AI and Ethics
https://doi.org/10.1007/s43681-024-00572-w

ORIGINAL RESEARCH

Uncovering labeler bias in machine learning annotation tasks

Luke Haliburton[1,2] · Jan Leusmann[1] · Robin Welsch[3] · Sinksar Ghebremedhin[1] · Petros Isaakidis[1] · Albrecht Schmidt[1,2] · Sven Mayer[1,2]

Received: 15 April 2024 / Accepted: 24 August 2024

Abstract

As artificial intelligence becomes increasingly pervasive, it is essential that we understand the implications of bias in machine learning. Many developers rely on crowd workers to generate and annotate datasets for machine learning applications. However, this step risks embedding training data with labeler bias, leading to biased decision-making in systems trained on these datasets. To characterize labeler bias, we created a face dataset and conducted two studies where labelers of different ethnicity and sex completed annotation tasks. In the first study, labelers annotated subjective characteristics of faces. In the second, they annotated images using bounding boxes. Our results demonstrate that labeler demographics significantly impact both subjective and accuracy-based annotations, indicating that collecting a diverse set of labelers may not be enough to solve the problem. We discuss the consequences of these findings for current machine learning practices to create fair and unbiased systems.

Keywords Labeler bias · Machine learning · Bias · Crowd workers · Annotation

1 Introduction

With the rapidly increasing prevalence of artificial intelligence (AI), it is more important than ever to understand how bias is embedded in intelligent systems. Data annotation, typically conducted by crowd workers, is a crucial step in generating labeled training data for most contemporary AI models. However, datasets embed the cognitive biases of crowd workers [1], which can propagate through machine learning (ML) systems [2]. Most data annotation processes do not consider *who* is doing the labeling [3], and crowd worker platforms have non-representative demographics [4, 5], which bears the risk of creating training datasets based on input from biased populations of labelers, consequently creating biased systems. We refer to this category of bias, that is, any bias embedded by labelers when annotating data as *labeler bias*. With recent developments in generative models, the issue has only been elevated, c.f., Abid et al. [6]. Thus, understanding labeler bias is an open problem crucial to the development of fair AI systems.

Researchers have attempted to characterize [7–9] and correct for [10–12] labeler bias in multiple ML domains. For instance, prior research has demonstrated evidence for the existence of labeler bias in Natural Language Processing (NLP) annotation [13] and labeling toxic texts [8, 14]. A recent study characterized labeler bias in terms of the Stereotype Content Model (SCM) and showed that annotations vary depending on ethnicity and stereotypes held by the labelers [15]. However, their work used the FairFace dataset [16], which does not contain self-described ethnicity and sex labels but triple-coded ones. The task targeted subjective secondary

✉ Luke Haliburton
luke.haliburton@ifi.lmu.de

Jan Leusmann
jan.leusmann@ifi.lmu.de

Robin Welsch
robin.welsch@aalto.fi

Sinksar Ghebremedhin
sinksar.ghebremedhin@ifi.lmu.de

Petros Isaakidis
petros.isaakidis@ifi.lmu.de

Albrecht Schmidt
albrecht.schmidt@ifi.lmu.de

Sven Mayer
sven.mayer@ifi.lmu.de

[1] LMU Munich, Munich, Germany

[2] Munich Center for Machine Learning (MCML), Munich, Germany

[3] Aalto University, Helsinki, Finland

Published online: 16 September 2024

characteristics that are not in the picture (e.g., estimates of income), which past work called into question for facial recognition tasks [17]. As such, there remains a need to investigate face-labeling tasks with knowledge of the ground truth and, further, to characterize labeler bias in more objective tasks.

In this paper, we investigated labeler bias in two different labeling tasks. In the first phase, we investigated a secondary labeling task, reproducing the study by Haliburton et al. [15] with ground truth labels to verify their results. For this, we created a new face dataset where the subjects provided their own demographic information, addressing a key limitation of other existing datasets, e.g., the FairFace dataset [16]. We recruited 98 participants from seven ethnicities and two sexes to label the portraits for income and recorded their stereotype perceptions via perceived warmth, competence, status, and competition ratings. These labels represent secondary characteristics (i.e., the labels can be inferred from the image but do not appear in the image). We then conducted a second study investigating whether labeling bias also exists in a primary labeling task (i.e., the information to be labeled directly appears in the image). We recruited 210 participants to complete a bounding box task using ten images from the Waymo Open Dataset [18]. These tasks are intentionally quite different, one involves faces and potential interactions with the SCM, while the other involves labeling objects that would not typically be associated with stereotypes. In this manner, we aim to investigate whether there are multiple types of bias that contribute to labeler bias in ML.

Our results demonstrate that labelers exhibit bias in both primary and secondary annotation tasks. Through our first study, we found that stereotype variables and income estimations vary with labeler and portrait ethnicity and sex in a face-labeling task. Our second study demonstrated that bounding box accuracy varies with labeler ethnicity. Our results indicate that labeler bias stems from both characteristics of labelers and stereotype bias depending on the labeling task. These results motivate a critical reworking of the standard labeling process, including collecting demographic information from labelers to ensure a diverse labeling pool. This paper makes three concrete contributions: (1) a publicly available dataset of 56 face images with subject-labeled demographic information, (2) a characterization of labeler bias in a secondary face-labeling task, and (3) a characterization of labeler bias in an objective, primary bounding box task.

2 Related work

This section highlights research on crowdsourcing in ML and discusses labeler bias. We then introduce the SCM and its relation to Human–Computer interaction (HCI).

2.1 Crowdsourcing in machine learning

Many ML systems rely on crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) to recruit large numbers of participants to label information in datasets [19], primarily because crowdsourcing cost-effective and scalable [20]. Past work has shown that assembling labels from untrained crowd workers is comparable to an expert labeler for certain tasks [21–23]. Due to the unsupervised nature of crowdsourced annotations and the prevalence of spam submissions, many researchers have investigated quality control methods [24], including designing questions to control cheating [25] and determined topic affinity from social media profiles [26]. Several researchers have also investigated cognitive biases in crowd workers [1, 27]. Population studies have found that most MTurk workers are from India and the United States [28–30] and that over two-thirds of MTurk workers identify as white [5]. This lack of diversity risks embedding biases into the resulting datasets, which are then used to train ML models.

Crowdsourced labels are a relevant and evolving topic. With the growth in quality control methods, past work has identified that additional documentation can increase user trust in AI systems as long as they do not show bias [31]. Recent research has investigated new paradigms for generating crowdsourced labels using tags [32] and hierarchies [33]. Prior work has also designed and evaluated novel interfaces to increase crowd worker annotation speeds [34]. Importantly, there is evidence that both task and interface design can have a significant impact on quality [35–37]. We aim to add to this body of research by investigating how different estimates vary with labeler demographics.

2.2 Labeler bias in machine learning

Our work focuses on bias introduced by labelers who annotate datasets for ML applications. Prior research has made attempts to characterize the existence of labeler bias. Hube et al. [38] showed that even highly-experienced labelers make biased annotations. In the domain of toxic language, past work has found that labeler identity influences how an individual rates the toxicity of a text [8, 14]. This includes both demographic characteristics (e.g., ethnicity [8]) and social tendencies (e.g., conservatism [14]). Labelers also embed bias in their annotations based on socio-economic conditions and power relationships in labeling companies [7]. Rather than blaming labelers, past research has called for an increased investigation into the social and power dynamics present in annotation employment [3] and for increased consideration of economic conditions and fair compensation [37]. In some cases, biased labels are caused by the instructions given to labelers [36, 39]. Together, these works show that labeler bias is complex and important but not yet fully understood. We aim to add to this body of work by investigating how the demographic characteristics of labelers impact their annotations in ML labeling tasks.

