Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning

Natã M. Barbosa
Syracuse University, Syracuse, NY 13244
Figure Eight Inc., San Francisco, CA 94103
nmbarbos@syr.edu

Monchu Chen
Figure Eight Inc., San Francisco, CA 94103
monchu.chen@figure-eight.com

ABSTRACT

The increased use of machine learning in recent years led to large volumes of data being manually labeled via crowdsourcing microtasks completed by humans. This brought about dehumanization effects, namely, when task requesters overlook the humans behind the task, leading to issues of ethics (e.g., unfair payment) and amplification of human biases, which are transferred into training data and affect machine learning in the real world. We propose a framework that allocates microtasks considering human factors of workers such as demographics and compensation. We deployed our framework to a popular crowdsourcing platform and conducted experiments with 1,919 workers collecting 160,345 human judgments. By routing microtasks to workers based on demographics and appropriate pay, our framework mitigates biases in the contributor sample and increases the hourly pay given to contributors. We discuss potential extensions and how it can promote transparency in crowdsourcing.

CCS CONCEPTS

• **Information systems → Crowdsourcing**;

KEYWORDS

Crowdsourcing; Bias; Ethics; Machine Learning.

ACM Reference Format:
Natã M. Barbosa and Monchu Chen. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300773

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300773

1 INTRODUCTION

With the growing popularity and use of crowdsourcing platforms e.g., Amazon Mechanical Turk [22] (AMT), Figure Eight [10], Task Rabbit [37], came along what researchers refer to as dehumanization effects in crowdsourcing [11, 14, 20]. These occur when task requesters overlook the human aspects of those working on the tasks, also known as workers, due to the short time commitment between requesters and workers, and the very nature of crowdsourcing. As a result, issues such as underpayment [13], "boring" tasks [20, 29], and difficulties finding "good" work [13, 14, 40], have been observed in these platforms. These have serious consequences because many depend on doing tasks to secure their income [30], which leads to more competition for tasks, and more often than not, a small number of workers (i.e., the most active) submitting a large fraction of the work available [13].

The recent surge of machine learning applications in many domains resulted in increasing demand for manually labeled data used to train algorithms for a variety of purposes, from recognizing speech, to moderating Internet content, to developing self-driving cars. However, as machine learning models become ubiquitous, so do their impacts on people's lives. For example, a biased machine learning model can make unfair decisions about a person, such as preventing them from being contacted for a job interview, classifying their gender incorrectly, or inhibiting their own personal voice assistant from recognizing their speech due to their accent, age, or gender. These issues raise the question of bias and ethics and how they can be addressed in the many stages of developing and using machine learning in the wild.

Rightfully so, the issue of algorithmic bias has received much attention lately from the perspective of when a model does not learn or cover enough different cases [2, 19, 27, 39, 42]. Nonetheless, with a few exceptions (e.g., [8, 9, 34]), another perspective on bias remains largely unexplored: potential biases introduced in the process of labeling training data. For example, it is known that the demographics of crowd workers may skew toward female and people in developing countries, given the opportunity to earn money in more valuable, foreign currencies [30]. Thus, training data can carry implicit biases from these subgroups because they are the majority available to provide labels, which can lead to most judgments being made by people who speak the same language, share the same gender, or are in a timezone where it is business hours when a task is launched. These biases can greatly impact different use cases, such as audio collection, content moderation, sentiment analysis, among other tasks where subjective judgments from humans are needed [3, 6, 8, 26, 34], and may be perpetuated via transfer learning [19, 27, 39], a popular practice in deep learning where a model can be repurposed and reused [27]. Therefore, a way to mitigate these biases must be developed so that requesters can create unbiased datasets for machine learning, since gold-standard datasets obtained via crowdsourcing can carry cultural biases dependent on crowd demographics [34].

