In her Shoes: Gendered Labelling in Crowdsourced Safety Perceptions Data from India

A streetview from New Delhi, India.

In recent years, there has been a proliferation of women's safety mobile applications in India that crowdsource street safety perceptions to generate 'safety maps' which are used by policy makers for urban design and academics for studying mobility patterns. However, men and women's differential access to information and communication technologies (ICTs), and the distinctions between their social and cultural subjective experiences may mitigate the value of crowdsourced safety perceptions data and the predictive ability of machine learning (ML) models utilizing such data. We explore this by collecting and analyzing primary data on safety perceptions from New Delhi, India. Our curated dataset consists of streetviews covering a wide range of neighborhoods for which we obtain subjective safety ratings from both female and male respondents. Simulation experiments where varying proportion of ratings from each gender are assumed missing demonstrate that the predictive ability of standard ML techniques relies crucially on the distribution of data producers. We find that obtaining large amounts of crowdsourced safety labels from male respondents for predicting female safety perceptions is inefficient in a number of scenarios and even undesirable in others. Detailed comparisons between female and male respondents' data demonstrate significant gender differences in safety perceptions and their associated vocabularies. Our results have important implications on the design of platforms relying on crowdsourced data and the insights generated from them.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: crowdsourced ratings, safety, gender, algorithmic bias, India

Author's address:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

ACM Reference Format:
. 2018. In her Shoes: Gendered Labelling in Crowdsourced Safety Perceptions Data from India. 1, 1 (February 2018), 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

India, with a population of over 1.3 billion and over 30 thousand reported rapes annually, ranks as one of the least safe countries for women in the world [2, 16]. Research indicates that roadsides are amongst the locations rated most unsafe by women from the point of view of sexual harassment [29, 30]. The horrific gang-rape and murder of a 23 year old female while travelling on a bus in December 2012 catapulted women's street safety issues to the center of attention for both citizens and policy makers. The increasing visibility of this issue led to public demand for better measurement and tracking of safety parameters on Indian streets.

In response to these demands, a number of mobile applications targeted towards women's safety have emerged in recent years. One of the tools used to capture the level of safety of a local context is crowdsourcing. For instance the popular Safetipin platform collects user-inputed annotations of safety perceptions and related factors on local streets. Another widely used platform, Safecity crowdsources locations where individuals experienced or observed incidents of sexual harassment. In 2015, the central government of India launched the Himmat (which translates to Courage) platform for women to report sexual harassment incidents that reportedly get relayed directly to police control rooms.

The aggregated information from such apps can be used to generate 'safety maps' of different neighborhoods and streets in the city which are shared with policy think tanks and government departments as an input for urban policy. Such safety ratings data have also been used in mainstream academic research as an input for modelling women's mobility decisions. For instance [5], using a combination of data sources including from Safetipin and Safecity, finds that young females in Delhi make a trade off between college quality and safety concerns that arise while travelling to and from educational institutions. She further finds that women are willing to pay more than men for transportation costs to allay such safety concerns.

The use of crowdsourcing of subjective safety data is not restricted to the Indian context. Around the world too, with the proliferation of mobile applications and cheaper data plans, the collection of crowdsourced safety perceptions has been growing in popularity in a variety of contexts. Applications range from the reporting of street harassment (for instance 'Safe and the City' which is headquartered in the United Kingdom and 'Hollaback!', an online community originating from New York City USA) to the reporting of police brutality (for instance the 'SafeSpace' app originating from Minneapolis USA whose tagline is 'police the police') and general neighborhood safety (the 'Citizen' app which claims over 9 million users across more than 60 cities, mainly across USA). In the context of academic research, crowdsourced street safety perceptions data collected via MIT Media Lab's 'PlacePulse' platform has been used to map urban perceptions in major cities across the globe [19].

From a computational perspective, the resulting datasets consisting of crowdsourced safety perceptions can be used to train machine learning models which predict 'safety ratings' in out-of-sample streetviews. For example, the 'Streetscore' model utilizes data from the 'PlacePulse' platform to predict street safety scores for a million cityscapes using a support vector machine model [19]. A subsequent version uses Deep Learning [8]. These models have since been used as inputs to study other socio-economic phenomena including analyzing the physical evolution of neighborhoods [18] and estimating their levels of gentrification [11].

There is by now a substantial literature on algorithmic and labelling bias across multiple contexts utilizing big data and machine learning (ML) models with the largest prediction errors typically being associated with minorities and

under-represented communities. For instance [4] analyze gender differences in the context of content moderation on online platforms. [6] evaluate bias present in automated facial analysis algorithms and find that the lowest prediction errors are associated with lighter male individuals and the highest predictions are associated with darker females. Via an ethnographic analysis of Delhi police's data collection practices [17] elaborate on the types of biases that can become embedded into a predictive policing setup. [14] find that women are far less likely to be shown a gender-neutral ad for STEM jobs on social media. Such biases however have not been widely studied in the context of street safety perceptions which are increasingly relying on crowdsourced data (see [25] for a notable exception).

In this paper we address the important research gap on algorithmic bias in crowdsourced safety data, while focusing on a Global South context, that of India. We argue that while perceptions about the safety of a particular location may be useful in deciding whether it is generally safe or unsafe, if such decision making is automated, then differing gender perceptions about safety need to be taken into account and that demographic differences need to be carefully considered during the process of data collection. If not, the resulting automation would be biased towards one or the other demographic segment.

