Ethical Dilemmas in Algorithmic Decision-Making: An Algorithm Discriminates

Madhur Dixit
College of Engineering
Computer Science
North Carolina State University
Raleigh, North Carolina 27606
Email: www.mvdixit@ncsu.edu

***Abstract*—This case study examines the ethical challenges faced by Sandra, a software developer at Emporia, who discovers unintended disparate impacts in a resume screening software she designed. Despite reducing salesperson attrition, the software exhibits bias favoring white applicants due to a flawed proximity metric to Emporia stores. The case explores Sandra's ethical responsibility, professional and legal obligations at Emporia, and Timothy's reluctance to address the disparate impact issue. It questions the use of objective criteria, proposing alternative metrics for fair recommendations. Positioned within the context of algorithmic ethics, the study emphasizes the need for non-technical solutions and diverse expertise to address underlying issues in hiring fairness. The case prompts reflection on the ethical dimensions of software development and the potential implications of biased algorithms for employment practices.**

I. Introduction

We're looking at a challenging scenario in the rapidly evolving fields of technology and employment. Software developer Sandra from Emporia created a tool to aid in the hiring of more qualified salesmen. With a twist: minority job seekers may experience some unforeseen issues as a result of its success. This paper examines Sandra's problems and considers ethical issues while developing computer programs that impact people's jobs. We will discuss workplace policies, ethical obligations of IT professionals like Sandra, and the consequences of biased computer programs. We will unravel the moral implications of employing software and algorithms to make judgments or their effects on the actual world, exploring the unforeseen problems that may arise, using the Emporia case as an example.

mds
November 16, 2023

II. Case Study

*1) **Identifying the problem**:* Sandra's present dilemma is on the inadvertent emergence of moral dilemmas throughout Emporia's employment process—a problem that the company is unaware of. According to Sandra's study, there is a major imbalance. Even though 80% of applicants for sales roles are Black and Latino, 92% of the recently employed staff are White. The Principal Component Analysis (PCA) features were the source of the algorithm's racial prejudice.

*2) **Identifying the non-ethical issues**:* The software's hiring algorithm created some immoral challenges in addition to the obvious ethical ones, putting the organization in danger of employment discrimination lawsuits. The Civil Rights Act of 1964's disparate impact principle, which forbids employers from using any employment practices that have an unwarranted negative impact on members of protected classes, such as women, people of color, or those from lower-income backgrounds, may have been inadvertently broken by the company. A smaller pool of candidates may be discouraged from applying as a result of such policies' lack of diversity, which could have negative effects on the company's reputation among prospective applicants in addition to legal ramifications.

*3) **Identifying the stakeholders**:* The head of the Human Resources (HR) department, Sandra, whose employment may be in jeopardy as a result of her actions, and the firm Emporia, along with its wider range of stakeholders, are the main stakeholders involved in this circumstance. The choices made about the algorithm may have serious repercussions, such harming the company's brand or causing losses in money. If the algorithm is kept in place, it could cause legal problems, criticism, and damage to the company's reputation because of bias. It might also result in Sandra and Timothy, the HR department head, being fired. However, adjusting the algorithm's parameters to account for bias may have an adverse influence on how effective it is, possibly producing less than ideal outcomes and making the issue that the algorithm was meant to address—high sales department turnover rates—worse. Moreover, the existing algorithm unintentionally leads to the hiring of white people or people who live close to the organization with preference. Because of this unintentional prejudice, worthy individuals who might not fit these requirements are suffering unfair outcomes. It is not the best practice to base recruiting decisions solely on a candidate's proximity to the company—this will omit qualified applicants who could live further away. To ensure a fair and inclusive hiring process where talent is prioritized over arbitrary considerations, it is imperative to address this issue.

*4) **Possible Options or Alternatives**:*

- **External Audit by Experts:** Propose conducting an external audit of the software by independent experts specializing in algorithmic bias and fairness. This audit

can provide an unbiased evaluation of the software's impact on diverse candidate pools and recommend adjustments to mitigate any unintentional discrimination while maintaining the positive outcomes achieved in terms of retention and sales.