Beyond understanding the impact, prior work has also worked to correct for labeler bias. Geva et al. [13] recommend recruiting separate groups of training and testing labelers because natural language processing (NLP) annotations do not generalize across groups. Past research has modeled labeler bias using multi-task Gaussian Processes [12], Bayesian methods [11], and ground truth knowledge [10]. However, there is currently no industry standard method to account for bias introduced in the labeling phase.

2.3 The stereotype content model

The SCM is a psychological theory that explains how individuals develop stereotypes about others. The SCM posits that people assess others based on their perceived warmth and competence [40, 41]. This theory is well-established in social psychology and is commonly employed in research on social perceptions and interactions [42–44]. In the warmth-competence model, warmth is an evaluation of perceived friendliness, while competence is a measure of how capable a person is perceived to be. The SCM claims that, in general, we react to people based on how warm or competent they appear. An individual who is both warm and competent, for example, is admired, while a warm but incompetent person is pitied [40]. As an extension of the warmth-competence model, the SCM can also include competition and status elements to help explain how people in one group react to individuals from another group [40]. In this paper, we use all four dimensions to assess the stereotypes of labelers.

In HCI, the SCM has been successfully used to investigate how people react to various stimuli such as digital avatars [45], to understand social acceptability for mobile devices [46], and to characterize stereotypical portrayals in personas [47]. Past work has also used the SCM to tackle stereotypical language with anti-stereotypes [48] and to detect stereotypes in the news [49]. In line with Haliburton et al. [15], we apply the SCM to characterize the stereotypical perceptions of labelers in a face-labeling task.

3 Methodology

This paper presents two distinct phases that, together, form an investigation into labeler bias. The overall structure of our methodology is outlined in Fig. 1. In Phase 1, we first created a dataset of face images with subject-labeled demographic information. We then re-created the face-labeling task from Haliburton et al. [15] with our new face image dataset, thereby addressing the primary limitations of their investigation. In Phase 2, we evaluated a bounding box task to understand whether bias persists in primary tasks that do not include labeling human faces.

Demographic information is crucial to our investigation of labeler bias. Hence, we collected ethnicity and sex information from the participants in each phase of this investigation. Since ethnicity categories are not objectively defined [50], we align with prior related work [15, 16]. As such, we used the following seven groups, which will be referred to as ethnicities in this work: (1) Black, (2) East Asian, (3) Indian, (4) Latino Hispanic, (5) Middle Eastern, (6) South East Asian, and (7) White. We also categorized

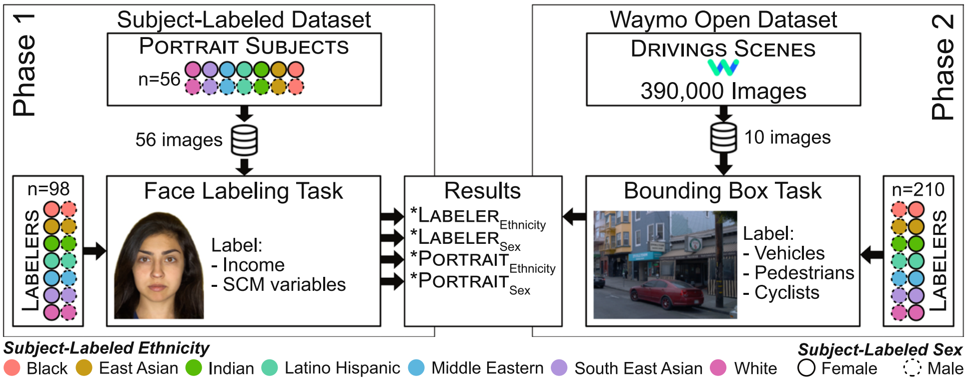
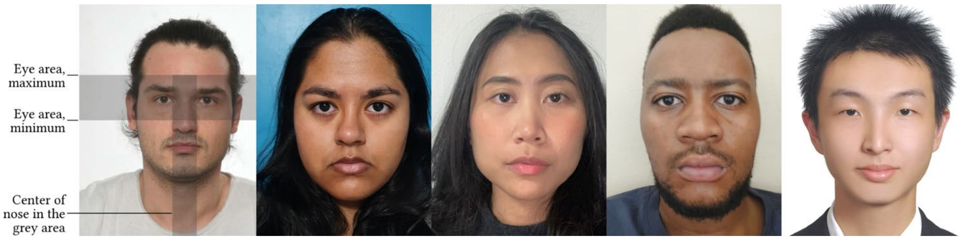


Fig. 1 Phase 1: We generated a face dataset with equal, self-labeled representation from 7 ethnicities and 2 sexes. We then recruited participants with the same demographics to label secondary characteristics. Phase 2: We recruited participants with the same demographics to complete a bounding box task. LABELER$_{Ethnicity}$ was significant in both tasks. LABELER$_{Sex}$, PORTRAIT$_{Ethnicity}$, and PORTRAIT$_{Sex}$ were significant for face labeling



(a) Example provided to the participants (b) Indian female (c) South East Asian female (d) Black male (e) East Asian male

Fig. 2 Sample images from the dataset with subject-labeled ethnicity and sex information. All subjects consented to the publication of their images

the participants according to sex assigned at birth (male or female), in line with prior work [15, 16].

4 Phase 1: face labeling task

In Phase 1, we intend to reproduce and verify the results by Haliburton et al. [15] while addressing their major limitation of using portraits without ground-truth labels. In general, past work commonly uses publicly available datasets for ML applications [15, 51, 52], but the representation in these datasets is typically skewed toward White people [16]. The FairFace dataset was developed with a balanced representation across seven ethnicities [16] in response to this trend. However, the ethnicity, sex, and age labels in the FairFace dataset were annotated by MTurk workers rather than being subject-labeled and, as such, do not represent the ground truth. Hence, we elected to generate a new dataset for this study.

4.1 Face dataset generation

We generated a dataset for this study and future research with subject-labeled ethnicity, sex, age, income, employment status, and education level.

4.1.1 Data collection

We recruited participants using Prolific[1] as it is possible to obtain demographic information about participants directly from the platform. Additionally, we asked participants for their demographic information in the survey and only accepted them if their responses matched the information provided by Prolific. We asked participants to carefully read and sign a consent form granting permission for their images to be used in future research applications. Furthermore, we asked participants to provide a passport-like frontal face photo with Fig. 2a as an example, with the following requirements: (1) Minimum 600 × 600 pixels in dimension, (2) in JPEG (.jpg) or PNG (.png) file format, (3) in color, (4) clear and in focus, (5) face must take up to 70–80% of the photo, (6) face must be centered in the photo, (7) facial expression must be neutral, (8) hats or head coverings are not allowed, (9) unaltered by computer software, and (10) file size: at least 50 KB and no more than 10 MB.

4.1.2 Resulting dataset

We collected 56 face images with subject-labeled demographic information, four from each ethnicity and sex category, c.f., [15]. Three authors reviewed the photos and unanimously agreed that they satisfied the requirements, see Fig. 2 for example photos. We use the images from this dataset in Sect. 4, and they are available upon request for future research purposes.

4.2 Evaluation method

4.2.1 Participants

We recruited $N = 98$ participants (49 female and 49 male) aged 19–49 years ($M = 27.7$, $SD = 6.51$) using Prolific, making sure to equally balance participants across the seven ethnicity categories defined in Sect. 3. We compensated participants at a rate of 10€/h for a total of 1.67€ The study was approved by the ethics committee within the University Faculty.[2]

4.2.2 Procedure

After a brief introduction to the task, the participants provided demographic information. We only accepted participants if their responses matched the information provided by Prolific. Next, we presented participants with one randomly chosen image from the dataset from each ethnicity and sex group at a time for a total of 14 images per participant. Participants completed SCM questions [40] for each image and estimated the income. The participants responded to a visual analog scale to estimate the income of each portrait. The

[1] Prolific: https://www.prolific.co.

[2] Details removed for anonymization purposes.

scale ranged from “low” to “high” to prevent bias or confusion due to currencies or country of origin.