Our methodology consists of obtaining safety perceptions ratings and related data from respondents on a fixed set of previously collected streetviews from Delhi. Participants in the survey consist of undergraduate and postgraduate students at a reputed engineering institute in New Delhi. Using the primary data thus collected, consisting of $11,200$ safety perception ratings data from 224 respondents over 50 streetviews each, we run extensive simulation experiments where varying proportion of labels from each gender are assumed missing. We compare predictive abilities of standard matrix factorization models from the ML literature and demonstrate that predictions of safety perceptions depend crucially on the distribution of the underlying (and observed) training data.

We find that obtaining large amounts of crowdsourced labels from male respondents has limited utility for predicting female safety perceptions and in certain situations more labels from male respondents *reduces* prediction accuracy for females. We also find significant gender differences in individual ratings (which are exacerbated for night-time images), response times and vocabularies used to describe visual safety. Overall our findings demonstrate the pitfalls of using crowdsourced data from mobile applications as an input to both public policy as well as empirical academic research.

Our findings have particularly important implications for public policy, especially given the increasing incorporation of crowdsourced data as well as algorithms into public policy processes [22]. We demonstrate that there is an urgent need to analyse potential bias in crowdsourced data [9], particularly from the perspective of the Global South which has received woefully less attention in the scientific discourse around algorithmic fairness [23]. This becomes crucial in economies such as India where access to Information and Communication Technologies (ICTs) is highly skewed towards the male population [1].

The remainder of the paper is organized as follows. We describe our research design, including data collection and analysis designs respectively in sections 2 and 3. Section 4 presents our detailed results and section 5 concludes with a discussion on policy implications and recommendations for platform design, along with directions for future research.

2 DATA COLLECTION

Our research design involved the collection of street safety perceptions from a sample of female and male participants over a set of fixed streetviews. In machine learning parlance we refer to the subjective safety ratings as "labels" and the streetviews for which ratings were elicited as "examples". Data collection for the study consisted of 1) building a repository of examples (streetviews) with visual and geographic variation and 2) collecting gender dis-aggregated labels (safety perception ratings) from a sample of female and male respondents.

2.1 Repository of Streetviews

Ratings for safety were collected for a set of images taken across various locations in New Delhi. These were images of streets, localities, highways, markets and residential areas which were collected using district-wise crime statistics available in the public domain[1]. District-wise crime statistics pertaining to cases of outdoor criminal activity including stalking, voyeurism, rape, kidnapping, acid attacks, robbery, auto-theft and others were aggregated and each district was ranked according to reported criminal activity. Using this ranking, four Delhi districts were shortlisted: North, North-East, West and New Delhi. While New Delhi and North were the 'most safe' based on the crime data, North-East and West were the 'least safe'. These four districts were also found to be accessible and located within Delhi (e.g. any 'Outer' districts or 'special zones' were excluded). When taking the images, unofficial information about the safest and most unsafe streets or localities in a district was solicited via informal conversations with locals including security and police personnel. Thus the set of locations were characterized as follows: 1) safer streets in the safest districts (labelled as 'safe-safe'), 2) more unsafe streets in the safest districts (labelled as 'unsafe-safe') , 3) safer streets in the most unsafe districts (labelled as 'safe-unsafe') and 4) more unsafe streets in the most unsafe districts (labelled as 'unsafe-unsafe').

Images were taken from these areas for our main survey, all during the day. The geo-location, date, time, district name and street name were recorded at the time of taking the image using the Open Data Kit software [2] which was installed on a tablet. The image quality of the resulting pictures was sometimes poor, hence identical pictures taken using a better quality camera at the same time were used in the survey. Out of the number of factors affecting female mobility that have been identified in the literature [15, 26], our study focuses on visual characterstics, particularly heterogenity in the built environment. Thus, while taking pictures, the guidelines was to take street-level views (no aerial shots or views from a height) and to the extent possible include a range of visual characteristics taken from the literature on pedestrian safety [13, 20, 28] such as buildings, walls, greenery, parks, traffic, bus-stops, markets, disarray, footpaths etc. A total of 50 images were curated for the main survey consisting of a range of spatial and visual characteristics (comprising of 14 'safe-safe', 12 'safe-unsafe', 12 'unsafe-safe' and 12 'unsafe-unsafe' streetviews). The list of localities from where streetviews were obtained is presented in Table 1 and the spatial distribution of the localities is presented in Figure 1. For a subsidiary analysis, we obtained streetviews of 20 additional locations during the day, as well as streetviews taken at the same location and perspective during night-time (i.e. about 2-3 hours after sunset). We refer to this subset of images as the Day/Night set.

2.2 Gender disaggregated safety perceptions

Participants for the survey were recruited from among undergraduate and postgraduate students at a reputed engineering institute in New Delhi and exclude anyone unable to read English, lacking access to the internet or having any visual impairment. From the list of respondents to an institute-wide email soliciting interested students, a randomly selected subset of 125 participants who self-reported as female and another 125 participants who self-reported as male were invited for the survey (no participant self-reported as "non-binary" or preferred not to answer). Of these a total of 106 participants who self-reported as female and 118 who self-reported as male took the final survey which was conducted online and designed using `jsPsych` software. A pilot version of the survey was first conducted before finalizing the set of questions for the main survey. A description of the different types of data that were collected in the study is presented next.

