There Are No Colorblind Models in a Colorful World: How to Successfully Apply a People Analytics Tool to Build Equitable Workplaces

David R. Anderson
Assistant Professor of Business Analytics
Villanova University

Margrét V. Bjarnadóttir
Associate Professor of Management Science and Statistics
University of Maryland

David Ross
R. Perry Frankland Associate Professor
University of Florida

Abstract

The field of People Analytics is exploding, with algorithms increasingly having life-changing impacts on employees. The data-driven decision support systems fueled by these algorithms have the potential to revolutionize operations across every facet of HR: improving hiring decisions; identifying high-performing workers; identifying skill gaps and optimizing workforce training; and ensuring compensation is fair and competitive. However, entrusting black-box algorithms with supporting such critical decisions poses the risk of unintended consequences. If the retrospective data supporting these innovative algorithmic tools reflect a status quo of biased decisions, algorithms may do harm: rather than improving decisions, they could simply entrench and automate biases while giving them a veneer of objectivity. In this paper, we examine the potential pitfalls of blindly applying standard analytical models to HR data, highlight best practices, and discuss how to build a bias-aware analytical process that can prevent machine learning biases from creeping into HR practices. We show that it is not enough for algorithmic models to simply be race-blind or gender-blind; rather, these models need to affirmatively identify and correct for unwanted biases.

There Are No Colorblind Models in a Colorful World: How to Successfully Apply a People Analytics Tool to Build Equitable Workplaces

Artificial intelligence and machine learning are revolutionizing the practice of HR management. By automating the collection and analysis of large datasets, People Analytics offers the promise of improving every phase of the people pipeline, from recruitment and compensation to promotion, training, and evaluation. Startups, consultants, and internal HR analysts are drawing on a wide array of scientific domains, including natural language processing, statistical modeling, and applied psychology, to make better decisions about how to recruit, train, and manage effective workforces.

People Analytics thus offers tremendous promise in employing automated machine learning algorithms to help managers measure productivity and make important personnel decisions. Algorithms are currently being used to support hiring decisions, promotion and training opportunities, and compensation decisions, all of which may be life-changing for employees. This human impact makes it all the more crucial that the decision input provided by these algorithms is fair and transparent. However, this is not always the case. Researchers and practitioners across application areas have voiced concerns about the harm that may be caused by decision support tools using automated machine learning algorithms and about the lack of accountability associated with them. The popular press provides multiple examples of people analytics leading managers astray.

For instance, Amazon made news headlines when it had to throw away a resume screening tool built by its engineers to select the best job candidates from the applicant pool. This followed the discovery that the tool had penalized resumes of women applicants for technical job roles by basing decisions on phrases like “women’s chess team.” [1] Because these technical jobs had been dominated by men, a model trained by looking at previously hired and successful candidates learned to read factors correlated with maleness as indicators of potential success.

LinkedIn also made headlines [2] when the auto-complete feature on its website’s text search box suggested replacing female names such as “Stephanie” with male names like “Stephen.” The reason given by the company was that the auto-complete suggestions were based on volume and there were simply more people named Stephen than Stephanie on the platform.

Finally, a study [3] highlighted an ad for Science, Technology, Engineering and Math (STEM) field opportunities that had been carefully designed to be gender neutral, yet a social media algorithm resulted in this gender-neutral ad being shown to disproportionately more men than women. The reason was that the algorithm was optimized to maximize cost-effectiveness by showing the ad to as many people possible with the given budget. Because this social network charged a higher cost to show ads to young women (who typically buy more through ads) than it did to show ads to young men, the algorithm decided to target the ad to men.

Each of these examples highlights a breakdown in the analytical process. Yet these breakdowns can be prevented, and prevention begins with an understanding of the root causes of algorithmic bias. In a nutshell, the possible causes include biased retrospective data, surrogates for actual outcomes when predicting success, underrepresentation of key groups in the data, and failure in the machine learning process.

Over the past few years, there has been a proliferation of studies focusing on understanding why algorithms behave the way they do. Here, we highlight two insights from the literature, which we think are critical to understanding why algorithms can produce biased results. These insights should inform how we think about the data analytics process and how we apply machine learning models to support decision making in HR.