4.2.3 Measures and analysis

We combined the SCM questions according to the original documentation [40], resulting in numerical values for warmth, competence, status, and competition. We then analyzed the relationship between stereotype variables and income estimates using Pearson correlations. To investigate the relationship between demographics and labels, we performed two-way ANOVA models (Type III, $\alpha = 0.05$) using Mauchly corrections on the *df* where the sphericity assumption was violated.

4.3 Results

4.3.1 The impact of stereotypes on estimations

We conducted a Pearson correlation analysis on income and all the SCM variables. The results, shown in Table 1, reveal that warmth, competence, status, and competition are all significantly positively correlated with income. Status has the strongest positive correlation with income and is shown as an example plot in Fig. 3. Plots for all dependent variables are included in the supplementary material.

Table 1 The Pearson correlations for each of the stereotype variables and the estimated income

	p	*r*
Warmth	**<0.001**	0.21
Competence	**<0.001**	0.53
Status	**<0.001**	0.78
Competition	**<0.001**	0.36

Significant *p* values are in bold font. All SCM variables have a significant positive correlation with the income estimation

4.3.2 The impact of demographics on estimations

We conducted an ANOVA on the income, warmth, competence, status, and competition estimations using the interaction effects of $\textsc{Labeler}_{\mathit{Ethnicity}}$, $\textsc{Labeler}_{\mathit{Sex}}$, $\textsc{Portrait}_{\mathit{Ethnicity}}$, and $\textsc{Portrait}_{\mathit{Sex}}$, c.f., Table 2. The results reveal a significant main effect of $\textsc{Labeler}_{\mathit{Ethnicity}}$ on income and competence with a large effect size, and status with a medium effect size. There is a significant main effect of $\textsc{Labeler}_{\mathit{Sex}}$ on competence with a small effect size. There is also a significant main effect of $\textsc{Portrait}_{\mathit{Ethnicity}}$ on all dependent variables, with a small effect size for warmth, medium for competence, and a large effect size for all others. There is a significant main effect of $\textsc{Portrait}_{\mathit{Sex}}$ with a medium effect size on all dependent variables as well as an interaction effect between $\textsc{Portrait}_{\mathit{Ethnicity}}$ and $\textsc{Portrait}_{\mathit{Sex}}$ on all variables with medium effect sizes, except warmth. Finally, there is an interaction effect on competition for both $\textsc{Labeler}_{\mathit{Ethnicity}}$ and $\textsc{Portrait}_{\mathit{Ethnicity}}$ as well as $\textsc{Labeler}_{\mathit{Sex}}$ and $\textsc{Portrait}_{\mathit{Sex}}$, both with medium effect sizes. All other effects are not significant. Figure 4 shows income as a function of $\textsc{Labeler}_{\mathit{Ethnicity}}$ and $\textsc{Portrait}_{\mathit{Ethnicity}}$ as an example. Plots for all dependent estimate variables are included in the supplementary material.

4.4 Discussion

Our results show that all SCM variables are correlated with income estimations to varying degrees, with Status and Competition showing the strongest correlations (see Table 1). We can thus assert that Haliburton et al. [15] results are robust, although they did not find a correlation for warmth. All of the correlations are positive (e.g., high status correlates with high income), which matches expectations from prior work [40]. The results also demonstrate that variation in labeler and portrait demographics both impact estimations for income and SCM variables. In line with previous work [15], we found $\textsc{Portrait}_{\mathit{Ethnicity}}$ to be a highly influential factor, as it significantly impacts all variables in our results.

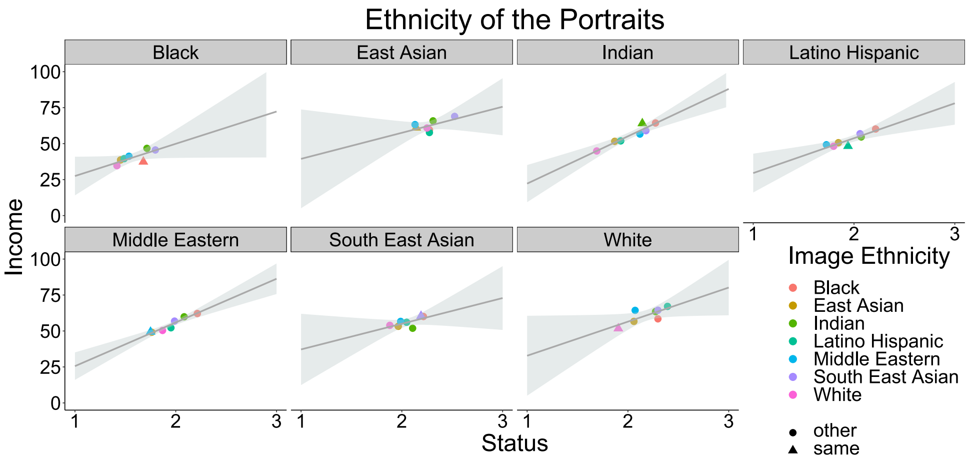


Fig. 3 Correlation between mean status and income. Each subplot represents a portrait ethnicity, and the points in each plot show how labelers of each ethnicity rated the portraits. The shaded gray area in each plot represents the 95% confidence interval for the correlation

Table 2 The four-way ANOVA results for the income estimates and the stereotype variables for Labeler$_{Ethnicity}$, Labeler$_{Sex}$, Portrait$_{Ethnicity}$, and Portrait$_{Sex}$

	Income					Warmth					Competence				
	df	*df*	*F*	*p*	η_p^2	*df*	*df*	*F*	*p*	η_p^2	*df*	*df*	*F*	*p*	η_p^2
Labeler$_{Eth}$	6	84	2.31	**0.04**	0.14	6	84	1.00	0.430	0.07	6	84	4.39	**<0.001**	0.24
Labeler$_{Sex}$	1	84	1.89	0.173	0.02	1	84	3.12	0.081	0.04	1	84	3.95	**0.04**	0.04
L$_{Eth}$ × L$_{Sex}$	6	84	0.46	0.835	0.03	6	84	1.17	0.330	0.08	6	84	1.50	0.187	0.07
Portrait$_{Eth}$	6	504	30.5	**<0.001**	0.27	5.21	437	2.64	**0.021**	0.03	6	504	6.78	**<0.001**	0.07
Portrait$_{Sex}$	1	84	17.4	**<0.001**	0.17	1	84	15.4	**<0.001**	0.15	1	84	8.17	**0.005**	0.09
P$_{Eth}$ × P$_{Sex}$	5.17	434	8.54	**<0.001**	0.09	5.13	431	1.88	0.94	0.02	6	504	7.99	**<0.001**	0.09
L$_{Eth}$ × P$_{Eth}$	36	504	1.20	0.206	0.08	31.2	437	1.12	0.301	0.07	36	504	1.02	0.447	0.07
L$_{Sex}$ × P$_{Sex}$	1	84	0.03	0.868	0.01	1	84	3.06	0.084	0.04	1	84	0.04	0.841	0.01
	Status					Competition									
	df	*df*	*F*	*p*	η_p^2	*df*	*df*	*F*	*p*	η_p^2					
Labeler$_{Eth}$	6	84	2.99	**0.011**	0.18	6	84	1.57	0.167	0.10					
Labeler$_{Sex}$	1	84	0.83	0.366	0.01	1	84	1.79	0.185	0.02					
L$_{Eth}$ × L$_{Sex}$	6	84	0.45	0.840	0.03	6	84	1.02	0.415	0.07					
Portrait$_{Eth}$	6	504	30.4	**<0.001**	0.27	4.99	419	15.8	**<0.001**	0.16					
Portrait$_{Sex}$	1	84	11.2	**0.001**	0.12	1	84	12.7	**<0.001**	0.13					
P$_{Eth}$ × P$_{Sex}$	5.22	438	7.82	**<0.001**	0.09	6	504	2.55	**0.019**	0.03					
L$_{Eth}$ × P$_{Eth}$	36	504	0.76	0.843	0.05	29.9	419	1.51	**0.043**	0.10					
L$_{Sex}$ × P$_{Sex}$	1	84	0.02	0.875	0.01	1	84	8.81	**0.004**	0.09					