We also believe the very nature of crowdsourcing for machine learning contributes further to dehumanization effects because the crowd is used for simply "filling in" labels to an unstructured, unlabeled dataset. In addition, a (much needed) common practice in these platforms is to use historical accuracy on "gold" units (i.e., units for which a label is previously known) or acceptance rate of previous tasks as a proxy for quality in order to filter out bad actors. In doing so, crowdsourcing becomes more and more *process-oriented* rather than *person-oriented*, although the latter has been deemed a better alternative for several reasons (e.g., [11, 21]). Consequently, a potential solution to address ethical issues in crowdsourcing for machine learning must encompass ethical issues in the *product* (i.e., the dataset) and the *process* of collecting data, which include but are not limited to implicit biases, underpayment, boredom, and incompatible tasks.

With this in mind, we propose a framework considering human factors in the process of labeling data for machine learning, addressing issues of crowd bias and ethics. Our framework was evaluated on our platform: Figure Eight [10] [1] (f.k.a. CrowdFlower), a platform primarily designed to support crowdsourcing tasks for gold-standard data used for machine learning. Requesters upload otherwise unstructured and/or unlabeled data and launch labeling tasks to the crowd.

We present the design and evaluation of the framework, which allocates labeling tasks to workers, referred hereafter as "contributors," based on different human-centric criteria such as contributor demographics and minimum wage in their country. We implemented the framework into the Figure Eight platform and evaluated it using three different machine learning use cases, with different requirements, namely, image categorization, content moderation, and audio transcription. We show that the use of our framework can mitigate demographic biases in contributor samples and increase contributor hourly pay. We discuss how our framework can be extended and used to promote transparency of human factors and "rehumanize" crowdsourcing.

2 RELATED WORK

Dehumanization in Crowd Work. Several issues related to dehumanization effects in crowdsourcing have been observed and addressed in prior works, including underpaid contributors (e.g., [13, 14, 20, 30, 40]), incompatible tasks (e.g., [4, 7, 15–17, 20, 21, 36]), tedious work (e.g., [12, 15, 29, 35]), and power imbalance (e.g., [32, 41]). As a result, human factors in crowdsourcing have been increasingly discussed by researchers. For instance, in reviewing human-centric issues in crowdsourcing, Gadiraju *et al.* [11] argue that human factors must be considered so that the humans behind the crowdsourcing tasks can be properly accounted for. These issues are largely overlooked by designers of crowdsourcing tasks, also known as requesters, as the established practice is to consider quality alone (e.g., accuracy, work approval rates), although human-centric approaches have been shown to improve contribution quality (e.g., [21]).

Nonetheless, solutions proposed in the past have addressed these individual ethical issues in isolation, whereas we hope the design of our framework will accommodate solutions for most of these issues simultaneously.

Biases from Crowd Work. In regards to biases originating from crowd work, several prior works have looked at biases introduced by the process of labeling via crowds (e.g., [3, 5, 6, 8, 9, 23, 24, 26, 34]). More related to the biases of interest in this work, it has been suggested that cultural differences in the crowd can affect algorithmic accuracy when gold-standard datasets used in machine learning applications are created via such crowds [6, 34], and that such differences may be introduced by implicit associations from different demographics. For example, Dong and Fu found that European-Americans and Chinese contributors can tag images differently [8]. In regards to gender, Otterbacher *et al.* [26] found that subjective judgments can be affected by contributor attitudes, showing that sexist people are less likely to detect and report gender biases in image search results. In another example, Nguyen *et al.* [23] showed that gender detection is difficult because of implicit associations and social constructions that take place in the annotation process. These issues are especially relevant when crowdsourcing is used to label data that are used by machine learning models, and therefore are addressed by our framework.

While the issue of dataset bias has been extensively investigated from the perspective of the data samples themselves (e.g., [2, 19, 27, 39, 42]), for example, when a facial recognition dataset does not include samples of faces from people of different races and ethnicities, prior works so far have only hinted and encouraged researchers to study the impact of biases in contributor demographics in crowdsourcing (e.g., [34, 38]). Our framework considers these human factors to enable requesters to obtain unbiased contributor samples.