In 2016, ProPublica released the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) data. The COMPAS model is a proprietary algorithm used to evaluate the risk of recidivism. The corresponding risk scores can be used in court to help determine parole decisions. What the data showed was that black defendants were more likely to be falsely labeled as risky than white defendants. Conversely, white defendants were more likely to be falsely labeled as being at low risk of recidivism than black defendants. [4] This is in conflict with the goal of group fairness, that one group should not benefit or be harmed more than another by the application of a risk model. The developers of COMPAS argued that their model was well calibrated, meaning that for any given risk of recidivism, approximately that proportion of subjects went on to commit future crimes. For example, considering all the parolees with risk scores of roughly 20%, on average 20% went on to commit crimes. [5]

The discussion and subsequent academic publications that followed the ProPublica article highlighted that if the underlying risk of two groups (here, black and white defendants) is not the same, then it is (almost) mathematically impossible for a model to be well calibrated and have equal probabilities of false positives and false negatives for each group[1]. [6,7] Thus, people of the race with a higher rate of recidivism are more subject to one type of misclassification by the algorithm (false positive), and people of the race with a lower rates of recidivism are subject to another type of misclassification (false negative). Translated to the context of people analytics, a well-constructed model to identify future managers, where most managers in the past have been male, will be more likely to falsely identify male employees as management candidates (in addition to correctly identifying many male candidates) compared to females, and similarly, the algorithm is more likely to miss deserving female candidates.

The second important finding from the literature is the fact that there is no such thing as a truly “race-blind” or “gender-blind” model. Multiple studies have shown that as long as the features used to construct the model are correlated with the demographic characteristics of the employees, simply leaving certain demographic characteristics out of the model is not enough to eliminate bias. For example, in lending, a model that does not include variables for race may still include information, like ZIP code, that carries information about the applicant’s race. If the original human decisions included racial bias and those decisions are reflected in the data used to train the model, then including ZIP code in a model will allow that bias to remain. Therefore, in order to ensure that a machine learning model is truly fair, additional steps need to be taken beyond simply removing demographic variables.

In what follows, we highlight the modeling process, offer examples of some of the issues that may arise, and discuss potential solutions. We then provide a checklist for developing fair People Analytics algorithms and summarize the path forward.

[1] To see why, suppose that a machine learning algorithm determines that certain socio-economic factors like residential neighborhood, income, and education level strongly predict recidivism. Then further suppose that these factors also strongly predict race. Then, individuals of that race will tend to get labeled as high risk for recidivism. When the algorithm is “wrong” about a person of that race (i.e., when it misclassifies them), it labels that person as high-risk even though they do not commit another crime (a false positive). Conversely, when the algorithm misclassifies someone of the other race, it labels that person as low-risk even though they *do* commit another crime (a false negative).

From Data to Decision

Behind every decision support tool is a process that starts with "raw data." This data is then cleaned and transformed, then fed into machine learning models that predict outcomes. Finally, these predictions are then used as bases for decision making, either by fully automated "artificial intelligence" tools or by tools designed to support human decision makers. Figure 1 illustrates this process.