Significant *p* values are in bold font. Note that *df* is degrees of freedom, *F* is the F-statistic, and η_p^2 denotes effect size

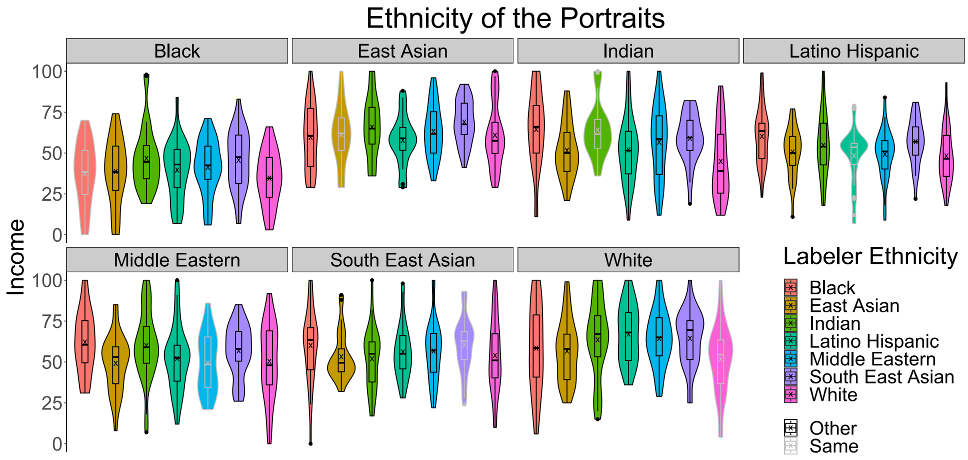


Fig. 4 Estimated income as a function of Labeler$_{Ethnicity}$ and Portrait$_{Ethnicity}$. Grey borders indicate the cases where Labeler$_{Ethnicity}$ and Portrait$_{Ethnicity}$ match

However, we also found some impacts of Labeler$_{Sex}$ and found that Portrait$_{Sex}$ impacted all dependent variables, which differs from the results in prior work.

Our findings support those of [15], suggesting that demographic-based labeler bias exists in face-labeling tasks and that stereotype content helps explain this bias [53]. Although there may have been biases propagated through the FairFace dataset [16] due to the labeling process, the results are mostly consistent with our subject-labeled images. These results are also in line with recent work in other domains, such as findings from Goyal et al. [8] suggesting that toxicity labels depend on labeler identity.

4.4.1 Stereotype bias is present independent of labeler demographics

While our results indicate that labeler demographics impact estimations, we have also found that the demographics of a person displayed in an image have an even more prominent impact. As seen in Table 2, Portrait$_{Ethnicity}$, and Portrait$_{Sex}$ have significant main effects on every dependent variable. These findings indicate that results in a face-labeling task will contain bias regardless of who is doing the labeling. As a consequence, datasets for ML should be more diverse. As Karkkainen and Joo [16] point out, most datasets containing images of people have a skewed representation of demographics. Unequal representation in datasets has been the subject of strong criticism in recent work [54], and while current sentiment is trending towards rectifying this issue (e.g., the FairFace dataset [16]), it is not yet solved. Future work in data generation should make a concerted and deliberate effort to include diverse subjects as an initial step to mitigate this issue. This recommendation echoes calls in the literature for increased demographic documentation [3].

4.4.2 Primary and secondary characteristics

The task in this phase asked participants to label secondary information (i.e., information that does not directly appear in the image). Regardless of skill, the participants had no method to determine the correct level of income objectively, and all labels were an inference. Past work has questioned whether secondary characteristics are appropriate tasks for facial recognition [17] since the true answer does not appear in the image. In response to this, as well as the clear influence of stereotype content in this study, we conducted a bounding box labeling task where participants were asked to label primary characteristics. Section 5 details this task.

5 Phase 2: bounding box task

In this section, we investigate labeler bias in a highly relevant ML task, namely object recognition for autonomous driving. This section aims to understand whether the labeler bias identified in the face-labeling task extends to other tasks that do not involve human faces.

5.1 Dataset curation

We selected images from the Waymo Open Dataset [18, 55] for the bounding box task. This dataset contains 390,000 images in a variety of cities, lighting conditions, and environments and has previously been used for training autonomous driving systems [18, 56]. The dataset features labeled vehicles, pedestrians, and cyclists. For consistency, we selected a subset of images from the dataset, each containing 5 objects to be labeled. Three authors reviewed and unanimously selected 10 images to be used in the study where the 5 labeled objects are distinct and recognizable. The selected images are included in the supplementary material.

5.2 Evaluation method

5.2.1 Participants

We recruited $N = 210$ participants (105 female and 105 male) from MTurk. Participants were between 19 and 66 years old ($M = 35.6$, $SD = 10.5$). We ensured that participants were equally distributed across the seven ethnicity categories defined in Sect. 3, resulting in 15 participants per ethnicity-sex intersection. We only included participants who had a minimum of 100 approved HITs with at least a 90% approval rate. We compensated participants at a rate of 10€ per hour for a total of 1.5€. The study was approved by the ethics committee within the University Faculty.[3]

5.2.2 Procedure

Participants first completed a demographics screening questionnaire where we collected ethnicity and sex information. Upon completion, participants were provided with a link to proceed to the bounding box survey as long as we had not yet reached capacity for that participant's ethnicity and sex intersection.[4] This procedure enabled us to collect equal numbers of participants from each category.

To annotate the images, we recruited crowd workers through MTurk. We selected MTurk over Prolific for this experiment because of the built-in bounding box interface, which we used for this task. Crowd workers from MTurk are commonly used for annotation in ML applications [57] and are required to pass a qualification task before they can complete bounding box tasks.

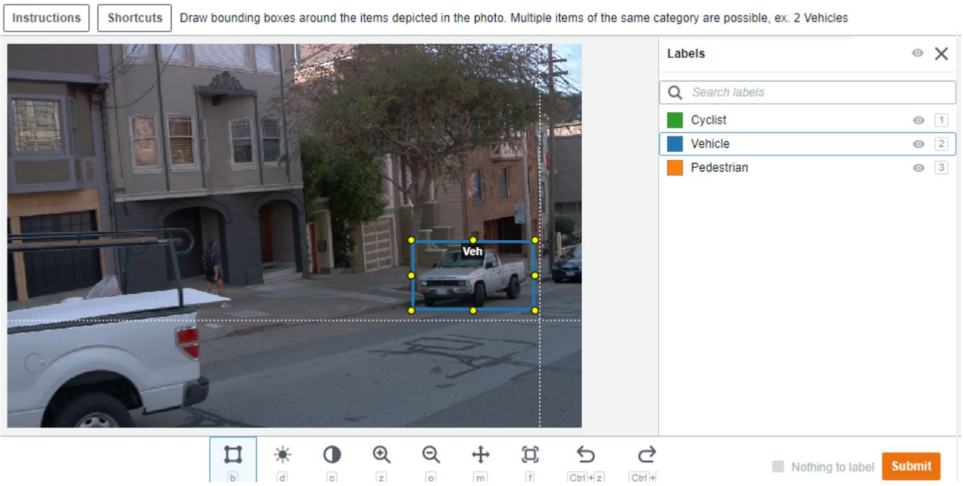
In the bounding box survey, participants were required to label all cyclists, pedestrians, and vehicles in a series of images. Each participant labeled all 10 images, each containing 5 objects to be labeled. Figure 5 shows the task interface with an example image. The participants were given "Cyclist," "Pedestrian," and "Vehicle" as annotation options.

