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](mailto:www.mvdixit@ncsu.edu)

1 ***Abstract*—This case study examines the ethical challenges faced**

2 **by Sandra, a software developer at Emporia, who discovers**

3 **unintended disparate impacts in a resume screening software**

4 **she designed. Despite reducing salesperson attrition, the software**

5 **exhibits bias favoring white applicants due to a flawed proximity**

6 **metric to Emporia stores. The case explores Sandra’s ethical**

7 **responsibility, professional and legal obligations at Emporia,**

8 **and Timothy’s reluctance to address the disparate impact issue.**

9 **It questions the use of objective criteria, proposing alternative**

10 **metrics for fair recommendations. Positioned within the context**

11 **of algorithmic ethics, the study emphasizes the need for non-**

12 **technical solutions and diverse expertise to address underlying**

13 **issues in hiring fairness. The case prompts reflection on the**

14 **ethical dimensions of software development and the potential**

15 **implications of biased algorithms for employment practices.**

16 I. Introduction

17 We’re looking at a challenging scenario in the rapidly

18 evolving fields of technology and employment. Software de-

19 veloper Sandra from Emporia created a tool to aid in the

20 hiring of more qualified salesmen. With a twist: minority job

21 seekers may experience some unforeseen issues as a result

22 of its success. This paper examines Sandra’s problems and

23 considers ethical issues while developing computer programs

24 that impact people’s jobs. We will discuss workplace policies,

25 ethical obligations of IT professionals like Sandra, and the

26 consequences of biased computer programs. We will unravel

27 the moral implications of employing software and algorithms

28 to make judgments or their effects on the actual world,

29 exploring the unforeseen problems that may arise, using the

30 Emporia case as an example.

31 mds

1. ***Identifying the non-ethical issues****:* The software’s hir- 42 ing algorithm created some immoral challenges in addition to 43 the obvious ethical ones, putting the organization in danger of 44 employment discrimination lawsuits. The Civil Rights Act of 45 1964’s disparate impact principle, which forbids employers 46 from using any employment practices that have an unwar- 47 ranted negative impact on members of protected classes, such 48 as women, people of color, or those from lower-income 49 backgrounds, may have been inadvertently broken by the 50 company. A smaller pool of candidates may be discouraged 51 from applying as a result of such policies’ lack of diversity, 52 which could have negative effects on the company’s reputation 53 among prospective applicants in addition to legal ramifications. 54
2. ***Identifying the stakeholders****:* The head of the Human 55 Resources (HR) department, Sandra, whose employment may 56 be in jeopardy as a result of her actions, and the firm Emporia, 57 along with its wider range of stakeholders, are the main 58 stakeholders involved in this circumstance. The choices made 59 about the algorithm may have serious repercussions, such 60 harming the company’s brand or causing losses in money. If 61 the algorithm is kept in place, it could cause legal problems, 62 criticism, and damage to the company’s reputation because of 63 bias. It might also result in Sandra and Timothy, the HR de- 64 partment head, being fired. However, adjusting the algorithm’s 65 parameters to account for bias may have an adverse influence 66 on how effective it is, possibly producing less than ideal 67 outcomes and making the issue that the algorithm was meant to 68 address—high sales department turnover rates—worse. More- 69 over, the existing algorithm unintentionally leads to the hiring 70 of white people or people who live close to the organiza- 71

tion with preference. Because of this unintentional prejudice,

32 November 16, 2023

72

worthy individuals who might not fit these requirements are 73

33 II. Case Study

suffering unfair outcomes. It is not the best practice to base 74

recruiting decisions solely on a candidate’s proximity to the

34 *1)* ***Identifying the problem****:* Sandra’s present dilemma is

35 on the inadvertent emergence of moral dilemmas throughout

36 Emporia’s employment process—a problem that the company

37 is unaware of. According to Sandra’s study, there is a major

38 imbalance. Even though 80% of applicants for sales roles

39 are Black and Latino, 92% of the recently employed staff

40 are White. The Principal Component Analysis (PCA) features

41 were the source of the algorithm’s racial prejudice.