[1] https://ncrb.gov.in/en/node/3009
[2] https://getodk.org/

District Name	Safe Localities	Unsafe Localities
New Delhi (Safe)	Inner Circle C.P. Parliament Road Jantar Mantar	Ridge Road
North (Safe)	Rajpura Road Roop Nagar Road Shakti Nagar Authority Road	Geeta Colony Flyover Outer Ring Road ISBT Inderlok
West (Unsafe)	Mayapuri Narayana Vikas Puri	Harinagar Punjabi Bagh Tilak Nagar
North East (Unsafe)	Shastri Park Sonia Vihar Gokalpuri	New Usmanpur Bhajanpura Khajuri Khas

Table 1. List of localities from which streetviews were obtained for the main survey

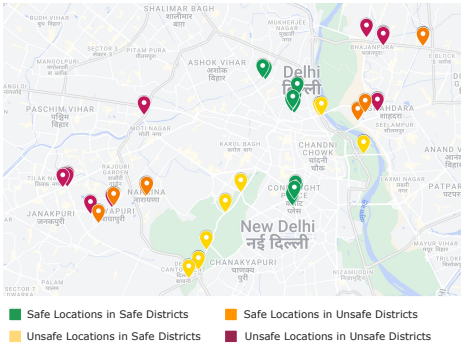

Fig. 1. Spatial distribution of localities from which streetviews were obtained for the main survey

- **Main survey:** Perceived safety ratings for the set of 50 curated images were collected on a 5 point Likert scale where 0 corresponded to 'very unsafe' and 4 to 'very safe'. Image order was randomized for each respondent to mitigate question ordering effects. Figure 2 presents a sample of safety ratings questions. The resulting data was used to generate a safety perceptions matrix whose rows represent the 224 respondents and columns represent the 50 streetviews. The entries in the matrix consisted of a total of $224 \times 50 = 11,200$ ratings.
- **Qualitative attributes:** Respondents were asked to provide a list of image attributes to justify their ratings for a subset of 10 streetviews each. This data was used to generate a qualitative dataset in the form of text data which was tagged by gender of respondent and their safety rating of the streetview.
- **Demographic Controls:** Along with the safety ratings, basic demographic information about the respondents including their gender, age, socioeconomic and educational background was also collected.
- **Meta Data:** The time taken to answer each question was recorded along with instances of participants exiting full screen and/or opening other tabs while taking the survey.
- **Day/Night survey:** Safety ratings were also collected for the set of 20 curated Day/Night streetviews. Every survey participant was randomly assigned to either the "Day" group or the "Night" group and was shown 20 additional images which were either taken during the day or during the night. As before, perceived safety ratings were collected on a 5 point Likert scale. The resulting data was used to generate 4 additional safety perceptions matrices – safety ratings from 53 female respondents in the 'Day' group (1,060 ratings), 53 female respondents in the 'Night' group (1,060 ratings), 63 male respondents in the 'Day' group (1,260 ratings) and from 55 male respondents in the 'Night' group (1,100 ratings).

3 DATA ANALYSIS

Using the main survey data, we first perform simulation experiments to analyze the predictive ability of standard ML models when different proportions of labels from each gender are assumed missing. We then contextualize the results of the simulations by analyzing gender differences in the associated quantitative and qualitative data from the survey. A detailed description of the methodology followed is presented next.

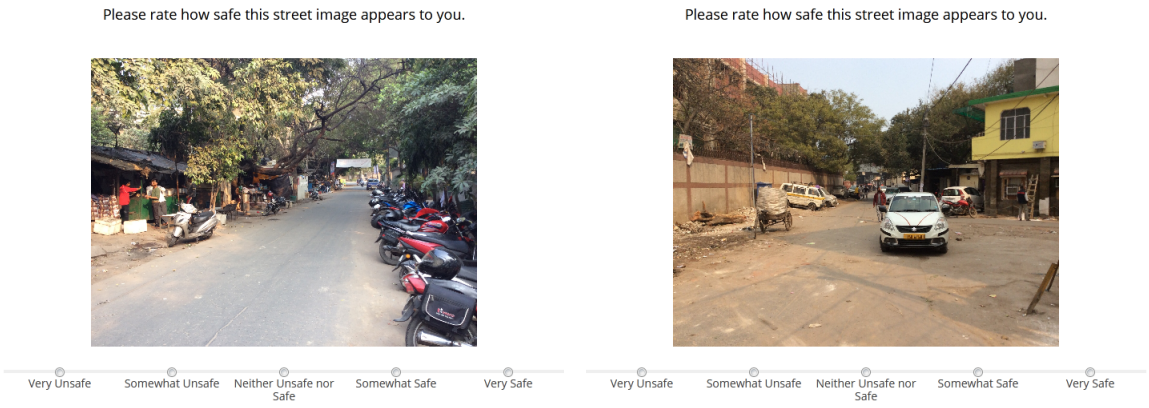

Fig. 2. Sample questions from the main survey.

3.1 Simulations

Drawing from the 11,200 safety perception ratings in our dataset, we simulate the downstream effects of gender imbalance in crowdsourced data on the prediction accuracy of standard ML models. To do so we utilize the safety ratings matrix (denoted X) corresponding to the main survey and carry out simulations in `R` using `softImpute`, a popular matrix factorization and missing values imputation package.

We fix a random subset of 25% female and male respondents each as 'test respondents' and the remaining respondents as 'train-sampling respondents'. $\rho_{miss,test}$ entries of the test respondents are assumed to be unobserved and constitute the test set. Results in the paper correspond to $\rho_{miss,test} = 0.50, 0.75, 0.90$[3]. We maintain a separate set of test respondents to ensure that changes in the predictive performance of matrix factorization algorithms are due to ratings elicited from other individuals and not due to observing more ratings of the test respondents themselves.

To simulate situations where crowdsourced data is not balanced between both genders, varying proportions of ratings from the train-sampling respondents are assumed to be observed, where the proportion of female train-sampling respondents varies from $\rho_{obs,female} = 10\%, 20\% \cdots 100\%$ and male train-sampling respondents varies from $\rho_{obs,male} = 10\%, 20\% \cdots 100\%$. Figure 3 presents a visualization of the simulated matrices thus generated.