5.2.3 Measures and analysis

We collected the bounding box coordinates for all participants across all ten images. We then calculated accuracy via Intersection over Union (IoU) [58], which has been used in prior work on object detection [59] and tracking [60]. IoU ranges from 0.0 to 1.0 (where 0.0 indicates no overlap and 1.0 is an exact match) and is calculated by dividing the sum of the overlapping areas of two boxes by their total combined

[3] Details removed for anonymization purposes.

[4] The demographics of the participants who responded to the pre-screening survey are as follows: Male 47.8%, Female 52.1%, White 63.8%, Middle Eastern 11.2%, Indian 8.22%, Latino Hispanic 5.91%, Black 4.87%, East Asian 3.58%, South East Asian 2.40%.



Draw bounding boxes around the items depicted in the photo. Multiple items of the same category are possible, ex. 2 Vehicles

Fig. 5 Example task from the survey for annotating images with bounding boxes. Image from the Waymo Open Dataset [18]

area. We calculated a mean IoU score for each participant-image pair, resulting in 2100 measurements.

We also calculated the Mean Average Precision (mAP) for each participant-image pair by calculating the Precision and Recall at 11 IoU threshold values (0 to 1 with steps of 0.1) and obtaining the area under the resulting curve. Precision is calculated as the number of True Positives (IoU above the threshold and correct label) divided by the number of True Positives and False Positives (IoU above the threshold and incorrect label), while Recall is the number of True Positives divided by the number of True Positives and False Negatives (failure to label an object).

We used ANOVA procedures to analyze the accuracy metrics. We used two-way ANOVA tests when the data are normally distributed according to Shapiro-Wilk testing [61] and used ART-ANOVA procedures otherwise.

5.3 Results

Figure 6 shows an example annotated figure from one participant with labeled ground truth boxes and IoU values. Figure 7 shows an example of the variation in bounding box coordinates for a single object.

We analyzed the bounding box accuracy using both IoU and mAP. The results for both IoU and mAP were not normally distributed according to Shapiro–Wilk testing (IoU: $W = 0.948, p < 0.001$; mAP; $W = 0.906, p < 0.001$). As such, we performed ART-ANOVA procedures to evaluate both dependent variables. The results, shown in Table 3, reveal a significant main effect of $\textsc{Labeler}_{Ethnicity}$ on both IoU ($p = .01$) and mAP ($p < 0.001$) with medium effect sizes. Figure 8 displays the IoU and mAP results as a function of $\textsc{Labeler}_{Ethnicity}$. As the mAP is calculated as the area under the Precision-Recall curve, we show the Precision-Recall curves by $\textsc{Labeler}_{Ethnicity}$ in Fig. 9.

5.4 Discussion

Our results indicate that $\textsc{Labeler}_{Ethnicity}$ has a significant main effect on bounding box accuracy, as measured by IoU and mAP metrics. We found no significant effect for $\textsc{Labeler}_{Sex}$, nor for the interaction between $\textsc{Labeler}_{Ethnicity}$ and $\textsc{Labeler}_{Sex}$. These results are in line with past work suggesting that identity impacts labeling [8] and that ethnicity impacts estimations while sex does not [15].

5.4.1 Labeler bias is present even in non-stereotype tasks

Our results imply that labelers produce biased labels even when the labeling task is not subject to stereotype effects. Current labeling pipelines often use MTurk for data labeling without collecting demographic information, and MTurk has a non-representative demographic distribution [28–30]. Our results, therefore, indicate that there is a systemic issue with the status quo in data labeling. To remedy this, it may be

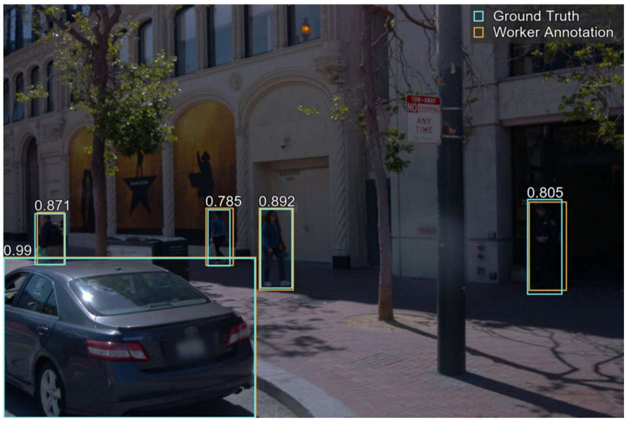


Fig. 6 Example annotated image with IoU scores containing four pedestrians and one vehicle, image from the Waymo Open Dataset [18]

Table 3 The ART-ANOVA results for the bounding box accuracy (*IoU* and *mAP*) by Labeler$_{Ethnicity}$ and Labeler$_{Sex}$

	Labeler$_{Ethnicity}$				Labeler$_{Sex}$				L$_{Ethnicity}$ × L$_{Sex}$			
	dF	*F*	*p*	η_p^2	*dF*	*F*	*p*	η_p^2	*dF*	*F*	*p*	η_p^2
IoU	(6,196)	2.82	**0.01**	0.08	(1,196)	0.856	0.356	0.04	(6,196)	1.94	0.0757	0.06
mAP	(6,196)	4.14	**<0.001**	0.11	(1,196)	1.08	0.300	0.01	(6,196)	1.19	0.312	0.04

Significant *p* values are in bold font

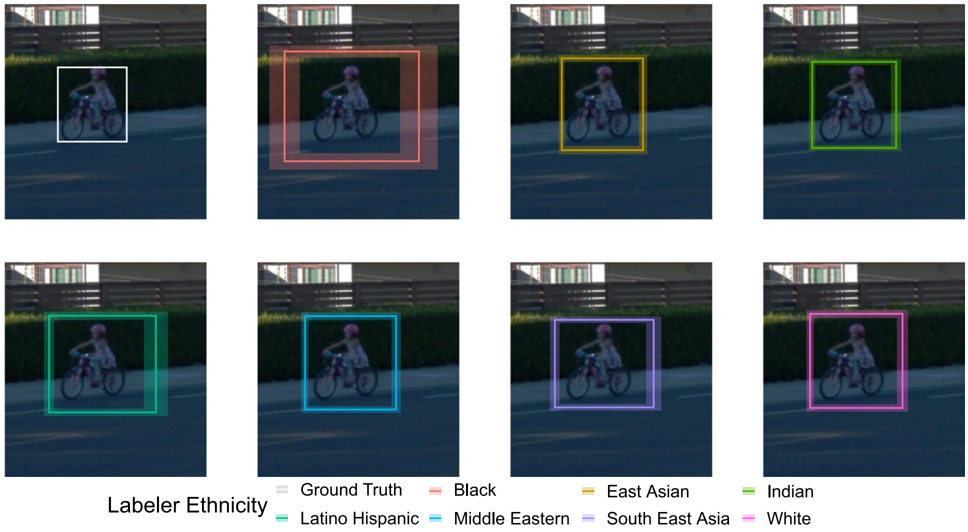


Fig. 7 One example object showing the ground truth bounding box and representative boxes with variation for each ethnicity. The solid line represents the mean box coordinates and the translucent box represents the standard error variation. Image from the Waymo Open Dataset [18]

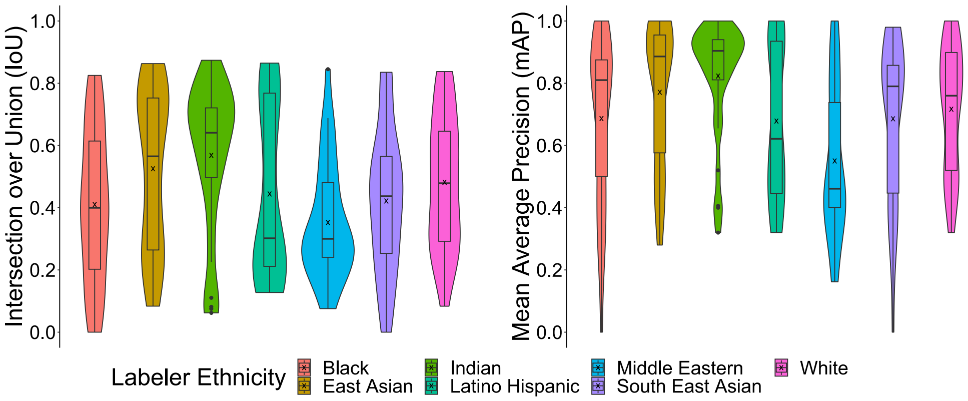


Fig. 8 Bounding box accuracy metrics reveal a significant main effect of Labeler$_{Ethnicity}$ for both *IoU* (left) and *mAP* (right)