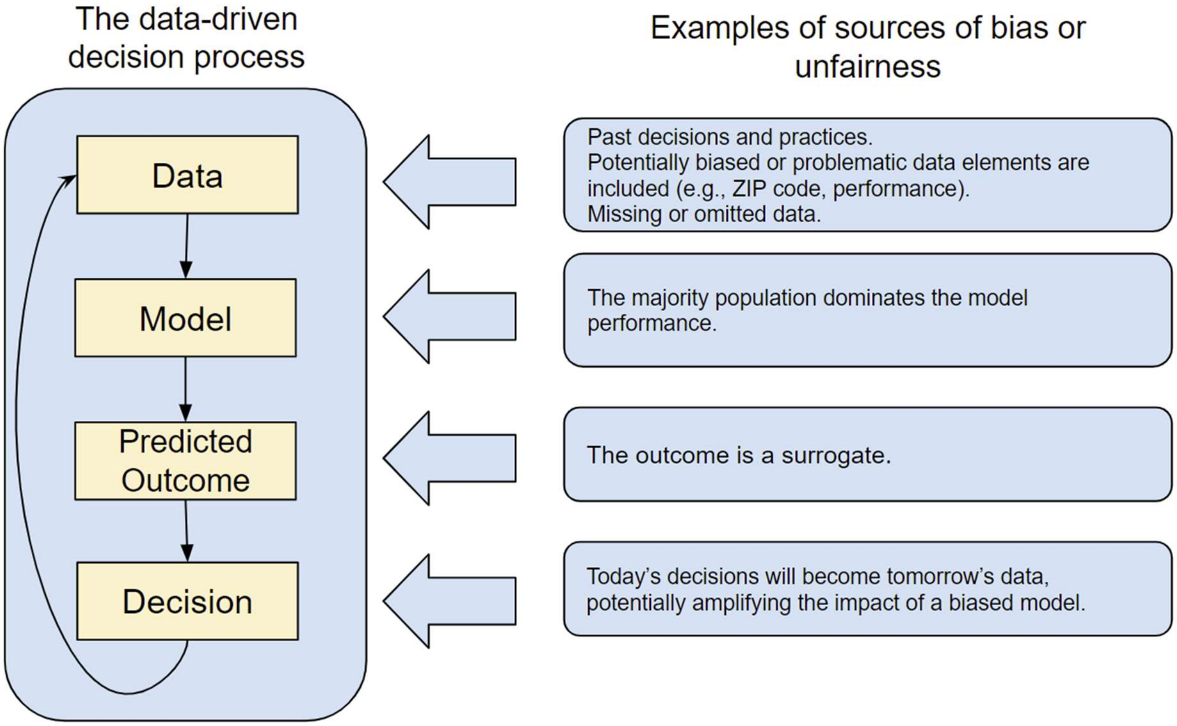

The data-driven decision process
Data
Model
Predicted Outcome
Decision
Examples of sources of bias or unfairness
Past decisions and practices.
Potentially biased or problematic data elements are included (e.g., ZIP code, performance).
Missing or omitted data.
The majority population dominates the model performance.
The outcome is a surrogate.
Today's decisions will become tomorrow's data, potentially amplifying the impact of a biased model.

Figure 1. Overview of the data-driven decision-making process

While this process may appear objective, it is filled with human judgment at multiple levels. How is data categorized? How is success defined? How is the model built? How is the predicted outcome used when making a decision? And perhaps most importantly, what data are being used?

Data

Retrospective data, which is often used to build People Analytics tools, will always reflect the decisions and attitudes of the past. Therefore, as we attempt to build the workplace of tomorrow, we need to be mindful of how our retrospective data may be biased and may not fully capture the complexities of people management. Below we discuss additional important considerations.

In People Analytics models, we often use features that are surrogates for the underlying employee qualities that we want to capture. For example, undergraduate GPA might be used as a proxy for intelligence, occupational licenses or certificates may be a measure of skills, or travel required may be a proxy for the time demands of the job. However, each of these is an incomplete measure of the target concept and often contain biases and distortions. For example, job applicants who had to work during college may have gotten lower grades, but they may actually make the best job candidates because they have demonstrated the drive to overcome obstacles. Understanding such potential mismatches between a proxy and the target construct is important for any user of a quantitative model, especially when the goal is to build a more diverse workplace.

Another way that the data can bias our models is that we only have partially observed data. For example, if we are trying to estimate how well an employee will perform as a manager, we only know the performance of those who have been promoted to managerial positions in the past; we do not observe how well those who did not get promoted would have performed. As a result, when we build a model to predict managerial performance to identify future candidates, our model will be biased in such a way that it predicts managerial potential based on the characteristics of employees who were, in the past, identified (by humans) to be likely to succeed.

Finally, it is important to note that some key features that drive the outcome (e.g., an employee's personality may affect their managerial performance) may not be captured in the data. These factors should be documented and accounted for in the decision-making process. While machines can help interpret past data and identify patterns, People Analytics is still a human-centered field, and in many cases the final decisions are still going to be made by humans (reflected in the current popular catch phrase "human-in-the-loop-analytics").