Missing values in a single simulated matrix (denoted $\tilde{X}$) are imputed by nuclear norm minimization with optimally selected tuning parameters using `softimpute`. SoftImpute solves the following optimization problem

$$\min ||\tilde{X} - M||_o^2 + \lambda ||M||_*$$

where $||\cdot||_o$ refers to the Frobenius norm restricted to observed entries of $\tilde{X}$ and $||.||_*$ refers to the nuclear norm [12]. M is the imputed matrix and λ is the regularization tuning parameter. Mean Squared Error or MSE on the test set is computed as follows:

$$MSE = \frac{1}{|\Omega|} \sum_{i,j \in \Omega} \left(x_{ij} - m_{ij}\right)^2 .$$

where Ω refers to the entries in the test set, (i, j) refers to the row and column indices of these entries, x_{ij} refers to the true values from the complete matrix X and m_{ij} the imputed values from matrix M at these indices. 3 versions of Ω

[3]We do not consider situations where 100% of test respondent ratings are unobserved since matrix factorization algorithms require that every row has at least some observed entries.

are considered for each simulation, corresponding to a) all respondents, b) only male respondents and c) only female respondents in the test set.

A total of $N = 500$ simulations are run for each combination of $(\rho_{miss,test}; \rho_{obs,female}; \rho_{obs,male})$ values. The MSE values are averaged across simulations and compared via contour plots.

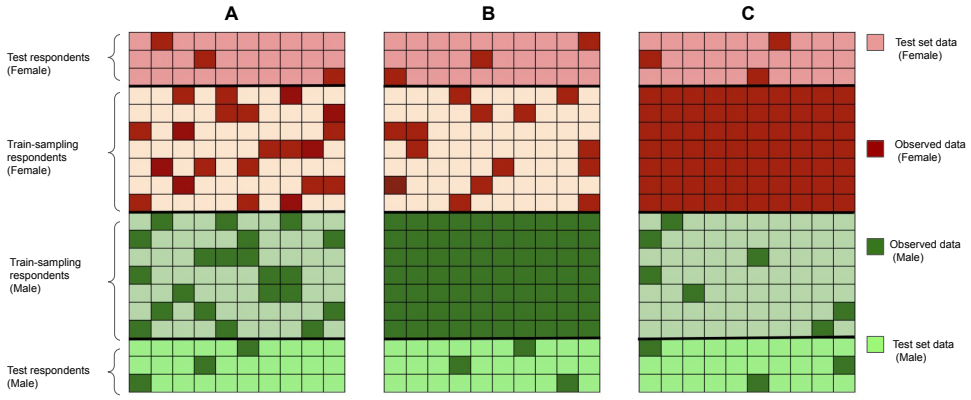

Fig. 3. Visualization of data generation in the simulations. The dataset X with fully observed ratings across all respondents is separated into a set of female test respondents, female train-sampling respondents, male test respondents and male train-sampling respondents. Dark colors represent the observed ratings in the simulated matrix $\tilde{X}$ and light colors represent ratings that are to be imputed. In each of the subfigures **A, B, C** $\rho_{miss,test} = 0.90$ i.e. 90% of ratings of the test respondents is assumed missing and constitute the test set. In subfigure **A**, $\rho_{obs,female} = \rho_{obs,male} = 0.30$ ie. 30% entries of both female and male train-sampling respondents is observed. In subfigure **B**, $\rho_{obs,female} = 0.20$ and $\rho_{obs,male} = 1$ and in subfigure **C**, $\rho_{obs,female} = 1$ and $\rho_{obs,male} = 0.1$

3.2 Quantitative Data Analysis

The following standard comparison of means t-tests were run to estimate gender differences in the main survey: a) gender differences in safety ratings by pooling all ratings in the main survey i.e. a single t-test comparing $106 \times 50 = 5300$ female ratings and $118 \times 50 = 5900$ male ratings, b) gender differences in safety ratings for each of the 50 images individually i.e. 50 t-tests each comparing 106 female ratings and 118 male ratings; c) gender differences in response times. From the Day/Night survey we compared the distributions of ratings from day-time and night-time images for each gender separately.

3.3 Qualitative Data Analysis

Text analysis of the qualitative attributes data was carried out in `python` using the package `nltk`. Textual attributes data was prepared for the anlaysis as follows: i) all words were converted to lower case and punctuation was removed, ii) typographical errors were removed by going through each set of words starting with a given alphabet and removing words starting with numbers, iii) stopwords were removed using the inbuilt list of English stopwords, iv) the data was tokenized and lemmatized. From the resulting data, the top attributes elicited for images rated 'very unsafe' and 'very safe' were compared for the two genders.