75

company—this will omit qualified applicants who could live 76 further away. To ensure a fair and inclusive hiring process 77 where talent is prioritized over arbitrary considerations, it is 78 imperative to address this issue. 79

1. ***Possible Options or Alternatives****:* 80
   * **External Audit by Experts:** Propose conducting an 81

external audit of the software by independent experts 82

specializing in algorithmic bias and fairness. This audit 83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

can provide an unbiased evaluation of the software’s impact on diverse candidate pools and recommend adjust- ments 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 ap- proach 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 unin- tended biases while maintaining positive outcomes.
* **Algorithmic Bias Mitigation Tool:** Implement an algo- rithmic 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 re- sources 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 cur- | | 140 |
| rent algorithm makes decisions and identify potential | | 141 |
| sources of bias. Explainable AI models enable developers | | 142 |
| to interpret and debug complex algorithms, making it | | 143 |
| easier to identify and address biases[6]. | | 144 |
| *•* | **Re-evaluate Feature Selection Criteria:** Revisit the | | 145 |
|  | criteria used for feature selection, considering a broader | | 146 |
|  | set of factors beyond zip codes. Collaborate with domain | | 147 |
|  | experts and diversity teams to identify relevant features | | 148 |
|  | that better reflect qualifications and the potential for | | 149 |
|  | success. By reassessing feature selection criteria, Sandra | | 150 |
|  | can create a more inclusive model that avoids reliance on | | 151 |
|  | factors that may introduce bias. | | 152 |
| *5)* | ***Tests****:* | | 153 |
| *•* | **External Audit by Experts** | | 154 |
|  | **– Harm Test:** Moderate harm due to potential costs | | 155 |
|  | and time. | | 156 |
|  | **– Publicity Test:** Positive, demonstrating a commit- | | 157 |
|  | ment to transparency. | | 158 |
|  | **– Defensibility Test:** Highly defensible, relying on | | 159 |
|  | external experts. | | 160 |
|  | **– Reversibility Test:** Positive, as it aims for an unbi- | | 161 |
|  | ased evaluation. | | 162 |
|  | **– Virtue Test:** Shows a commitment to objectivity and | | 163 |
|  | improvement. | | 164 |
|  | **– Colleague Test:** Positive response for seeking exter- | | 165 |
|  | nal validation. | | 166 |
|  | **– Professional Test:** Generally supported as a rigorous | | 167 |
|  | approach. | | 168 |
|  | **– Organization Test:** May require explaining the need | | 169 |
|  | for external scrutiny. | | 170 |
| *•* | **Employee Input and Feedback** | | 171 |
| **– Harm Test:** Low harm, involving | | employees in | 172 |
| decision-making. | |  | 173 |
| **– Publicity Test:** Positive, emphasizing a democratic 174 | | | |

approach. 175

* + **Defensibility Test:** Easily defensible, showing a 176

commitment to employee voices. 177

* + **Reversibility Test:** Positive, demonstrating adapt- 178

ability based on feedback. 179

* + **Virtue Test:** Fosters a culture of listening and valu- 180

ing employee opinions. 181

* + **Colleague Test:** Positive response for employee en- 182

gagement. 183

* + **Professional Test:** It is generally supported as it 184

aligns with employee-centered ethics. 185

* + **Organization Test:** Supports a culture of openness 186

and improvement. 187

# Fairness-Adjusted Weighting 188

* + **Harm Test:** Low harm involves adjusting existing 189

criteria. 190

* + **Publicity Test:** Positive, as it demonstrates a com- 191

mitment to fairness. 192

* + **Defensibility Test:** Easily defensible as a targeted 193

adjustment. 194

195

196

197

198

199

200

201

202

203

204

* **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.
  + **Colleague Test:** Positive response to incorporating 250

fairness frameworks. 251

* + **Professional Test:** Generally supported for trans- 252

parency and fairness. 253

* + **Organization Test:** Aligns with industry standards 254

for ethical AI. 255

# Ethical AI Guidelines Integration 256

* + **Harm Test:** Low harm involves embedding ethical 257

considerations. 258

* + **Publicity Test:** Positive, emphasizing a commitment

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

# 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 re- finement.
  + **Virtue Test:** Fosters a culture of continuous im- provement and scrutiny.
  + **Colleague Test:** Positive response for ongoing mon- itoring.
  + **Professional Test:** Generally supported for continu- ous 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 respon- sible AI development.
  + **Colleague Test:** Positive response for using special- ized tools for fairness.
  + **Professional Test:** Generally supported for leverag- ing external tools.
  + **Organization Test:** Aligns with the trend of adopt- ing ethical AI practices.

# Fairness-aware Machine Learning Frameworks

* + **Harm Test:** Low harm, as it involves integrating fairness metrics.
  + **Publicity Test:** Positive, emphasizing transparency

