting. He specializes in statistical modeling and bridging the gap between theoretical statistics and industry practice. Brian earned his Hons. Bachelor of Science, Majoring in Statistics and Linguistics, as well as a minor in Computer Science from the University of Toronto in 2023.

Monika Dydynski. Monika Dydynski is a senior data consultant with Hamiltonian Path Consulting. She specializes in business analytics and communicating data driven insights to clients from various industries. Monika earned her Honours Bachelor of Science, Majoring in Statistics and Mathematics with studies in Computer Science and Machine Learning from the University of Toronto in 2021.

Eric Zhu. Eric Zhu is a senior consultant with Hamiltonian Path Consulting. He specializes in statistical modelling with an emphasis on visualization and predictive analytics. Eric earned his HBSc. in Computer Science and Statistics from the University of Toronto in 2023.

Code of ethical conduct

Hamiltonian Consulting Co. is dedicated to an ethical standard of statistical consulting. All members of our team agree to abide by the following Code of ethical conduct.

- Our statistical consultants are forthright about assumptions, limitations or biases that have been identified in the data that may jeopardize the integrity of data analysis. We will disclose to Black Saber Software (hereafter, the client) that our analysis is subject to any bias that may be inherent within analysis or the data collection process.
- Our team of statistical consultants are dedicated to a high standard of statistical practice and transparency. The client can trust that the methods we follow will provide objective and reproducible results, as well as honest data driven insights.
- Our statistical consultants swear to protect proprietary and confidential data and abide by additional confidentiality requirements stipulated in the contract between Hamiltonian Consulting Co. and the client. We will not attempt to identify the respondents within data obtained from the client, nor use the data for any other purpose without explicit permission from the client.