- **Employee Input and Feedback:** Advocate for gathering input from current employees, especially those hired after the implementation of the software. Conduct surveys or focus group discussions to understand their perspectives on the hiring process, ensuring that the software aligns with the company's commitment to fairness and diversity.
- **Fairness-Adjusted Weighting:** Suggest adjusting the weighting of features within the current algorithm to mitigate the impact of zip code on the recommendation scores. By assigning different weights to factors related to qualifications and tenure, the algorithm can be fine-tuned to be more inclusive and avoid favoring candidates based solely on their proximity to company locations.
- **Iterative Testing and Redesign:** Propose an iterative approach where the software design undergoes continuous testing and refinement. Regularly assess the algorithm's impact on diverse candidate pools, gather feedback, and make adjustments accordingly. This agile methodology allows for ongoing improvements to address any unintended biases while maintaining positive outcomes.
- **Algorithmic Bias Mitigation Tool:** Implement an algorithmic bias mitigation tool that can identify and address potential biases in the recommendation algorithm. These tools, such as AI Fairness 360 by IBM, can help Sandra detect and mitigate biases by providing metrics and algorithms designed for fairness testing and interventions. This tool can assist Sandra in quantifying and mitigating biases, aligning with best practices for responsible AI development[3].
- **Fairness-aware Machine Learning Frameworks:** Adopt fairness-aware machine learning frameworks, like Fairness Indicators by Google, that allow developers to visualize and evaluate model performance across different demographic groups. Sandra can use these frameworks to ensure fairness during the software development life cycle. Fairness indicators can provide a transparent view of how the model performs across various demographic groups, facilitating informed decisions during the design phase[4].
- **Ethical AI Guidelines Integration:** Incorporate ethical AI guidelines into the software development process. Utilize frameworks such as the Ethical AI Toolkit by Microsoft, which provides practical guidance and resources for embedding ethical considerations into AI projects. Following established ethical AI guidelines can help Sandra design software that aligns with industry best practices, ensuring fairness and minimizing unintended biases[5].
- **Explainable AI Models:** Use explainable AI models to enhance transparency in the decision-making process. Models like LIME (Local Interpretable Model-Agnostic Explanations) can help Sandra understand how the current algorithm makes decisions and identify potential sources of bias. Explainable AI models enable developers to interpret and debug complex algorithms, making it easier to identify and address biases[6].
- **Re-evaluate Feature Selection Criteria:** Revisit the criteria used for feature selection, considering a broader set of factors beyond zip codes. Collaborate with domain experts and diversity teams to identify relevant features that better reflect qualifications and the potential for success. By reassessing feature selection criteria, Sandra can create a more inclusive model that avoids reliance on factors that may introduce bias.

*5) **Tests**:*

- **External Audit by Experts**
  - **Harm Test:** Moderate harm due to potential costs and time.
  - **Publicity Test:** Positive, demonstrating a commitment to transparency.
  - **Defensibility Test:** Highly defensible, relying on external experts.
  - **Reversibility Test:** Positive, as it aims for an unbiased evaluation.
  - **Virtue Test:** Shows a commitment to objectivity and improvement.
  - **Colleague Test:** Positive response for seeking external validation.
  - **Professional Test:** Generally supported as a rigorous approach.
  - **Organization Test:** May require explaining the need for external scrutiny.
- **Employee Input and Feedback**
  - **Harm Test:** Low harm, involving employees in decision-making.
  - **Publicity Test:** Positive, emphasizing a democratic approach.
  - **Defensibility Test:** Easily defensible, showing a commitment to employee voices.
  - **Reversibility Test:** Positive, demonstrating adaptability based on feedback.
  - **Virtue Test:** Fosters a culture of listening and valuing employee opinions.
  - **Colleague Test:** Positive response for employee engagement.
  - **Professional Test:** It is generally supported as it aligns with employee-centered ethics.
  - **Organization Test:** Supports a culture of openness and improvement.
- **Fairness-Adjusted Weighting**
  - **Harm Test:** Low harm involves adjusting existing criteria.
  - **Publicity Test:** Positive, as it demonstrates a commitment to fairness.
  - **Defensibility Test:** Easily defensible as a targeted adjustment.