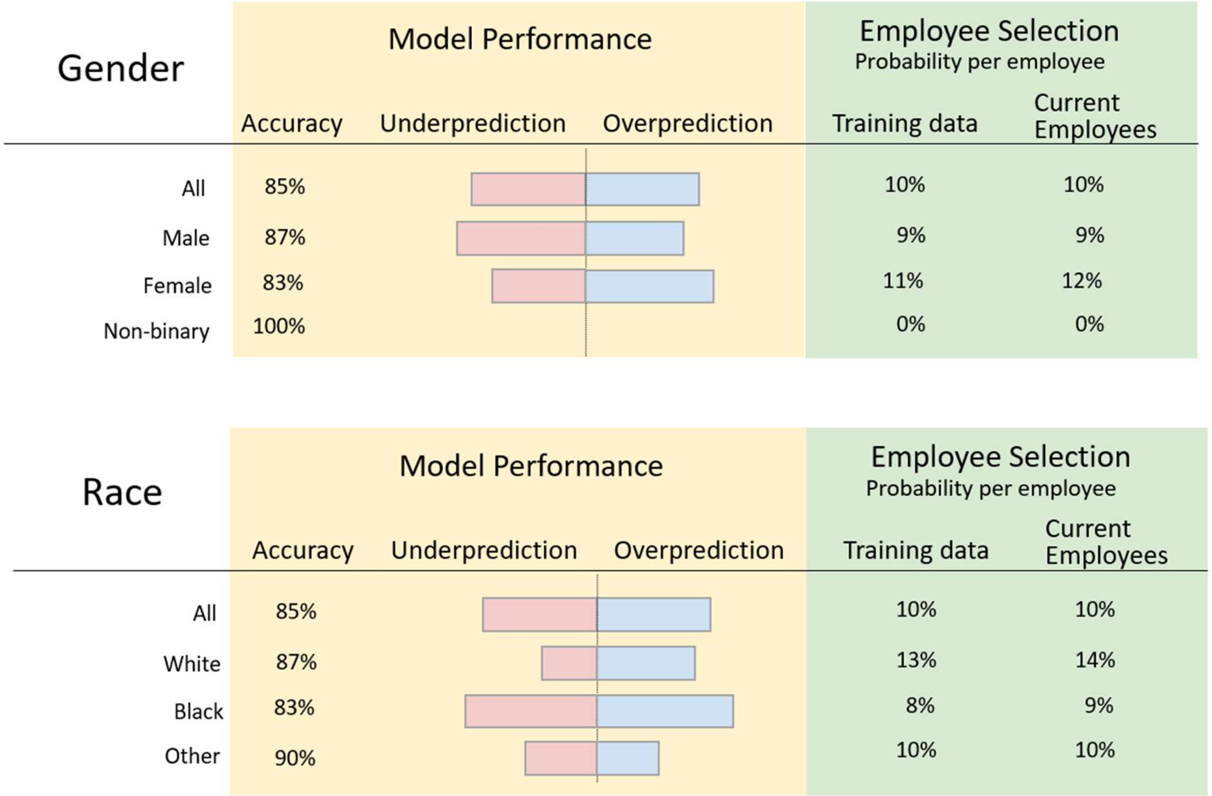
Modeling Choices

In building People Analytics models, analysts need to be mindful of the composition of the underlying data. A model that maximizes the overall quality of the prediction (which is the standard approach) is likely to perform best with regard to individuals in majority demographic groups, as discussed in the Amazon resume screening example. This is because the algorithms are typically maximizing overall accuracy, and therefore the performance for the majority population has more weight than the performance for the minority population.

Often, underrepresentation in data could be addressed by collecting more data. For example, a biased skin cancer diagnostic tool may be fixed by collecting additional images of dark-skinned people. In our context, organizations are unable to go back in time and hire a more diverse employee population to support today's decision making. The issue of underrepresentation in the training data must then be directly addressed in the modeling phase. Modelers need to understand the representation of the groups in the data and the model's performance for each group. If the performance is not adequate across groups, additional corrective steps need to be taken to ensure that the model performs well on each subgroup of the data.

