PARTNERSHIP ON AI

Participatory and Inclusive Demographic Data Guidelines

Case Study

Use in conjunction with:

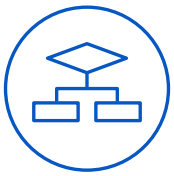
- Guidelines
- Implementation Workbook

The following description provides a hypothetical example in which an AI-developing and deploying organization conducts a fairness assessment of its algorithmic system following the Participatory and Inclusive Demographic Data Guidelines. This example is intended to better illustrate the implementation of the Guidelines for readers and is organized by the demographic data lifecycle phases.

Introduction

A US-based company built an algorithmic system that screens resumes to select candidates for an initial job interview. The system looks for specific keywords in resumes and then sorts applications into groups based on the results. It then matches qualified applicants to jobs for hiring managers to review. The system is trained on previously received resumes and does not collect any demographic information (i.e., race, ethnicity, gender, sexuality, etc.).

To ensure equitable product performance, the company decides to conduct an assessment to see if the system is operating fairly across demographic groups. To do so, they need to collect relevant demographic information from people impacted by the algorithmic system: in this case, job applicants who submit their resumes for review.



1 Planning & Design

Various internal actors are involved in designing the fairness assessment and demographic data collection process, including the Responsible AI team, the legal team (to ensure regulatory compliance during data collection efforts), the developers involved in building the resume screening system, and their product leads. Together, these internal actors will make up the fairness assessment team. Externally, the team assembles a compensated focus group of people who have previously submitted their resumes to the company and ten experts in issues related to discrimination and hiring practices (across a range of dimensions from racial and gender equity to disability rights). The team leading the fairness assessment obtains support from the company's leadership and robust and flexible funding. The company's leadership commits to stopping the use of the resume screening system if the assessment process finds the system to be fundamentally harmful or inequitable. In addition, all team members undergo training in participatory research methods and structural competency.

The team decides to work with the focus group and external experts to identify which communities may be adversely impacted by the resume screening system. They find that the system is most likely to discriminate against women, particularly women of color and Black women, given the long history of occupational segregation in the United States and that the company has very few women of color and Black women employees. The team then works to build relationships with relevant women and women of color advocacy

organizations (such as Black Girls Code or Until Freedom). It convenes to gather input from leadership and staff on the fairness assessment design process (see similar collaborations between the Danish Institute for Human Rights and technology companies on digital impact assessments). These workshops also help the team to clarify what demographic categories may be relevant to collect for the assessment.

2 Consent

After completing the Planning & Design phase, the team considers how they will collect relevant demographic information from users to assess their system for bias. The team works again with the advocacy organizations who participated in their workshop to get feedback on their proposed consent process and ensure the terms and format are accessible, particularly to women of color and Black women suspected to be most at risk of discrimination by the resume screening system. The team designs a variety of consent mechanisms for users to choose from, including a short video explainer and a form that can be re-administered throughout the fairness assessment process. They also design a mechanism that allows participants to revoke their consent at multiple points in the fairness assessment process. However, while they can remove participants' individual data points from the aggregate dataset at any point, they cannot delete the result of any analyses derived from the aggregate dataset. The team clearly communicates this limitation of consent revocation limitation during the opt-in process.

They recognize that the same group of people who are most likely experiencing harm from their system are also least likely to provide affirmative consent given the history of harm against women of color and Black women who participate in formal data collection processes in the US. To prevent having a non-representative sample, the team conducts additional outreach to these communities via the advocacy groups they worked with for the workshops. When they notice that they are still getting a lower opt-in rate among Latinx women, they deepen their outreach efforts among this community.

3 Collection

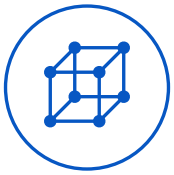
Now that the team has obtained affirmative, informed consent from users to collect their demographic information, they can begin actual collection to understand how the resume screener works across demographic groups. They collaborate with their focus group and equity experts to select multiple collection methods, including forms with pre-determined demographic categories and individual user interviews. With the help of their focus group, equity experts, and relevant advocacy organizations, the team identifies what relevant categories should be included in the survey to best capture data subjects' identities. This includes a list of racial and ethnic categories expanded from the US Census categories to include more granular ethnic categories for Asian Americans and Pacific Islanders

and gender categories that include non-binary, transgender, and gender non-conforming options. The team collects this demographic information from applicants. They also collect information regarding whether applicants were selected or rejected by the screener system. Throughout the data collection process, the team collects only the information they absolutely need (upholding the principle of data minimization). They also protect individual and group-level privacy during the collection process by applying data masking, which allows the data to be safely stored and used while maintaining the anonymity and safety of the applicants. Thanks to their extensive outreach efforts, the team is able to collect a reasonably representative sample of user demographic data. They rigorously validate the dataset for selection bias by comparing their dataset to other demographic information on applicants previously gathered by the company and findings from interviews and focus groups.