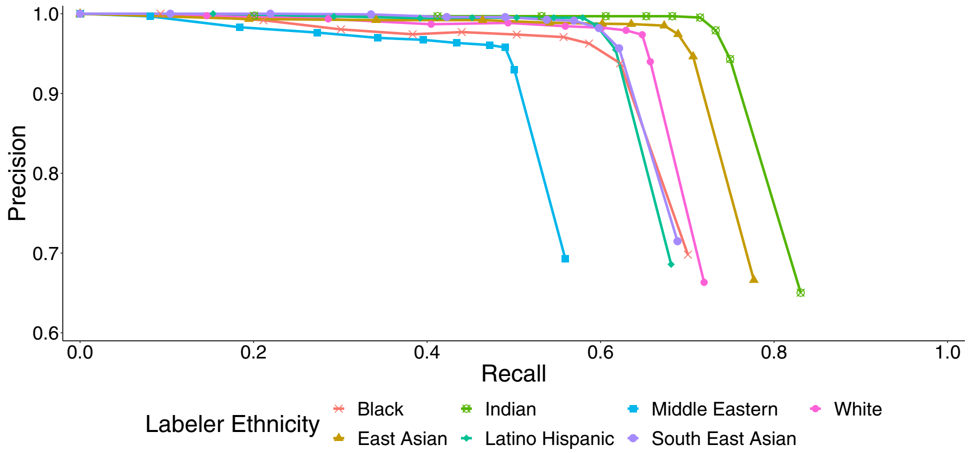


Fig. 9 Precision-recall curves for each Labeler$_{Ethnicity}$

useful to collect demographic information from labelers and subsequently use a balanced and diverse sample of labelers to counteract some of the bias effects. This recommendation is a reiteration from Sect. 4 as well as recent work in related literature [3].

5.4.2 Developers should balance ethnicity across train-validation-test splits

One implication of our results is that ground truth labels generated by labelers of different ethnicities are likely to be different. This implies that if a developer recruits participants of a single ethnicity to label a training set and recruits participants of a different ethnicity (wittingly or no) to label the test set, the model will appear to have a lower accuracy due to the ethnicity differences. This suggests that it is crucial to balance ethnicities across the train-validation-test split when training a model, as a different distribution of ethnicities in the two groups may artificially appear as reduced performance. Balancing the training and testing split in this manner would require developers to collect demographic data from labelers during the annotation process, which we suggest above, and past work has also recommended [3].

6 General discussion

We investigated labeler bias in two scenarios. First, we created a dataset of face images with subject-labeled demographic information and recruited participants to estimate the income and stereotype content of the images. We found that Labeler$_{Ethnicity}$, Labeler$_{Sex}$, Portrait$_{Ethnicity}$, and Portrait$_{Sex}$ all impact the estimation results to some degree. We also found that the income estimation significantly correlates with all SCM variables. We then recruited participants to conduct a bounding box task and found that Labeler$_{Ethnicity}$ had a significant effect on both *IoU* and *mAP* accuracy metrics, while Labeler$_{Sex}$ had no significant effect.

6.1 Where does labeler bias come from?

We found that labels significantly vary with labeler demographics in the face-labeling task (Sect. 4). However, we also found that the estimations depended on the demographic characteristics of the labeled images. Since the income estimations significantly covary with the SCM variables, we selected a bounding box task to investigate whether labeler bias persists in a task where stereotype content is not applicable. In Sect. 5, we found that labeler demographics still have a significant effect on annotations even when there is no stereotype content.

Based on these results, we hypothesize that two forms of bias may be at play here. One form is what is traditionally thought of as labeler bias, where the biases of labelers are propagating through the labeling process independent of the subject to be labeled. This bias accounts for variation in the bounding box task and explains some of the variation in the face-labeling task. The exact source of this inherent bias is unclear, as past work suggests that bias can stem from a variety of factors, such as socio-economic conditions [7, 37]. The other bias is a bias due to stereotypes, independent of the labelers. This stereotype bias is evident in the fact that Portrait$_{Ethnicity}$ and Portrait$_{Sex}$ have a significant effect on nearly every variable in the face-labeling task. These results suggest that labeling tasks that feature faces will have biased results even when developers recruit a diverse set of labelers.

6.2 Practical recommendations for dataset labeling

In the face-labeling task, we found that the demographics of the labelers had a significant impact on their estimations. However, since the task involved labeling secondary characteristics (i.e., features not physically present in the image), this implies that the estimations vary relative to one another, rather than varying from a ground truth value. The results from the bounding box task, however, are different. There are ground truth values in the bounding box task [18], so any variation in the results implies that *some groups are more accurate in this labeling task*.

We urge the reader not to over-generalize this finding. We do not suggest that labelers who belong to an ethnic group that performed well in our experiment will categorically perform better than other groups in all labeling tasks. However, it is interesting to note that in this particular scenario, recruiting a diverse set of labelers would *reduce* the accuracy when training ML models compared to choosing labelers from the highest-performing group. In this work, we have not investigated whether different ethnic groups perform better at different labeling tasks. Given the number and variety of possible labeling tasks, characterizing labeling performance in this way would likely be futile. Rather, we suggest recruiting labelers from diverse ethnicities to ensure that the highest-performing group is always included.

If technology designers aim to recruit a diverse set of labelers, they should collect demographic information when they recruit labelers. Although this is not currently common practice in the annotation process, this recommendation is supported by our results and prior work [3]. We also recommend increasing transparency in documenting the annotation process. Prior work has shown that communicating the credibility of training data increases trust in AI systems as long as that system does not show bias [31]. We additionally support incorporating holistic suggestions from prior work, including addressing socio-economic conditions and fairly compensating labelers [37].

Prior work has identified that the demographics of MTurk workers are categorically non-diverse. Two-thirds of MTurk workers identify as white [5], and most reside in the United States and India [28–30]. Our experience collecting data for the bounding box task aligns with these findings. This stark lack of diversity implies that randomly recruiting labelers from MTurk has a vanishingly small likelihood of being balanced across ethnicities.

In sum, although future work is needed to develop further solutions to this issue, the following steps can be taken immediately to begin moving towards a more fair and transparent data labeling process:

1. Record demographic information from labelers
2. Recruit a diverse sample of labelers
3. Report labeler demographics in the dataset documentation

6.3 Limitations and future work

Although we have conducted this investigation with the utmost care, limitations remain. First, while we investigated two disparate annotation tasks and uncovered evidence for labeler bias in each, it remains to be seen whether this bias generalizes to all annotation tasks. Continued research is required to characterize the nature of labeler bias in other tasks, which has been the subject of other investigations into NLP [13] and toxicity labeling [8], among others. Our work motivates the need for similar investigations to be carried out for any ML task that uses labeled data. For example, medical imaging [62] or spam detection [63]. For further examples please see reviews of supervised ML applications by Shetty et al. [64] and Sarker [65].

Additionally, although we make an effort to characterize the nature of labeler bias in face labeling and bounding box tasks, we have not explored whether this information is sufficient to account for the resulting bias in an ML system trained on this data. Future research should take our findings and apply corrections based on the measured biases. Furthermore, our results suggest that there are at least two sources of labeler bias, both dependent on, and independent of stereotypes. This strongly motivates the need for interdisciplinary research to investigate and characterize the nature of the various forms of bias that contribute to biased labels in ML datasets.