[1] https://www.figure-eight.com/platform

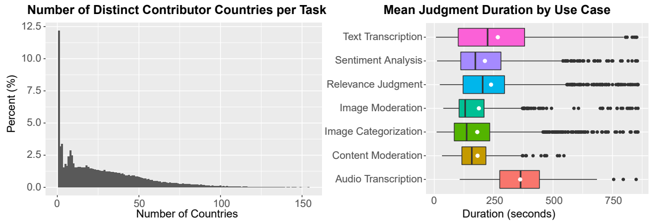

Figure 1: Left: Countries per task (only tasks not targeting a country): about 15% of all tasks are labeled by at most five distinct countries. Right: Average judgment time for seven use cases (white dot = average): time spent on task varies greatly by use case and individual.

3 THE FRAMEWORK

Motivation

In addition to prior works, several additional problematic scenarios involving human factors have been observed in our platform and thus have greatly motivated our design. For example, the hiring of contributors can be affected by temporal changes in the available crowd and/or the demographics of those drawn to each platform [24]. For instance, a major economic crisis in Venezuela caused many people to sign up to our platform in order to earn money in a more stable currency, biasing the available workforce when most contributors are from the same country, culture, and speak the same language, which can be problematic for Natural Language Processing (NLP) tasks. These can not only bias the contributor sample, but also cause underpayment and frustration when available tasks are incompatible with the skills of those attempting to complete them, with completion times varying greatly for any given task (see Figure 1). Biases can also be introduced because of the time of the day when a task is launched, being completed by those located where it is business hours. Such demographic biases can also pose issues to data relying on subjective judgments of a subgroup [23, 26] in tasks such as sentiment analysis and content moderation. For example, a machine learning model used for content moderation would be trained on data collected by a crowdsourcing platform which has over 70% of its available workforce as male. The model may be biased toward a male view of what may be considered offensive or inappropriate. In addition, cultural background can affect subjective judgments [8], which may pose issues when tasks involve judgments on politics and religion.

Our intention in showing these scenarios is to make clear where the motivation for our work comes from. It is also to show that more often than not, historical accuracy or acceptance rates alone are far from sufficient when dealing with tasks that aim at collecting training data for machine learning models. These scenarios are not hypothetical – they do take place often in our platform. In addition, as has been shown elsewhere [11, 20], relying on process-centric metrics such as accuracy leads to dehumanization effects in crowd work. With this in mind, we present our framework, which mitigates these issues in the process of data labeling for machine learning, but also in crowd work more broadly.

Design

The framework's ultimate goal is to help a requester meet desired arrangements for a task, allowing them to specify different settings related to human factors before launching it – with transparency. That is, the requester will be able to see how different arrangements for demographic distributions impact one another as well as what biases could be introduced in the training data or likely ethical issues (e.g., underpayment). Our philosophy is that, instead of taking sides and defining which biases are wanted and which are not, our approach is to let a requester decide how "diverse" or "skewed" the distribution of a certain contributor demographic must be for a given labeling task. To illustrate when certain biases may be desirable or undesirable, consider a requester who wants to label comments for an online discussion forum in which the number of male and female active users is close to equal. It is important for this requester that the training data for the model performing the content moderation include the perspective of both male and female contributors, or the dataset may be biased. Likewise, diversity may be needed when collecting training data for a personal voice assistant in form of audio, where variations of accent, gender, age, and native language are crucial. In a different scenario, consider a requester collecting training data for a search relevance model to be used in an online shopping website where 90% of the user base is female. For this requester, gender bias may be desirable in the training data.

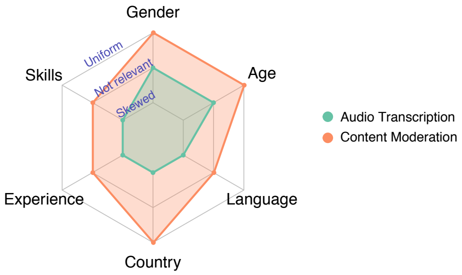

Figure 2: Example task configuration for two use cases. For audio transcription, gender and contributor age are not as relevant as contributor skills, experience, and language. For content moderation, diversity of gender, age, and country is more important than skills, experience, and language.