The optimal approach to such correction differs based on the application. In some cases, it is best to train different models for each subgroup. This is especially useful if, for example, the factors that indicate success among one group of candidates differ from the factors indicating success factors for the majority. If, in the past, a company's candidates came mostly from majority groups, it is necessary to take a step back and develop an understanding of appropriate indicators of success for the minority population and whether and how they differ from the majority population. If there is insufficient data, more human intervention may be required to evaluate minority candidates.

Lastly, as discussed above, it is typical in the world of People Analytics for some observable job-related variables like job role and education to be correlated with protected employee attributes (e.g., race and gender). As a result, algorithms that are trained on other employee features while ignoring protected demographic information can still be influenced by biases, so the related decision support tools may favor one group over another. To address this issue and the potential deficit in algorithm performance for smaller demographic groups, a critical part of any modeling framework should be a careful study of the performance of the models for different demographic groups. Figure 2 shows a dashboard that could help visualize these differences.[2]



Gender	Model Performance			Employee Selection Probability per employee	
	Accuracy	Underprediction	Overprediction	Training data	Current Employees
All	85%			10%	10%
Male	87%			9%	9%
Female	83%			11%	12%
Non-binary	100%			0%	0%

Race	Model Performance			Employee Selection Probability per employee	
	Accuracy	Underprediction	Overprediction	Training data	Current Employees
All	85%			10%	10%
White	87%			13%	14%
Black	83%			8%	9%
Other	90%			10%	10%

Figure 2. Model performance dashboard for a model tasked with identifying employees for a management training program, lending itself to two observations. First, the model shows 100% accuracy for non-binary employees, but at the same time, no non-binary employees have ever been selected for the training program, so this this result comes from not having collected data on non-binary status in the past. Second, the model is more likely to select white employees into the training program, especially when compared to black employees.

Decisions and Feedback

Just as our employee population of yesterday is used to train the People Analytics models of today, today's workforce will serve as the training data for tomorrow's models. Therefore, the decisions made today will shape the decisions of tomorrow, which can be problematic if the models deployed are biased. This unfortunate cycle is referred to as unintended feedback bias.

[2] We note that new solutions exist to assist with providing bias checks. Two of those tools that are relatively easy to use, offer good visualization capabilities, and propose adjustments to a model to counteract biases are the FairLearn toolkit, developed by Microsoft, and AI Fairness360, developed by IBM. Both offer dashboards with fairness metrics.

The example typically used to illustrate this phenomenon is that of predictive law enforcement. The idea is that by using arrest data to inform patrol deployment decisions, police departments can better serve communities by reducing crime rates. However, since arrests are more likely in heavily policed areas, the deployment decisions based on prediction of crime hotspots will focus policing efforts on already overly policed communities. [8] In the context of People Analytics, this could happen, for example, when hiring decisions are based on "cultural fit," thereby perpetuating demographic imbalances.

The Bias-Aware Data-Driven Decision Process

Given the discussion above, it is critical that the analytical process reflected in Figure 1 not be applied blindly as a data-driven bias-free process. Rather, we should apply a bias-aware process with a human in the loop. It is critical that care is taken to be aware of potential biases in the data collection process, what proxies are used, how the model performs for different groups, and how it is being deployed. In Figure 3, we present a bias-aware process that facilitates fairer application of people analytics tools.