Finally, while the bounding boxes in the Waymo Open Dataset [18] are reported as "ground truth" in the documentation, they were labeled by humans. As such, there is a possibility that there is bias embedded in these ground truth values. Past work has shown that even experienced labelers can embed bias in subjective tasks [38], and our current work indicates that bias can be embedded in primary tasks as well. This suggests that there is a possibility that experienced labelers could embed bias in primary tasks. While this question is interesting, we argue that our results remain valid. The positions of the ground truth bounding boxes act as intercepts against which we compare the annotations from our participants. The fact that we identified significant variation in the bounding box labels between groups of participants is the important aspect, and these between-group differences should be independent of the intercept. To confirm, we re-ran our analysis and calculated the *IoU* and *mAP* seven times with the responses from a randomly selected participant from each ethnicity as the ground truth. There were no changes in any conclusions (e.g. all significant results remained significant, and vice versa). The resulting table can be found in the supplemental material.

7 Conclusion

In this paper, we investigated labeler bias in two ML scenarios. We first created a dataset of face images with subject-labeled demographic information. Using this dataset, we then recruited a diverse sample of 98 participants to conduct a face-labeling task where they estimated the income and stereotype characteristics. Finally, we recruited 210 additional participants to perform a bounding box task. The results of the face-labeling task indicate that $\textsc{Labeler}_{Ethnicity}$, $\textsc{Portrait}_{Ethnicity}$, and $\textsc{Portrait}_{Sex}$ all have a significant effect on income estimations. We also found main and interaction effects for $\textsc{Labeler}_{Ethnicity}$, $\textsc{Labeler}_{Sex}$, $\textsc{Portrait}_{Ethnicity}$, and $\textsc{Portrait}_{Sex}$ on stereotype perceptions. The results of the bounding box task reveal a significant main effect of $\textsc{Labeler}_{Ethnicity}$ on accuracy. Together, our results indicate that it is important for developers to know *who* is doing the labeling. On the one hand, recruiting a non-diverse sample of labelers will likely lead to biased results. On the other hand, especially for primary tasks, our findings suggest that ethnicity should be balanced across the train-validation-test split when training ML models. Otherwise, the model may not generalize well and may artificially report reduced accuracy. We recommend that collecting demographic information to ensure a diverse sample should become standard procedure in data labeling pipelines as a small step towards fairer intelligent systems in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-024-00572-w.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by LMU Munich and the Munich Center for Machine Learning (MCML).

Data and materials availability Researchers who wish to access the dataset should contact the corresponding author and briefly describe their use case. The dataset is available for research purposes only.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

Ethics approval and consent to participate The study was approved by the ethics committee within the LMU Munich Faculty for Mathematics, Informatics, and Statistics

Consent to publish The authors affirm that human research participants provided informed consent for the publication of their images.

Materials availability Researchers who wish to access the dataset should contact the corresponding author and briefly describe their use case. The dataset is available for research purposes only.

References

1. Eickhoff, C.: Cognitive biases in crowdsourcing. In: Proceedings of the Eleventh ACM International Conference on Web Search And Data Mining. WSDM '18, pp. 162–170. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3159652.3159654
2. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. ACM, New York (2021). https://doi.org/10.1145/3442188.3445922
3. Miceli, M., Posada, J., Yang, T.: Studying up machine learning data: why talk about bias when we mean power? Proc. ACM Hum. Comput. Interact. **6**(GROUP), 34–13414 (2022). https://doi.org/10.1145/3492853
4. Moss, A.J., Rosenzweig, C., Robinson, J., Jaffe, S.N., LItman, L.: Is it ethical to use mechanical turk for behavioral research? Relevant Data from a Representative Survey of MTurk Participants and Wages. PsyArXiv (2020). https://doi.org/10.31234/osf.io/jbc9d
5. Levay, K.E., Freese, J., Druckman, J.N.: The demographic and political composition of mechanical turk samples. SAGE Open (2016). https://doi.org/10.1177/2158244016636433
6. Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES '21, pp. 298–306. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3461702.3462624
7. Miceli, M., Schuessler, M., Yang, T.: Between subjectivity and imposition: power dynamics in data annotation for computer vision. Proc. ACM Hum. Comput. Interact. **4**(CSCW2), 115–111525 (2020). https://doi.org/10.1145/3415186
8. Goyal, N., Kivlichan, I.D., Rosen, R., Vasserman, L.: Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. Proc. ACM Hum. Comput. Interact. (2022). https://doi.org/10.1145/3555088
9. Kaplan, S., Handelman, D., Handelman, A.: Sensitivity of neural networks to corruption of image classification. AI Ethics **1**(4), 425–434 (2021). https://doi.org/10.1007/s43681-021-00049-0
10. Jiang, H., Nachum, O.: Identifying and correcting label bias in machine learning. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 108, pp. 702–712. PMLR, Virtual (2020). https://proceedings.mlr.press/v108/jiang20a.html
11. Wauthier, F.L., Jordan, M.: Bayesian bias mitigation for crowdsourcing. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 24. Curran Associates, Inc., Granada (2011). https://proceedings.neurips.cc/paper_files/paper/2011/file/0768281a05da9f27df178b5c39a51263-Paper.pdf
12. Cohn, T., Specia, L.: Modelling annotator bias with multi-task Gaussian processes: an application to machine translation quality estimation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 32–42. Association for Computational Linguistics, Sofia (2013). https://aclanthology.org/P13-1004
13. Geva, M., Goldberg, Y., Berant, J.: Are we modeling the task or the annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets (2019). arXiv:1908.07898
14. Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., Smith, N.A.: Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection (2022). arXiv:2111.07997
15. Haliburton, L., Ghebremedhin, S., Welsch, R., Schmidt, A., Mayer, S.: Investigating labeler bias in face annotation for machine learning. In: HHAI 2024: Hybrid Human AI Systems for the Social Good. Frontiers in Artificial Intelligence and Applications, vol. 386, pp. 145–161. IOS Press, Amsterdam (2024). https://doi.org/10.3233/FAIA240191
16. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Winter Conference on Applications of Computer Vision (WACV). IEEE, Waikoloa (2021). https://doi.org/10.1109/WACV48630.2021.00159
17. Engelmann, S., Ullstein, C., Papakyriakopoulos, O., Grossklags, J.: What People think AI should infer from faces. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22, pp. 128–141. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3531146.3533080
18. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset, pp. 2446–2454 (2020). https://openaccess.thecvf.com/content_CVPR_2020/html/Sun_Scalability_in_Perception_for_Autonomous_Driving_Waymo_Open_Dataset_CVPR_2020_paper.html
19. Howe, J.: Why the Power of the Crowd is Driving the Future Of Business. Random House, New York (2008)
20. Eickhoff, C., Harris, C.G., Vries, A.P., Srinivasan, P.: Quality through flow and immersion: gamifying crowdsourced relevance assessments. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12, pp. 871–880. Association for Computing Machinery, New York (2012). https://doi.org/10.1145/2348283.2348400