  - **Reversibility Test:** Positive, as it aims for improved fairness.
  - **Virtue Test:** Encourages fairness and continuous improvement.
  - **Colleague Test:** Positive response for fine-tuning existing processes.
  - **Professional Test:** Generally supported for fairness adjustments.
  - **Organization Test:** Aligns with a commitment to fair practices.
- **Iterative Testing and Redesign**
  - **Harm Test:** Low to moderate harm involves ongoing adjustments.
  - **Publicity Test:** Positive, showcasing commitment to continuous improvement.
  - **Defensibility Test:** Easily defensible, aligning with iterative testing.
  - **Reversibility Test:** Positive, enabling ongoing refinement.
  - **Virtue Test:** Fosters a culture of continuous improvement and scrutiny.
  - **Colleague Test:** Positive response for ongoing monitoring.
  - **Professional Test:** Generally supported for continuous improvement.
  - **Organization Test:** Aligns with a commitment to ongoing scrutiny and adjustment.
- **Algorithmic Bias Mitigation Tool**
  - **Harm Test:** Low harm, as implementing a bias mitigation tool is a proactive step.
  - **Publicity Test:** Positive, showcasing commitment to addressing biases.
  - **Defensibility Test:** Highly defensible, leveraging external tools for fairness.
  - **Reversibility Test:** Positive, as it facilitates ongoing bias mitigation.
  - **Virtue Test:** Demonstrates a commitment to responsible AI development.
  - **Colleague Test:** Positive response for using specialized tools for fairness.
  - **Professional Test:** Generally supported for leveraging external tools.
  - **Organization Test:** Aligns with the trend of adopting ethical AI practices.
- **Fairness-aware Machine Learning Frameworks**
  - **Harm Test:** Low harm, as it involves integrating fairness metrics.
  - **Publicity Test:** Positive, emphasizing transparency and fairness.
  - **Defensibility Test:** Easily defensible, using established fairness frameworks.
  - **Reversibility Test:** Positive, fostering ongoing fairness evaluation.
  - **Virtue Test:** Aligns with best practices for transparent AI development.
  - **Colleague Test:** Positive response to incorporating fairness frameworks.
  - **Professional Test:** Generally supported for transparency and fairness.
  - **Organization Test:** Aligns with industry standards for ethical AI.
- **Ethical AI Guidelines Integration**
  - **Harm Test:** Low harm involves embedding ethical considerations.
  - **Publicity Test:** Positive, emphasizing a commitment to ethical AI. **Defensibility Test:** Easily defensible, following established guidelines.
  - **Reversibility Test:** Positive, promoting ethical AI principles.
  - **Virtue Test:** Demonstrates a commitment to ethical AI development.
  - **Colleague Test:** Positive response for adhering to ethical guidelines.
  - **Professional Test:** Generally supported for ethical AI practices.
  - **Organization Test:** Aligns with responsible AI development.
- **Explainable AI Models**
  - **Harm Test:** Low harm involves enhancing transparency.
  - **Publicity Test:** Positive, emphasizing transparency in decision-making.
  - **Defensibility Test:** Easily defensible, using interpretable models.
  - **Reversibility Test:** Positive, facilitating understanding and adjustments.
  - **Virtue Test:** Encourages transparency and accountability.
  - **Colleague Test:** Positive response for using explainable models.
  - **Professional Test:** Generally supported for transparency in AI.
  - **Organization Test:** Aligns with a culture of transparency and accountability.
- **Re-evaluate Feature Selection Criteria**
  - **Harm Test:** Low harm involves revisiting criteria based on expert input.
  - **Publicity Test:** Positive, emphasizing a commitment to inclusivity.
  - **Defensibility Test:** Easily defensible, considering diverse factors.
  - **Reversibility Test:** Positive, fostering adaptability based on expertise.
  - **Virtue Test:** Encourages collaboration and inclusivity.
  - **Colleague Test:** Positive response for involving domain experts.
  - **Professional Test:** Generally supported for inclusive feature selection.

– **Organization Test:** Aligns with a commitment to diversity and fairness.

*6) **Tentative Choice Based on the Above Information:**:* After evaluating the options and considering the ethical implications, the most tentative choice would be to implement **Iterative Testing and Redesign**. This option involves low to moderate harm as it entails ongoing adjustments, showcases a commitment to continuous improvement, and is easily defensible. It aligns with the virtue of fostering a culture of continuous improvement and scrutiny.