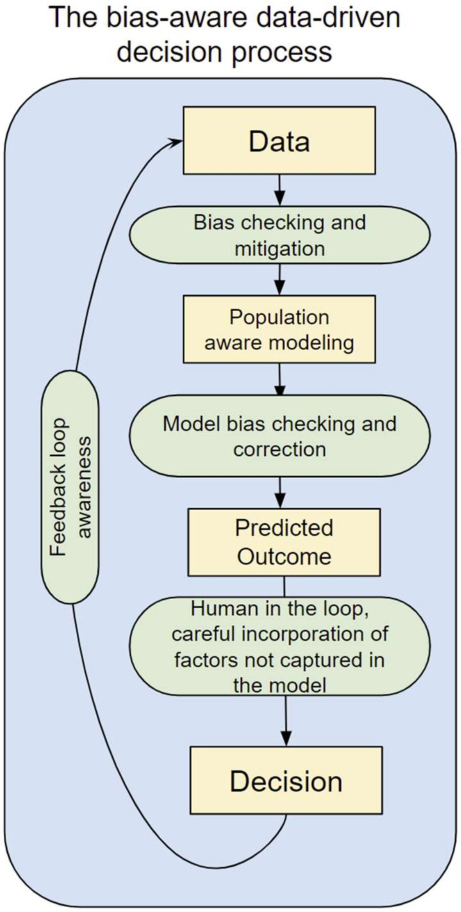
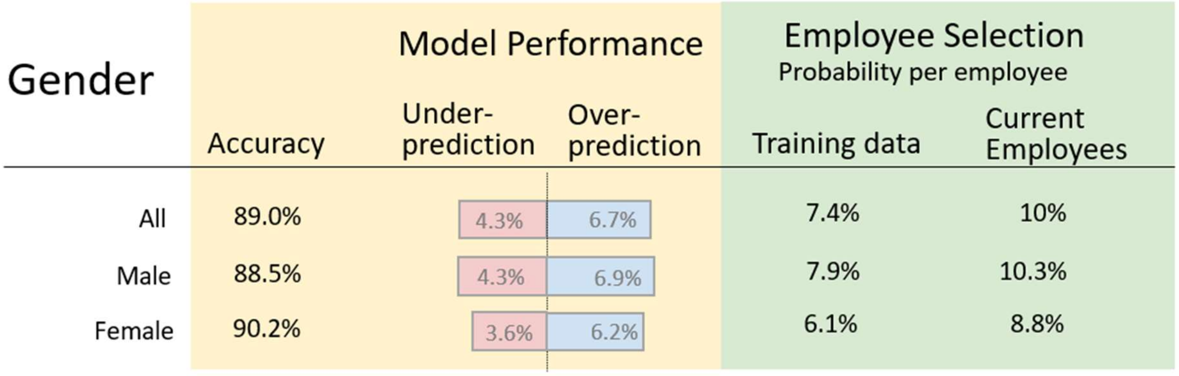

The bias-aware data-driven decision process
Data
Bias checking and mitigation
Population aware modeling
Model bias checking and correction
Predicted Outcome
Human in the loop, careful incorporation of factors not captured in the model
Decision
Feedback loop awareness

Figure 3. Bias-Aware Data-Driven Decision Making Progress

Empirical Example: Identifying Future IT Leaders

To illustrate some of the ideas discussed above, we use data from the IT department of the Veterans Administration (VA). This data was obtained through a Freedom of Information Act request and includes longitudinal data from the Office of Personnel Management (OPM) on every US federal government employee, excluding the Department of Defense. This area of the VA is male-dominated, with roughly 75% of the 5,800 employees being men. We want to clearly state that this is not an example of how the VA built and executed the analytical process. Rather, we use this data to demonstrate some of the challenges highlighted above by building a hypothetical model that demonstrates how applying a data-driven approach without bias awareness can amplify disparities. We use the organization's promotion history as the outcome we model, then use the model to identify current employees who should be considered for a management training program (and potentially an eventual promotion). We show that when there is a historical bias towards male employees, a so-called gender-blind model can simply entrench the bias. We then discuss how the resulting decision-making process can be improved and made more equitable.

Using the OPM data, we find all the employees in the VA IT department who were promoted between Q1 2005 and Q1 2006 (a raise of >10% and an increase in pay grade, roughly 7% of employees). We build a model that uses past promotion history as a signal of employee potential, then use it to identify 10% of employees for a management training program. We first build a gender-blind model and predict the likelihood that each employee in the organization will be promoted in the following year. The model identified employees with higher education, younger age, and lower job grade as more likely to be promoted. The bias dashboard of the resulting model is presented in Figure 4.