21. Alonso, O., Rose, D.E., Stewart, B.: Crowdsourcing for relevance evaluation. ACM SIGIR Forum **42**(2), 9–15 (2008). https://doi.org/10.1145/1480506.1480508
22. Grady, C., Lease, M.: Crowdsourcing document relevance assessment with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 172–179. Association for Computational Linguistics, Los Angeles (2010). https://aclanthology.org/W10-0727
23. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08, pp. 453–456. Association for Computing Machinery, New York (2008). https://doi.org/10.1145/1357054.1357127
24. Sheng, V.S., Zhang, J.: Machine learning with crowdsourcing: a brief summary of the past research and future directions. Proc. AAAI Conf. Artif. Intell. **33**(01), 9837–9843 (2019). https://doi.org/10.1609/aaai.v33i01.33019837
25. Eickhoff, C., Vries, A.P.: Increasing cheat robustness of crowdsourcing tasks. Inf. Retr. **16**(2), 121–137 (2013). https://doi.org/10.1007/s10791-011-9181-9
26. Difallah, D.E., Demartini, G., Cudré-Mauroux, P.: Pick-a-crowd: tell me what you like, and i'll tell you what to do. In: Proceedings of the 22nd International Conference on World Wide Web. WWW '13, pp. 367–374. Association for Computing Machinery, New York (2013). https://doi.org/10.1145/2488388.2488421
27. Fleischmann, M., Amirpur, M., Benlian, A., Hess, T.: Cognitive biases in information systems research: a scientometric analysis. In: Proceedings of the European Conference on Information Systems (ECIS) Tel Aviv, Israel (2014)
28. Difallah, D., Filatova, E., Ipeirotis, P.: Demographics and dynamics of mechanical turk workers. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18. ACM, New York (2018). https://doi.org/10.1145/3159652.3159661
29. Ipeirotis, P.G.: Demographics of mechanical turk. Technical Report 1585030, Rochester, New York (2010). https://papers.ssrn.com/abstract=1585030
30. Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers? shifting demographics in mechanical turk. In: CHI '10 Extended Abstracts on Human Factors in Computing Systems. CHI EA '10. ACM, New York (2010). https://doi.org/10.1145/1753846.1753873
31. Chen, C., Sundar, S.S.: Is this AI trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, pp. 1–11. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3544548.3580805
32. Fruchard, B., Malacria, S., Casiez, G., Huot, S.: User Preference and Performance using Tagging and Browsing for Image Labeling. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, pp. 1–13. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3544548.3580926
33. Stureborg, R., Dhingra, B., Yang, J.: Interface design for crowdsourcing hierarchical multi-label text annotations. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, pp. 1–17. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3544548.3581431 . https://dl.acm.org/doi/10.1145/3544548.3581431
34. Oyshi, M.T., Vogt, S., Gumhold, S.: TmoTA: simple, highly responsive tool for multiple object tracking annotation. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, pp. 1–11. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3544548.3581185
35. Hirth, M., Borchert, K., De Moor, K., Borst, V., Hoßßfeld, T.: Personal task design preferences of crowdworkers. In: 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2020). https://doi.org/10.1109/QoMEX48832.2020.9123094
36. Parmar, M., Mishra, S., Geva, M., Baral, C.: Don't blame the annotator: bias already starts in the annotation instructions (2022). arXiv:2205.00415
37. Barbosa, N.M., Chen, M.: Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, pp. 1–12. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3290605.3300773
38. Hube, C., Fetahu, B., Gadiraju, U.: Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, pp. 1–12. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3290605.3300637
39. Miceli, M., Posada, J.: The Data-Production Dispositif (2022). arXiv:2205.11963
40. Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J.: A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. J. Personal. Soc. Psychol. (2002). https://doi.org/10.1037/0022-3514.82.6.878
41. Fiske, S.T., Cuddy, A.J.C., Glick, P.: Universal dimensions of social cognition: warmth and competence. Trends Cogn. Sci. (2007). https://doi.org/10.1016/j.tics.2006.11.005
42. Cuddy, A.J.C., Fiske, S.T., Glick, P.: Warmth and competence as universal dimensions of social perception: the stereotype content model and the bias map. Adv. Exp. Soc. Psychol. **40**, 61–149 (2008). https://doi.org/10.1016/S0065-2601(07)00002-0
43. Durante, F., Tablante, C.B., Fiske, S.T.: Poor but warm, rich but cold (and competent): social classes in the stereotype content model. J. Soc. Issues (2017). https://doi.org/10.1111/josi.12208
44. Grigoryev, D., Fiske, S.T., Batkhina, A.: Mapping ethnic stereotypes and their antecedents in Russia: the stereotype content model. Front. Psychol. (2019). https://doi.org/10.3389/fpsyg.2019.01643
45. McKee, K., Bai, X., Fiske, S.: Understanding human impressions of artificial intelligence. Technical report, PsyArXiv (2021). https://doi.org/10.31234/osf.io/5ursp
46. Schwind, V., Deierlein, N., Poguntke, R., Henze, N.: Understanding the social acceptability of mobile devices using the stereotype content model. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York (2019). https://doi.org/10.1145/3290605.3300591
47. Marsden, N., Haag, M.: Stereotypes and politics: reflections on personas. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16, pp. 4017–4031. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2858036.2858151
48. Fraser, K.C., Nejadgholi, I., Kiritchenko, S.: Understanding and countering stereotypes: a computational approach to the stereotype content model (2021). arXiv:2106.02596 [cs]
49. Kroon, A.C., Trilling, D., Raats, T.: Guilty by association: using word embeddings to measure ethnic stereotypes in news coverage. J. Mass Commun. Q. (2021). https://doi.org/10.1177/1077699020932304
50. Phinney, J.S.: When we talk about American ethnic groups, what do we mean? Am. Psychol. (1996). https://doi.org/10.1037/0003-066X.51.9.918

51. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach (2018). https://openaccess.thecvf.com/content_eccv_2018_workshops/w5/html/Das_Mitigating_Bias_in_Gender_Age_and_Ethnicity_Classification_a_Multi-Task_ECCVW_2018_paper.html
52. Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N., Gummadi, K.: Who makes trends? Understanding demographic biases in crowdsourced recommendations. Proc. Int. AAAI Conf. Web Soc. Media **11**(1), 22–31 (2017). https://doi.org/10.1609/icwsm.v11i1.14894
53. Fiske, S.T., Taylor, S.E.: Social Cognition, 2nd edn. Mcgraw-Hill Book Company, New York (1991)
54. D'Ignazio, C., Klein, L.F.: Data Feminism. MIT Press, Cambridge (2020)
55. Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.-C., Kretzschmar, H.: Waymo open dataset: panoramic video panoptic segmentation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision—ECCV 2022, pp. 53–72. Springer, Cham (2022)
56. Gu, Z., Li, Z., Di, X., Shi, R.: An LSTM-based autonomous driving model using a Waymo open dataset. Appl. Sci. **10**(6), 2046 (2020). https://doi.org/10.3390/app10062046
57. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139–147. Association for Computational Linguistics, Los Angeles (2010). https://aclanthology.org/W10-0721
58. Rezatofighi, S.H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I.D., Savarese, S.: Generalized intersection over union: a metric and A loss for bounding box regression. CoRR (2019). arXiv:1902.09630
59. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR (2014). arXiv:1405.0312
60. Leal-Taixé, L., Milan, A., Reid, I.D., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. CoRR (2015). arXiv:1504.01942
61. Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J.: The aligned rank transform for nonparametric factorial analyses using only anova procedures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11, pp. 143–146. Association for Computing Machinery, New York (2011). https://doi.org/10.1145/1978942.1978963
62. Rahmani, A.M., Yousefpoor, E., Yousefpoor, M.S., Mehmood, Z., Haider, A., Hosseinzadeh, M., Ali Naqvi, R.: Machine learning (ML) in medicine: review, applications, and challenges. Mathematics **9**(22), 2970 (2021). https://doi.org/10.3390/math9222970
63. Ahsan, M.N.I., Nahian, T., Kafi, A.A., Hossain, M.I., Shah, F.M.: Review spam detection using active learning. In: 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1–7 (2016). https://doi.org/10.1109/IEMCON.2016.7746279
64. Shetty, S.H., Shetty, S., Singh, C., Rao, A.: Supervised machine learning: algorithms and applications. In: Fundamentals and Methods of Machine and Deep Learning, pp. 1–16. Wiley, New York (2022). https://doi.org/10.1002/9781119821908.ch1
65. Sarker, I.H.: Machine learning: algorithms, real-world applications and research directions. SN Comput. Sci. **2**(3), 160 (2021). https://doi.org/10.1007/s42979-021-00592-x