4 Pre-Processing

Now that the team has a dataset detailing a sample of applicants' racial, ethnic, and gender identities and whether or not the resume screener system selected them, they are ready to clean the dataset. As the team collected self-identified data, they don't need to work with an external party to annotate or label the dataset and instead conduct the Pre-Processing phase themselves. Due to the data masking applied during the Collection phase, the dataset is safely anonymized throughout the process.

After the pre-processing is complete, the team realizes they only gathered information from a small number of applicants who identified as gender non-conforming despite their extensive outreach to applicants during the Collection phase. The team recognizes that a small sample within a subgroup increases the risk of de-anonymization for the data subjects. They decide to handle this risk by deleting the data from this subgroup from the dataset that will be used for quantitative analysis and instead conducting focus groups and interviews with gender-nonconforming people. This will still allow the team to understand their experiences with the resume screener tool and identify vectors of discrimination experienced by this community.



5 Analysis

After the dataset is cleaned and made usable for analysis, the team again convenes a workshop with the external experts and advocacy groups they worked with during the previous Planning & Design phase to discuss the best analysis technique to assess for bias in the system (using the demographic dataset) and how bias will be defined. They also convene the focus group they previously assembled during the Planning & Design phase to assist in this process. After these consultations, the team decides that system fairness could be achieved when the system meets the conditions of demographic parity

and the interview findings reflect that applicants (particularly women of color and Black women applicants) experience fair treatment by the screener system. They plan to examine intersecting forms of discrimination (i.e. outcomes experienced by Black women vs. white men) in both the analysis of the dataset and via interviews. The team clearly documents all decisions made and steps taken during the analysis process, including their working definition of fairness, and shares this information with external experts, advocacy groups, and focus groups for their visibility.

Through the analysis and interviews, the team discovers that the screener system discriminates against Black women applicants. The analysis reveals that Black women applicants are rejected from the screener tool at a higher rate than white men applicants. Interviews and discussions with the focus group reveal that Black women applicants feel they are submitting high-quality resumes but are unable to get to the interview stage of the application process because they are rejected by the screener tool from the outset. Interviewees report that the general distrust of this system, based on previous experiences of discrimination with job-related algorithmic systems, has caused some Black women to not apply at all.

6 Bias Mitigation & Results-Sharing

Now that the team has the results from their fairness analysis, they begin the process of implementing their findings and sharing those steps with the organization, external stakeholders, and applicants. Based on input from the advocacy groups, focus groups, and external experts, the team decides to temporarily pause the use of the system until they can build a model that meets their agreed-upon definition of fairness. They recommend the organization hire additional HR personnel trained in equitable hiring practices, and have a particular sensitivity to the experiences of Black women applicants to assist in the resume screening process. The team produces an in-depth report, a short-form video, and a series of social media posts and press releases announcing how they arrived at this decision. They also set up a mechanism on their website that allows people to ask questions or get further information from the team about this decision. All feedback and steps taken to address this feedback are documented in the organization and shared publicly after anonymization.

7 Removal, Archive, & Destruction

Immediately following the end of the analysis, the team removes the dataset from all active environments and securely destroys all copies (as agreed upon by participants during the Consent phase). Back in the Planning & Design phase, the team built a tool on their website that allowed participants to indicate whether they'd like to revoke their consent and have the organization remove their information from the aggregate dataset. The tool also clearly communicated that depending on the timing of this request, the organization may be unable to delete the individual's data from the analysis findings. A couple of participants who initially opted into the collection process used the tool to indicate they want to be removed from the dataset. The organization deletes their individual data points from the aggregate dataset but notifies the participants that the analysis phase has already been completed so they cannot remove their data from the analysis findings.

8 Documentation & Data Governance

Throughout the data lifecycle, the team has dedicated time to ensuring that each of their steps and decisions are clearly and accessibly documented using the Partnership on AI Documentation Guidelines. A version of the documentation (abbreviated to ensure privacy) is publicly accessible to ensure transparency and allow for replicability. The team has also ensured that the dataset has been used in compliance with consent from the data subjects. They have also implemented strong data storage and privacy-preserving techniques (including data masking) to prevent data leakages and other privacy risks.