Simulations were then run to compare the overall distribution of words employed by female and male respondents. In the simulations, a set of 30 randomly selected female respondents was assigned as the reference group and two comparison groups consisting of 30 female and male respondents out of the remaining sample were assigned as the

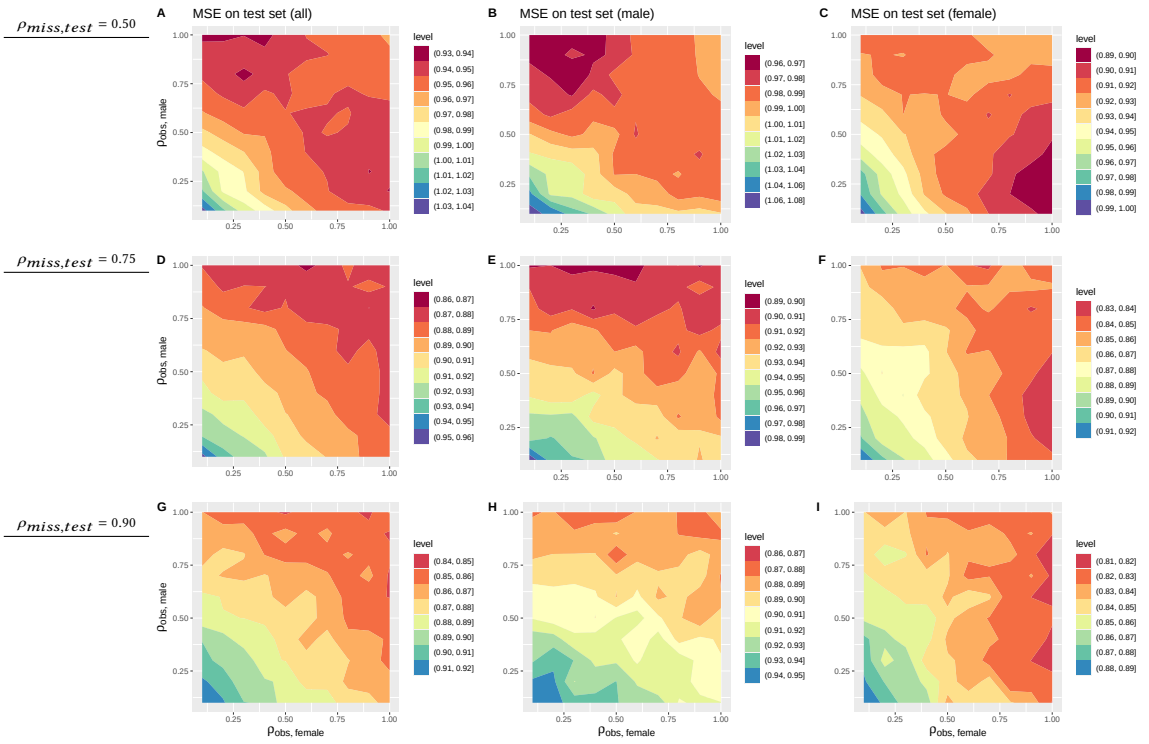

Fig. 4. Contour plots of MSE values. Columns correspond to MSE values for a) all respondents, b) only male respondents and c) only female respondents in the test set. Rows represent simulations where $\rho_{miss,test} = 0.50, 0.75, 0.90$ respectively ie increasing proportions of unobserved data within test respondents. Within each contour plot, the x-axis represents proportion of female train-sampling respondents' ratings observed $\rho_{obs,female}$ and y-axis represents the proportion of male train-sampling respondents' ratings observed $\rho_{obs,male}$.

female and male reference groups respectively. Kullback-Leibler or KL distance [3] between distributions was calculated between the reference group and each of the comparison groups. Means of the resulting KL distances from $N = 500$ simulations were compared.

4 RESULTS

4.1 Simulation Results

Figure 4 presents contour plots of MSE values on different test sets across the range of simulations considered. Columns represent a) all respondents, b) only male respondents only and c) only female respondents in the test set respectively. Rows represent increasing proportions of unobserved data of the test respondents $\rho_{miss,test} = 0.50, 0.75, 0.90$ respectively. Within each contour plot the x-axis corresponds to the proportion of female train-sampling respondents' ratings observed i.e. $\rho_{obs,female}$ and y-axis corresponds to the proportion of male train-sampling respondents' ratings observed i.e. $\rho_{obs,male}$. The lowest levels of MSE (or highest prediction accuracy) correspond to regions in the contour plots having darker shades of red.

When considering the top row (where $\rho_{miss,test} = 0.50$) we infer that the prediction accuracy on the test set can almost be decomposed into two gender specific components. From panel **C** we note that prediction accuracy for female respondents is highest when we observe a large proportion of female responses in the train-sampling data. Prediction accuracy even falls when a larger proportion of male labels in the train-sampling data is observed. Panel **B** shows a similar pattern for male respondents. This pattern however cannot be inferred from panel **A** where both female and male responses are pooled.

While less striking, similar patterns can be inferred from the second ($\rho_{miss,test} = 0.75$) and third ($\rho_{miss,test} = 0.90$) rows. In both cases MSE across all respondents appears to fall with any additional data from either gender, however if we split the MSE values by gender we find that MSE values for a specific gender fall more when we observe more ratings from that particular gender. This is demonstrated in the visualizations by the contour regions becoming more parallel to the $x - axis$ for MSE on test sets of male respondents and more parallel to the $y - axis$ for MSE on test sets of female respondents.

At low levels of observed ratings, additional ratings from either gender improve prediction accuracy. However as more ratings from a particular gender are observed, obtaining additional ratings from the other gender is at best inefficient and at worst undesirable. For instance in panels **F** and **I** we note that when $\rho_{obs,female} \approx 0.80$, obtaining more ratings from males doesn't reduce MSE values for the female test set. In these scenarios collecting more crowdsourced data from male respondents is inefficient.

More crucial however is the result that in certain cases, obtaining more ratings from male respondents actually *increases* the MSE values on the female test set. In both panels **C** and **F** we see that MSE values for the female test set is lowest when $\rho_{obs,female} = 100\%$ and $\rho_{obs,male} < 50\%$. This increases when $\rho_{obs,male} > 75\%$ indicating a situation where observing more data from male respondents *worsens* the predictive accuracy for out-of-sample female ratings.