In being transparent, not only requesters can make decisions about trade-offs, but they are also made aware of potential biases or ethical issues that may be introduced if a task is launched at any point in time, thus mitigating potential dehumanization effects. We believe that mitigating biases and ethical issues in this process are parallel goals to rehumanizing crowd work, and that such a framework will ultimately contribute to rehumanizing crowd work via increased transparency in regards to how human factors affect the work to be completed on the platform and the resulting labels that will later be used in machine learning models.

Figure 2 shows examples of different arrangements for two different labeling tasks. By setting Experience to uniform, the framework will show the tasks to contributors with different levels of experience in the platform. By setting Skills to uniform, contributors with different sets of skills obtained through working on different types of tasks will be selected. When specifying a skewed metric, the requester is asked to select which values to be used along with the percentage (e.g., 80% male, 20% female). While targeting workers by demographics has been a feature of crowdsourcing platforms, differently, our framework attempts to automatically optimize the final distribution of demographics (e.g., uniform).

Once an initial configuration is given, our framework then attempts to "approximate" the desired configuration for the task, essentially translating the task assignment process into a multi-objective optimization problem, doing the best possible to achieve the desired distribution of contributors working on a task. Besides allowing requesters to specify arrangements, the framework can also have defaults for which it always optimizes, such as minimizing underpayment and maximizing the historical accuracy of those who are selected. The framework also allows different "goals" for different use cases (e.g., sentiment analysis, audio transcription, image moderation) as well as for new metrics to be added easily in the future (e.g., optimize for task novelty, optimize for learning). We note that a limitation of such framework is that in promoting transparency, requesters could intentionally misuse it to do the opposite of what the framework strives for, for example, by allowing them to exclude people who identify with one gender, or less experienced contributors. However, we believe that more often than not, doing so may result in machine learning models that do not perform well on the intended target audience. In our study, we used the framework to ensure diversity of contributor gender, age, country, minimize the pay gap, and maximize historical accuracy of contributors hired for the tasks.

System Implementation

We implemented the proposed framework as a live system in the back-end of our crowdsourcing platform: Figure Eight [10]. We created a system with Python that automatically and iteratively selects contributors who are online for a task so that the "right" contributors (based on the desired arrangement) for the task are hired at every optimization step until the task is complete (i.e., all needed labels are provided).

An optimization step consists of obtaining data from our databases about what contributors are online at a given time, along with their demographics, as well as data about contributors who already worked on the task, along with their demographics, and then selecting suitable contributors so that the desired arrangement is successfully achieved via a multi-objective optimization algorithm. At each step, after identifying optimal contributors for the task, the system attempts to hire contributors by creating Manual Custom Channels [2], which is a feature in our platform to target contributors by their respective contributor IDs. For example, if a uniform distribution of gender is desired and currently more males than females completed the task, in the next step, the framework will automatically attempt to hire more females in order to "approximate" the desired configuration.

The demographics we used were voluntarily provided by contributors when they created an account on our platform and were obtained through protected and authenticated access to our own databases. Such data about our platform's contributors were not publicly available to requesters at the time of our study. The selection process consists of qualifying a contributor for a task so that only those selected at that step can work on the task next. In our study, the selection of contributors took place every 20 minutes after the task was launched, until the task was complete. Algorithm 1 describes the underlying task assignment procedure:

Algorithm 1 Contributor-Task Assignment

```
while task not complete do
    $current$ ← set of contributors who worked on task
    $online$ ← set of online contributors
    $selected$ ← $\emptyset$
    $states$ ← set of states $\{State_0 \dots State_n\}$ if $online_i$ is added to $current$ for each contributor in $online$
    $selected$ ← Pareto-optimal set of best $n$ online contributors when $n$ contributors are added to $current$ as given by Pareto-optimal states in $states$
    recruit($selected$)
end while
```

where n is in the range [0, sizeof(online)) and $state_i$ is defined by points in multidimensional space composed of metrics to be optimized, such as normalized entropy of a probability distribution (e.g., distribution of gender, country,

[2] https://success.figure-eight.com/hc/en-us/articles/115005698186-Custom-Channels-Feature