This choice allows for ongoing monitoring of the software's impact on diverse candidate pools, regular assessment of the algorithm, gathering feedback and making adjustments accordingly. The iterative testing and redesign approach follows the principles of ethical AI development and ensures that any unintended biases are addressed while maintaining positive outcomes. The problem of unintentional bias in the hiring algorithm has not been fully resolved. While **Iterative Testing and Redesign** is a tentative choice, it is part of an ongoing process to address and mitigate the biases in the software. Continuous monitoring and refinement are necessary to ensure a fair and inclusive hiring process.

*7) **Final Choice**:* After a comprehensive review of the information and considering ethical implications, the final choice is to prioritize **Explainable AI Models**. This option involves low harm by enhancing transparency in decision-making, which is crucial for understanding and addressing biases. It aligns with the virtues of encouraging transparency and accountability.

Implementing explainable AI models, such as LIME, can help Sandra understand how the current algorithm makes decisions and identify potential sources of bias. This choice supports a culture of transparency and accountability in AI development.

1) **What could make it less likely you would have to make such a decision again?**
   Establishing a comprehensive framework for ethical AI development within the organization, including ongoing training on algorithmic bias and regular external audits, could reduce the likelihood of facing similar decisions in the future.
2) **What precautions can you take as an individual (announce policy on question change jobs, etc)?**
   As an individual, Sandra can proactively advocate for the integration of ethical AI guidelines into the organization's policies. This may involve collaborating with the HR department to develop and communicate clear policies regarding algorithmic fairness and diversity.
3) **What could you do to have more support next time (e.g., seek future allies on this issue)?**
   Sandra can work towards building alliances with key stakeholders within the organization, including members of the diversity and inclusion teams, to ensure a collective understanding and commitment to ethical AI practices. Seeking support from colleagues who share similar concerns can strengthen her position.
4) **What can you do to change the organization (e.g., suggest policy change at the next department meeting)?**
   Sandra can propose the establishment of a dedicated ethics committee or working group within the organization that focuses on monitoring and addressing ethical considerations in AI development. This committee can play a role in shaping and revising policies related to algorithmic fairness.
5) **What can you do to change larger society (e.g., work for a new statute or EPA regulation)?**
   Sandra can actively participate in industry conferences, engage with professional organizations like ASIST, and contribute to discussions on ethical AI at a broader societal level. Collaborating with external groups advocating for responsible AI practices can contribute to shaping industry standards and regulations.

III. Conclusion

In summary, Emporia's hiring algorithm presents a challenge with unintended biases, particularly in racial and socio-economic aspects. Sandra, as the HR head, is tasked with addressing these issues while maintaining the algorithm's effectiveness.

The problem is evident – biases in zip codes result in unfair treatment of minority candidates, posing legal risks and harming the company's reputation. Stakeholders, including Sandra, Timothy, and the broader Emporia community, require a solution that balances fairness and efficiency.

A collaborative approach involving HR, software, and diversity teams is essential. External audits and employee input provide valuable perspectives. Continuous adjustments and fairness-aware frameworks are crucial to meet evolving ethical standards.

The chosen solution should not only rectify biases but establish a foundation for responsible AI at Emporia. By addressing these challenges, Sandra and the team can transform this situation into an opportunity for positive change, reinforcing Emporia's commitment to fairness, diversity, and innovation in hiring. The lessons learned here can guide other organizations in navigating the intersection of technology and ethics in the workforce.

References

[1] Jason Ludwig and Kendall Darfler, *An Algorithm Discriminates* 2017.
[2] Barocas, Solon, and Andrew D. Selbst, *"Big Data's Disparate Impact."* California Law Review 104.3 (2016): 671-732.
[3] IBM Developer Staff *"AI Fairness 360 by IBM"* November 14, 2018.
[4] Google *"Fairness Indicators by Google"* May 3, 2023.
[5] Microsoft *"Responsible AI Toolbox Microsoft - https://github.com/microsoft/responsible-ai-toolbox"* 2023.
[6] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin *'"Why Should I Trust You?": Explaining the Predictions of Any Classifier'* August 9, 2016.