These plots demonstrate that crowdsourced data collection can lead to algorithmic bias in terms of differing prediction accuracy for the two genders – large proportions of male data are associated with better prediction accuracy for males and may even *worsen* predictions for females and vice versa. This result is crucially important within this particular context i.e. when data are collected with the express purpose of benefiting a particular group – as is the case for safety perceptions data in India for women's safety measurements. As a bare minimum therefore it is essential to have a metric that captures the underlying gender mix of crowdsourced data in such contexts before proceeding with any further downstream use of the data.

4.2 Gender differences in quantitative data.

Figure 5(a) plots the distribution of female and male ratings across all images. While the distributions appear to be overlapping, a comparisons of means t-test estimates that average female safety perception ratings are about 0.12 lower than average male ratings (statistically significant at $\alpha < 0.0001$) on a 5 point scale. Mean response times by gender also differ – average response time corresponding to a single rating for females is about 1.64 seconds higher than for males (statistically significant at $\alpha < 0.0001$) indicating that the cognitive load and decision processes for females when making subjective safety perceptions is more involved. The distribution of median response times by gender is plotted in Figure 5(b).

On running comparison of means t-tests for each image separately we find significant differences in average ratings by gender for 11 of the 50 images – and for each of these images the average female rating is *lower*. In other words, whenever there are significant differences between the two genders, females have lower safety perceptions ratings on

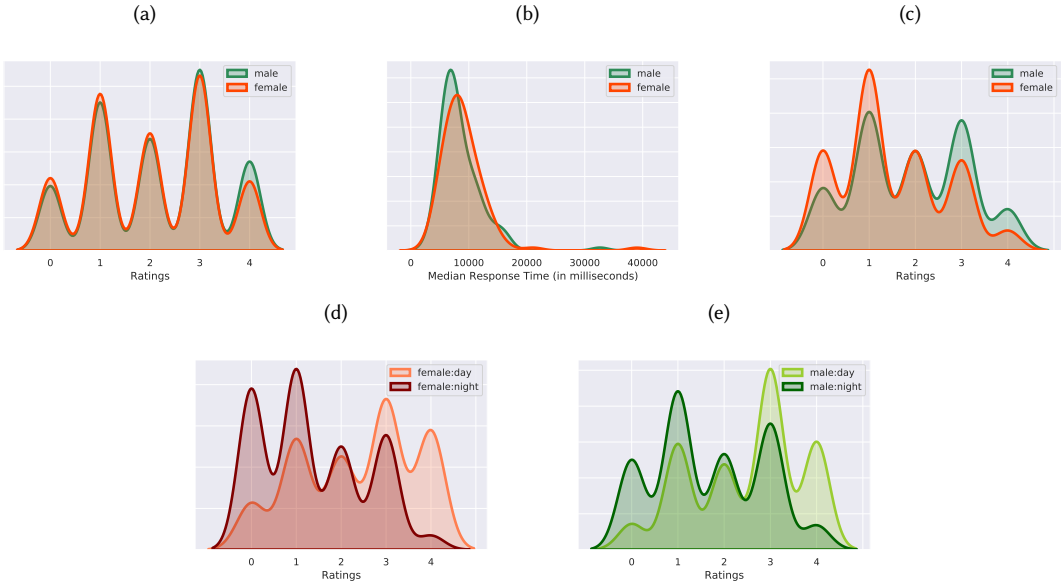

Fig. 5. Gender Differences in quantitative data. Subfigure (a) plots the distributions of ratings across all images in the main survey, (b) plots the distributions of median response times by gender, (c) plots the distributions of ratings for only those images where there is a significant gender difference in ratings. Subfigures (d) and (e) plot distributions for the Day/Night survey for female and male respondents respectively.

average. Figure 5 (c) plots the distribution of female and male ratings across only these 11 images – demonstrating that the distributions differ visually too, unlike in the case where ratings for all streetviews was pooled.

Figure 5 (d) and (e) plot the distribution of safety ratings from the Day/Night survey for female and male respondents respectively. These plots represent ratings given to a set of 20 streetviews taken from the same perspective and same location at both day-time (lighter shades) and night-time (darker shades). Comparing the two plots we note that for both female and male respondents the distribution of safety ratings shifts left (i.e. towards being more unsafe) for night-time images. For both genders, the largest mass in the distribution of safety ratings for day-time images is on Rating 3 (somewhat safe) and 4 (very safe). The leftward shift in night-time images however is more pronounced for female respondents with the largest mass for the night-time ratings being on Rating level 1 (somewhat unsafe) and 0 (very unsafe). For male respondents there is a visibly lower mass at the lowest safety rating (very unsafe). The distribution of the day/night survey ratings hints that the gender bias in predictions we saw in the simulations (which were conducted on a set of day-time images) may get exacerbated further if we were to extend our analysis to include night-time images.

A visual inspection of the 11 streetviews from the main survey where female ratings were significantly lower than male ratings on average can be used to tease out attributes for which female respondents may have greater negative associations with safety (Figure 6). Walls or fences appear in multiple photos, as do parked and/or abandoned cars. High speed roads and disarray are other attributes that are potentially important. Another attribute that is common across all of these images is the lack of passersby. When present, all but one of the passersby in the frames is male, pointing to the importance of gender mix of a scene as an important predictor of visual safety [21]. In the next subsection we analyze gender differences in explicitly elicited safety attributes from respondents.

Fig. 6. Streetviews where mean female safety perceptions are significantly lower than mean male safety perceptions.