Use Case	# contributors (# judgments) **Original**	 **Baseline**	 **Framework**	Country	Gender	Age	Accuracy	% Min. Wage
Image Cat. (L1)	308 (50,090)	430 (52,213)	297 (17,171)	⊕	⊕	⊕	↑	
Image Cat. (L2)			184 (7,980)	⊕	⊕	⊕	↑	↑
Content Moderation	565 (18,139)	205 (6,275)	241 (6,705)		⊕		↑	
Audio Transcription	47 (880)	11 (382)	24 (510)				↑	↑
	920 (69,109)	646 (58,870)	746 (32,366)					**2,312 (160,345)**

Table 1: Study Design. In the end, 2,312 (1,919 unique) contributors participated in the crowdsourcing experiments, providing 160,345 judgments. ⊕ = attempt to approximate uniform distribution ↑ = attempt to maximize.

and age) and continuous variables (e.g., mean historical contributor accuracy, percentage of minimum hourly wage, size of set *current* ∪ *selected*, used as objectives to be minimized or maximized. Objectives could have different weights, but we used equal weights in our study. The normalized entropy of a probability distribution was used because it grows as the distribution gets closer to uniform. We used the algorithms provided by PyGMO [28] to obtain Pareto-optimal solutions. In the very first time, *current* is a random sample of online contributors. In subsequent steps, until no one worked on the task, this random sample is used as *current*.

Due to the fact that not all contributors have provided demographic data when they signed up, local optima could occur when a distribution is already uniform, leading the framework to recruit those who did not provide a demographic as to not change an already optimal distribution. To deal with this, the framework temporarily disables the goal for which the distribution is already uniform at a given step. Nonetheless, demographics for about 65% of contributors in our platform have been voluntarily provided.

4 EVALUATION

Study Design

The goal of our study was to evaluate the impact of the framework in mitigating demographic bias (e.g., gender, age, country) in the resulting contributor sample, while optimizing the pay according to minimum wage and keeping comparable contribution quality. We evaluated our framework using a "within-task" and a between-subject design. That is, we selected one previously completed task from three popular use cases in our platform, namely image categorization, content moderation, and audio transcription, and relaunched these tasks in our platform under two conditions: *without* our framework (i.e., the baseline condition) and *with* our framework. The task for each use case is an actual task which was completed in our platform in the past, created by different requesters. The original tasks were launched about 3 months prior to the baseline and framework tasks. By relaunching them, we repeated the tasks with the same set-up (i.e., the same data provided in the original task, the same pay, the same number of judgments requested). Two of the three tasks were created by academic institutions while the other was created by an Internet company. The choice of use cases and tasks was also influenced by their potential to evaluate different goals for each task. For example, in the image categorization task, we set up the configuration to approximate uniform distribution of countries, gender, and age, whereas for the audio transcription task, which consisted of transcribing audio to text, we set up the framework to maximize the percentage of the pay according to minimum hourly wage in each country. In our study, we took the role of the requester by setting up these different configurations for each task, since our system was implemented in the back-end of our platform. This means we did not involve any requesters in our study. In considering different conditions, we also considered the original task in our analysis as a condition which we refer to as *original*, ultimately comparing three conditions: (a) the original task, (b) the baseline task (without the framework), and (c) the framework task. To maintain contribution quality, by default, all tasks were set to maximize the mean historical accuracy of contributors on ground-truth units. Tasks (b) and (c) were launched on the same week (which led to 393 contributors working on at least two tasks). We included task (a) (i.e., the original task) for each use case for a more conservative analysis in which we expected tasks (a) and (b) to produce similar results. Table 1 shows the configurations we used for each task. For the Audio Transcription baseline and framework tasks, we filtered contributors so that only those from countries whose English is one of the official languages were considered (90 countries). We did this for two reasons: (1) it makes sense for the task, and (2) we wanted to observe the framework at work when a filter was also in place.

Tasks

The image categorization task consisted of showing profile photos of users to contributors and asking them to provide the gender, ethnicity, and an emoji that closely matched the skin tone of the person in the photo, via multiple choice questions, for 10,000 data units. Contributors were paid $0.01 (USD) per judgment provided for this task, doing 10 judgments at a time. If this task ends up with skewed contributor demographics, a model classifying people would carry biases from that group. For example, the judgment of Black or White may differ between contributors in India and the U.S.