Gender	Model Performance			Employee Selection Probability per employee	
	Accuracy	Under-prediction	Over-prediction	Training data	Current Employees
All	89.0%	4.3%	6.7%	7.4%	10%
Male	88.5%	4.3%	6.9%	7.9%	10.3%
Female	90.2%	3.6%	6.2%	6.1%	8.8%

Figure 4. Bias dashboard showing the results of the gender-blind model. We note that the accuracy is higher for female employees and that the promotion rates in the past (reflected in the training data) are higher for men. As a result, when this model is applied to select 10% of employees as candidates for managerial training, the probability of a male employee being selected is much higher than the probability of a female employee being selected.

The figure demonstrates that while the prediction accuracy is similar for men and women, the employee selection probabilities vary, both in our training data and in our model when the data is applied to future employees. This raises an alarm about potential bias and should prompt the modelers to dive in deeper in order to gain an understanding of the issue. It could be that male candidates are more likely to have a higher education, which may justify the higher promotion rates. In order to exclude this possibility, we build a second model including gender as a predictor variable for promotion. We find that in this gender-aware model, men are significantly more likely to be promoted than similar women are, with on average 23% higher odds of promotion.

It is then the job of the human analyst to consider whether there may be a bias in the promotion process or whether there may be objective explanations for the difference in promotion rates. For instance, is there some missing variable that should be captured in the modeling process (e.g., annual review feedback – while acknowledging but sidelining the potential biases in performance data)? Are there outside factors that explain the difference (e.g., VA centers with onsite daycare had similar promotion rates between genders)? In that case, if the organization is willing to make changes that support equal promotion rates, the answer may be a logistical one (expand childcare access). If all possible causes are ruled out, and if it is confirmed that the difference is due to bias, the model and/or its outputs must be addressed directly. We will assume that this is the case in our example.

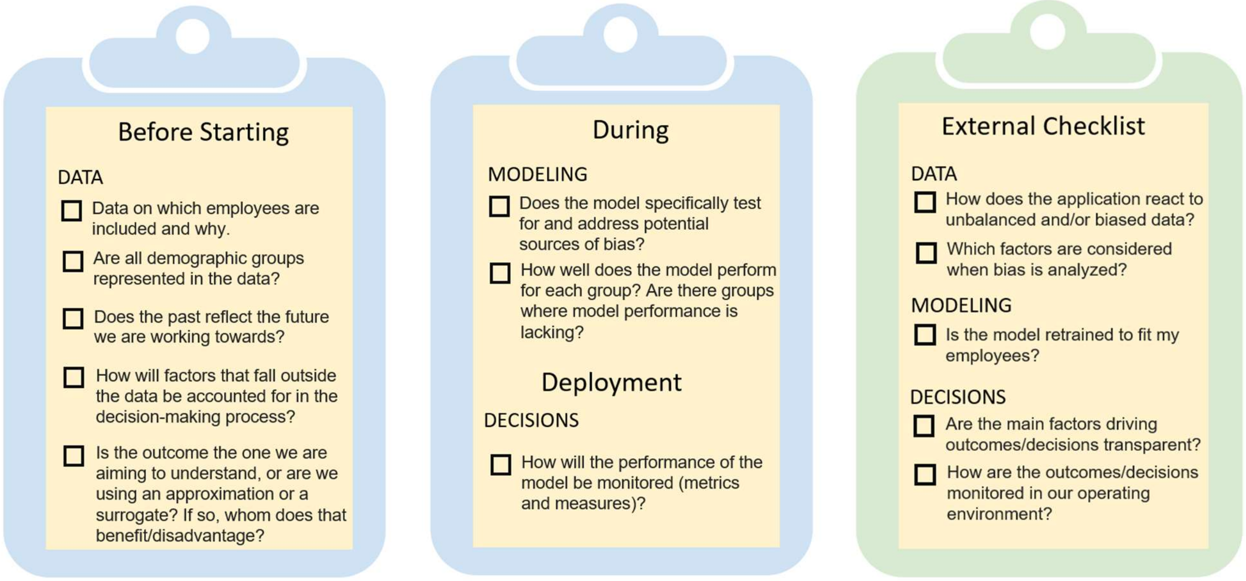
To begin to address this issue, the organization could look at a number of technical modeling solutions, including increasing the weight of particular employees in the modeling process or applying de-biasing techniques in the model building. Because here, we have explicitly measured the gender bias in the historical data, we will account for this bias by adjusting the predictions accordingly. In other words, since the model has exhibited a hidden, implicit bias, we remove the bias from the model by testing for it, identifying it, and quantifying it—in short, we make the bias explicit so we can rectify it. We calculate that female employees' scores give them 23% reduced odds of promotion, so we add this difference to the predicted score for each female employee. This reduces the gender disparities in the employees identified as promotion candidates and ensures that a similar proportion of each gender are selected. After applying this process, we have the same model, but it now includes an explicit bias correction. This model identifies 9.3%