259

to ethical AI. **Defensibility Test:** Easily defensible, 260

following established guidelines. 261

* + **Reversibility Test:** Positive, promoting ethical AI 262

principles. 263

* + **Virtue Test:** Demonstrates a commitment to ethical 264

AI development. 265

* + **Colleague Test:** Positive response for adhering to 266

ethical guidelines. 267

* + **Professional Test:** Generally supported for ethical 268

AI practices. 269

* + **Organization Test:** Aligns with responsible AI de- 270

velopment. 271

# Explainable AI Models 272

* + **Harm Test:** Low harm involves enhancing trans- 273

parency. 274

* + **Publicity Test:** Positive, emphasizing transparency 275

in decision-making. 276

* + **Defensibility Test:** Easily defensible, using inter- 277

pretable models. 278

* + **Reversibility Test:** Positive, facilitating understand- 279

ing and adjustments. 280

* + **Virtue Test:** Encourages transparency and account- 281

ability. 282

* + **Colleague Test:** Positive response for using explain- 283

able models. 284

* + **Professional Test:** Generally supported for trans- 285

parency in AI. 286

* + **Organization Test:** Aligns with a culture of trans- 287

parency and accountability. 288

# Re-evaluate Feature Selection Criteria 289

* + **Harm Test:** Low harm involves revisiting criteria 290

based on expert input. 291

* + **Publicity Test:** Positive, emphasizing a commitment 292

to inclusivity. 293

* + **Defensibility Test:** Easily defensible, considering 294

diverse factors. 295

* + **Reversibility Test:** Positive, fostering adaptability 296

|  |  |  |
| --- | --- | --- |
| 243 and fairness. | based on expertise. | 297 |
| 244 **– Defensibility Test:** Easily defensible, using estab- | **– Virtue Test:** Encourages collaboration and inclusiv- | 298 |
| 245 lished fairness frameworks. | ity. | 299 |
| 246 **– Reversibility Test:** Positive, fostering ongoing fair- | **– Colleague Test:** Positive response for involving do- | 300 |
| 247 ness evaluation. | main experts. | 301 |
| 248 **– Virtue Test:** Aligns with best practices for transpar- | **– Professional Test:** Generally supported for inclusive | 302 |
| 249 ent AI development. | feature selection. | 303 |

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

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

1. ***Tentative Choice Based on the Above Information:****:* After evaluating the options and considering the ethical im- plications, 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 defensi- ble. 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 accord- ingly. 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.

1. ***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.

# 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.

# 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 orga- nization’s policies. This may involve collaborating with the HR department to develop and communicate clear policies regarding algorithmic fairness and diversity.

# 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 mem- bers 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.

# What can you do to change the organization (e.g., 360 suggest policy change at the next department meet- 361 ing)? 362

Sandra can propose the establishment of a dedicated 363

ethics committee or working group within the organiza- 364

tion that focuses on monitoring and addressing ethical 365

considerations in AI development. This committee can 366

play a role in shaping and revising policies related to 367

algorithmic fairness. 368

# What can you do to change larger society (e.g., work 369

# for a new statute or EPA regulation)? 370

Sandra can actively participate in industry conferences, 371

engage with professional organizations like ASIST, and 372

contribute to discussions on ethical AI at a broader soci- 373

etal level. Collaborating with external groups advocating 374

for responsible AI practices can contribute to shaping 375

industry standards and regulations. 376

III. Conclusion 377

In summary, Emporia’s hiring algorithm presents a chal- 378 lenge with unintended biases, particularly in racial and socio- 379 economic aspects. Sandra, as the HR head, is tasked with 380

addressing these issues while maintaining the algorithm’s 381

effectiveness. 382

The problem is evident – biases in zip codes result in 383

unfair treatment of minority candidates, posing legal risks 384

and harming the company’s reputation. Stakeholders, including 385

Sandra, Timothy, and the broader Emporia community, require 386

a solution that balances fairness and efficiency. 387

A collaborative approach involving HR, software, and di- 388

versity teams is essential. External audits and employee input 389

provide valuable perspectives. Continuous adjustments and 390

fairness-aware frameworks are crucial to meet evolving ethical 391

standards. 392

The chosen solution should not only rectify biases but estab- 393

lish a foundation for responsible AI at Emporia. By addressing 394

these challenges, Sandra and the team can transform this 395

situation into an opportunity for positive change, reinforcing 396

Emporia’s commitment to fairness, diversity, and innovation in 397

hiring. The lessons learned here can guide other organizations 398

in navigating the intersection of technology and ethics in the 399

workforce. 400

References 401

1. Jason Ludwig and Kendall Darfler, *An Algorithm Discriminates* 2017. 402
2. Barocas, Solon, and Andrew D. Selbst, *”Big Data’s Disparate Impact.”* 403

California Law Review 104.3 (2016): 671-732. 404

1. IBM Developer Staff *”AI Fairness 360 by IBM”* November 14, 2018. 405
2. Google *”Fairness Indicators by Google”* May 3, 2023. 406
3. Microsoft *”Responsible AI Toolbox Microsoft -* 407

*https://github.com/microsoft/responsible-ai-toolbox”* 2023. 408

1. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin *‘”Why Should I* 409

*Trust You?”: Explaining the Predictions of Any Classifier‘* August 9, 410

2016. 411