4.3 Gender differences in qualitative data.

An analysis of textual data by gender reveals that female and male respondents value different attributes while making subjective safety perception ratings (Table 2). The majority of the top 10 attributes associated with the lowest levels of perceived safety ('very unsafe') elicited from female respondents' demonstrates that they have a much richer lexicon when it comes to describing scenes which are lacking in people: '[absence of] people', 'deserted', 'isolated', 'lonely' and 'secluded'. In the list for males, only 'lonely' appears lower down the ranking. The global literature on urban safety frequently refers to the importance 'eyes on the street' [10] – our results verify that this is a predominant concern for females in the Indian context. Another visual attribute frequently cited by females, 'wall' does not show up at all in the top 30 ranked attributes by males. Interestingly, on the other hand, the [presence of an] 'animal' is frequently cited by male respondents but does not show up in the top 30 ranked attributes by females. This analysis indicates that female and male respondents have very different mental models of the lack of safety or potential danger.

Very Unsafe (Rating 0)		Very Safe (Rating 4)	
Female	Male	Female	Male
'[absence of] people' (24)	'vehicle' (27)	'[presence of] people' (34)	'[presence of] people' (50)
'deserted' (17)	'narrow' (16)	'residential' (20)	'police' (17)
'isolated', 'narrow' (14)	'car', 'garbage' (14)	'crowded' (17)	'park' (15)
'garbage', 'wall' (13)	'animal' (13)	'open' (13)	'car' (14)
'vehicle' (12)	'parked' (12)	'busy', 'park' (12)	'locality', 'residential', 'clean' (13)
'lonely', 'secluded' (11)	'space', 'dirty' (11)	'traffic', 'vehicle', 'police' (11)	'vehicle', 'public' (11)
'small' (10)	'[car] parking', 'lonely', 'walking' (10)	'station' (9)	'society' (10)

Table 2. Top 10 attributes corresponding to streetviews rated 'very unsafe' and 'very safe' by gender.

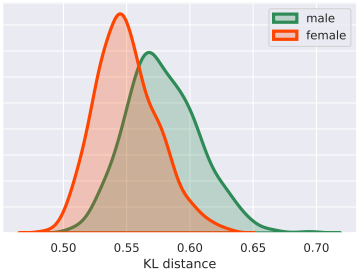

Fig. 7. Distributions of KL distances of vocabularies for female and male comparison groups with a female reference group across $N = 500$ simulations

There is more overlap between genders in the top attributes corresponding to the highest levels of perceived safety ('very safe'). The most common of these is expectedly the same for both genders: '[presence of] people'. Another common characteristic is streets that appear to be within a 'residential' 'society'. The ranking of the other top attributes indicates that female respondents are more preoccupied by crowds and openness whereas male respondents are more preoccupied by the presence of authority such as police and by cleanliness.

In order to go beyond comparing only the most frequently elicited attributes, we also run simulation experiments with textual data to compare the distributions of overall vocabularies associated with safety by gender. Our simulation experiments reveal that, when compared to a random reference subset of female respondents, the distribution of words employed by another random subset of female respondents is significantly closer ($\alpha < 0.0001$) than that of a random subset of male respondents. We measure these distances between distributions by computing their KL distances. Figure 7 plots the distributions of KL distances for female and male comparison groups across $N = 500$ simulations. Overall the simulations provide further evidence that the two genders utilize systematically different vocabularies when describing safety perceptions.

5 DISCUSSION

Intuitively we would expect that predictions and measurements made for a specific population should be drawn primarily from data that is representative of that population. Despite this, a number of crowdsourced applications do not explicitly require disclosure of respondent characteristics. The crowdsourced data thus generated may then be used in downstream ML models which can in turn influence public perceptions and policies, encoding the biases in data collection into the policy making exercise [7]. [24] refer to such situations as part of the problem of 'data cascades' or

'compounding events causing negative, downstream effects from data issues' which occur due to a lack of emphasis of data quality in ML models. In this paper we contextualize potential data cascades from the use of crowdsourced safety perceptions data for predicting measures of women's safety.

The main result in this paper demonstrates via extensive simulations that obtaining large amounts of crowdsourced labels from male respondents for predicting female safety perceptions is likely to be an inefficient way to collect data, and in some cases may even be undesirable. We find that the highest levels of prediction accuracy for safety ratings on a test set of female respondents corresponds to large number of ratings from other female respondents. In many cases adding more crowdsourced labels from male respondents does not lead to any gains in prediction accuracy and importantly in certain cases, having a very high proportion of male respondents' ratings actually leads to *worse* prediction accuracy for women.

We contextualize and unpack the results of our simulations by doing a detailed investigation of the ratings data and find that a) female respondents on average systematically rate streets as less safe than male respondents, b) female respondents on average systematically take longer to make subjective safety ratings hinting at a more complicated mental model of safety, c) female and male respondents utilize significantly different vocabularies when describing safe and unsafe locations, d) the underlying notion of safety itself is likely to be gendered. Additionally our analysis indicates that differences between genders are likely to be exacerbated further when considering night-time images.

Based on our analysis, we caution both researchers and policy makers against the widespread use of data generated via such crowdsourced data in downstream models without first ensuring that the context of data collection matches the context within which the models are applied. Consider the potential use-case of developing an algorithm that generates 'safe-routes' from point A to point B. If the data fed into the algorithm is crowdsourced predominantly from male respondents then it is likely that the resulting routes do not capture aspects of safety that females value.

We recommend the following design choices for platforms utilizing crowdsourced data: 1) explicitly collect basic demographic information from respondents; 2) crowdsource from specific sub-groups towards whom the platform is targetted and when appropriate predict for those specific sub-groups only and 3) in the absence of representative data apply appropriate weights to crowdsourced data to mitigate embedded bias in the system to the extent possible.