of women and 10.1% of men as candidates for promotion, reflective of differences in the characteristics of the male and female employees. We note that this approach is not a quota-based approach, which constrains the model to pick equal proportions of each demographic group. Rather, it simply measures and accounts for the bias in the estimation of promotion potential.

The Path Forward

As is evident from the above discussion, there is no silver bullet that will make each algorithm fair and every People Analytics tool unbiased. The adoption and utilization of People Analytics tools needs to be based on the principles of equity, transparency and fairness. The best weapon in that fight is an understanding of the underlying biases, how modeling approaches address them, and how human-aided analytics (human in the loop) can overcome them. In Figure 5, we provide a checklist that summarizes key questions to ask at each step in the analytical process, as well as a second checklist to ask any external analytical vendor before implementing their tool.

There are some issues that even conscientious modeling cannot address. However, being aware of the issues and taking affirmative steps to address them, rather than simply trusting that models that do not explicitly include demographic information are unbiased, can go a long way towards mitigating the potential harm of algorithmic bias. By asking the right questions of the data, the model, the decisions, and the software vendors, managers can successfully harness the power of People Analytics to build the high-achieving, equitable workplaces of tomorrow.



Before Starting

DATA

□ Data on which employees are included and why.

□ Are all demographic groups represented in the data?

□ Does the past reflect the future we are working towards?

□ How will factors that fall outside the data be accounted for in the decision-making process?

□ Is the outcome the one we are aiming to understand, or are we using an approximation or a surrogate? If so, whom does that benefit/disadvantage?

During

MODELING

□ Does the model specifically test for and address potential sources of bias?

□ How well does the model perform for each group? Are there groups where model performance is lacking?

Deployment

DECISIONS

□ How will the performance of the model be monitored (metrics and measures)?

External Checklist

DATA

□ How does the application react to unbalanced and/or biased data?

□ Which factors are considered when bias is analyzed?

MODELING

□ Is the model retrained to fit my employees?

DECISIONS

□ Are the main factors driving outcomes/decisions transparent?

□ How are the outcomes/decisions monitored in our operating environment?

Figure 5. High-level checklist to bring bias awareness to the modeling process

Bibliography

[1] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.com. Accessed online on 3/11/2021 at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[2] Matt Day. How LinkedIn's search engine may reflect a gender bias. Seattle Times. Accessed online on 3/12/2021 at: https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/.

[3] Anja Lambrecht and Catherine Tucker. "Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads." *Management Science* 65.7 (2019): 2966-2981.

[4] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. Machine Bias. ProPublica. Accessed online 3/14/2021 at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[5] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." Federal Probation, 80(2): 38-46, 2016.

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2):153–163, 2017.

[7] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In 8th Innovations in Theoretical Computer Science Conference, ITCS, 2017.

[8] Kristian Lum and William Isaac. To predict and serve? Significance, 13(5):14–19, 2016

Recommended reading

An issue related to algorithmic bias is the issue of the ethical application of algorithmic decision making. We highly recommend these two papers as an inroad into that discussion.

Martin, Kirsten E. "Designing ethical algorithms." *MIS Quarterly Executive June* (2019).

Martin, Kirsten. "Ethical implications and accountability of algorithms." *Journal of Business Ethics* 160.4 (2019): 835-850.

And as one additional cautionary tale of AI gone wrong, we recommend this amazing (for all the wrong reasons) story.

https://web.br.de/interaktiv/ki-bewerbung/en/