The following directions for future research emerge. First, similar empirical exercises in other geographic contexts where crowdsourced safety perceptions have become more mainstream is imperative, specifically United States where there is a proliferation of applications that crowdsource safety data and where *race* is likely to be an important moderating factor. Second, within India too, other dimensions beyond gender which can lead to crowdsourcing biases such as *caste* and *economic class* merit investigation. Third, case studies of the potential real world consequences of mainstreaming models built on biased datasets should be conducted: for instance, the possibility of safety score becoming a proxy for sensitive variables such as race, caste, religion or gender. Finally, a thorough investigation of what the term 'safety' means in different contexts for different populations is warranted. This will contribute to our understanding of the limits (and opportunities) of utilizing ML models for predicting subjective values.

To our knowledge, our study is among the first attempts to empirically measure potential algorithmic bias in the context of crowdsourced safety ratings using real world data. In doing so we employ a mixed methods approach, combining simulations using ML models, standard statistical comparisons and an analysis of qualitative attributes data. The paper also turns the spotlight on algorithmic bias questions from the perspective of the Global South. This is important, especially given that the majority of empirical evaluations of algorithmic and machine learning systems utilize datasets from the US and Europe, despite the fact that many countries in the Global South are now witnessing massive deployment of AI/ML products [27, 31]. We hope that with this piece helps to shift the discourse towards more

empirically rigorous evaluations of algorithmic systems utilizing a multi-disciplinary lens, particularly in the Global South.

REFERENCES

[1] Amy Antonio and David Tuffley. 2014. The gender digital divide in developing countries. *Future Internet* 6, 4 (2014), 673–687.

[2] Rituparna Bhattacharyya. 2016. Street violence against women in India: Mapping prevention strategies. *Asian Social Work and Policy Review* 10, 3 (2016), 311–325.

[3] Brigitte Bigi. 2003. Using Kullback-Leibler distance for text categorization. In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings*. Springer, 305–319.

[4] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*. Springer, 405–415.

[5] Girija Borker et al. 2021. *Safety first: Perceived risk of street harassment and educational choices of women*. World Bank.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[7] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[8] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 196–212.

[9] Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 162–170.

[10] Nicholas R Fyfe and Jon Bannister. 1998. THE EYES UPON THE STREET'. *Images of the street: planning, identity, and control in public space* 254 (1998).

[11] Edward L. Glaeser, Hyunjin Kim, and Michael Luca. 2018. Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change. *AEA Papers and Proceedings* 108 (May 2018), 77–82. https://doi.org/10.1257/pandp.20181034

[12] Trevor Hastie, Rahul Mazumder, and Maintainer Trevor Hastie. 2013. Package 'softImpute'.

[13] Jagori. 2010. Safe Cities Free of Violence against Women and Girls Initiative: Report of the Baseline Survey Delhi.

[14] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science* 65, 7 (2019), 2966–2981.

[15] Robin Law. 1999. Beyond 'women and transport': towards new geographies of gender and daily mobility. *Progress in human geography* 23, 4 (1999), 567–588.

[16] RN Mangoli and Ganapati N Tarase. 2009. Crime against women in India: A statistical review. *International Journal of Criminology and Sociological Theory* 2, 2 (2009).

[17] Vidushi Marda and Shivangi Narayan. 2020. Data in New Delhi's Predictive Policing System. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 317–324. https://doi.org/10.1145/3351095.3372865

[18] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and César A. Hidalgo. 2017. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences* 114, 29 (2017), 7571–7576. https://doi.org/10.1073/pnas.1619003114 arXiv:https://www.pnas.org/content/114/29/7571.full.pdf

[19] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore – Predicting the Perceived Safety of One Million Streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 793–799. https://doi.org/10.1109/CVPRW.2014.121

[20] Mangai Natarajan. 2016. Rapid assessment of "eve teasing"(sexual harassment) of young women during the commute to college In India. *Crime Science* 5, 1 (2016), 1–11.

[21] Jason Patch. 2008. "Ladies and gentrification": New stores, residents, and relationships in neighborhood change. In *Gender in an urban world*. Emerald Group Publishing Limited.

[22] Matthew J Salganik and Karen EC Levy. 2015. Wiki surveys: Open and quantifiable social data collection. *PloS one* 10, 5 (2015), e0123483.

[23] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 315–328.

[24] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[25] Reka Solymosi, Kate J Bowers, and Taku Fujiyama. 2018. Crowdsourcing subjective perceptions of neighbourhood disorder: Interpreting bias in open data. *The British Journal of Criminology* 58, 4 (2018), 944–967.

[26] Tanu Priya Uteng. 2012. Gender and mobility in the developing world. (2012).

[27] Shashi Shekhar Vempati. 2016. *India and the artificial intelligence revolution*. Vol. 1. JSTOR.

[28] Kalpana Viswanath and Ashish Basu. 2015. SafetiPin: an innovative mobile app to collect data on women's safety in Indian cities. *Gender & Development* 23, 1 (2015), 45–60.
[29] Kalpana Viswanath and Surabhi Tandon Mehrotra. 2007. 'Shall we go out?'Women's safety in public spaces in delhi. *Economic and political weekly* (2007), 1542–1548.
[30] Jagori & UN Women. 2011. Safe cities free of violence against women and girls initiative: Report of the baseline survey Delhi 2010.
[31] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. 2021. The AI index 2021 annual report. *arXiv preprint arXiv:2103.06312* (2021).

Received ; revised ; accepted