The content moderation task involved contributors judging whether a response to a forum post contained toxic content, also asking contributors to mark whom the attack was targeted to (e.g., a user on the thread, a group of people), with multiple choices, for 4,022 data units. In this task, contributors were paid $0.25 (USD) per judgment, doing 5 judgments at a time. If this task has skewed contributor gender, a model may miss out on content deemed toxic to the other gender.

Finally, the audio transcription task consisted of having participants provide text to match audio recordings which they listened to (e.g., "Tom is talking about the fee"), for 119 data units. Participants were compensated with $0.10 (USD) per judgment, giving 5 judgments at a time. Given the time to do this task, if pay gap is not optimized, contributors from countries where the minimum wage is far greater than what the tasks pays will be largely underpaid.

Data Analysis

In our data analysis, we compared the distribution of demographics of individual contributors who worked on the tasks as well as the continuous variables to be maximized (e.g., historical contributor accuracy, percent of minimum hourly wage). We also evaluated the impact of the framework on contribution quality by comparing accuracy of judgments provided to gold-standard data. In addition, we evaluated the difference in the distribution of labels given by individual contributors in each condition. This was so that we could understand whether the demographic optimization made by the framework would change the number of decisions made in favor of one label or the other. In the future, rather than counting the number of contributors in each subgroup, it may also be beneficial to control by number of judgments, if some contributors can provide more judgments than others.

In order to assess a contributor's quality/trust, our platform keeps track of historical accuracy on ground truth/gold units provided by requesters to validate the quality of their work as they undertake tasks. These data instances are called "test units" and are randomly picked and presented to contributors as Test Questions [3] in "quiz mode" (i.e., before labeling begins as a qualification step) and "test mode" (i.e., as attention checks during labeling). Our platform integrates these ground-truth units automatically in the process of completing the task in order to assess the quality of the contributions and determine whether the work will be accepted, also using the all-time accuracy of contributors on these data units as an indicator of a contributor's work quality/reputation [1].

Therefore, in order to assess and compare the quality of contributions when using the framework, we considered two metrics. The first is the historical accuracy of a contributor on ground-truth units of all previously completed tasks. This was automatically optimized for along with the other metrics in the system (see Table 1). In addition to using contributors' historical accuracy as a metric for their work quality, we also evaluated the percentage of "incorrect" judgments on each test unit (i.e., ground-truth data) provided in each study task for quality control. These units were originally provided by the task requesters and were used in all of the conditions evaluated in our study. The content moderation task had 84 of such units, the image categorization task 87, and the audio transcription had 20. Accordingly, we evaluate and compare the mean percentage of incorrect responses in all conditions to assess contribution quality.

Due to its experimental and exploratory nature, our system was not implemented in our main production technology stack and therefore it did not have access to our production databases in real-time. Nonetheless, our system is a live system and integrated into our platform using our data and infrastructure. Due to not being part of our production systems, our system only had access to a data warehouse that was "behind" at least 8 minutes – the delay to migrate from production to the data warehouse – with no guarantee of synchronization. This caused a throughput issue specific to our implementation. More specifically, this caused our task assignments to target some contributors who may have gone offline at each step, greatly impacting throughput of the framework tasks. For example, the original image categorization task took 25.5 hours to complete, with the baseline finishing in 33.4 hours. We stopped collecting judgments for the framework task after 145 hours (29.4% complete). This task also had the slowest throughput due to the low pay assigned by the original requesters. Similarly, the content moderation task had the original complete at 34 hours, while we stopped collecting judgments after 87.7 hours (36% complete) in the framework task.

To accommodate for the throughput limitation in our data analysis, we capped the number of contributors in the original and baseline conditions to match the number of contributors who contributed to the framework task, taking the first *n* contributors from the other conditions, where *n* is the number of contributors who worked on the framework task before we paused it. Doing so is also beneficial for visualizing how biases may start taking place as soon as the tasks are launched without the framework, when most contributors

[3] https://success.figure-eight.com/hc/en-us/sections/200596719-Test-Questions

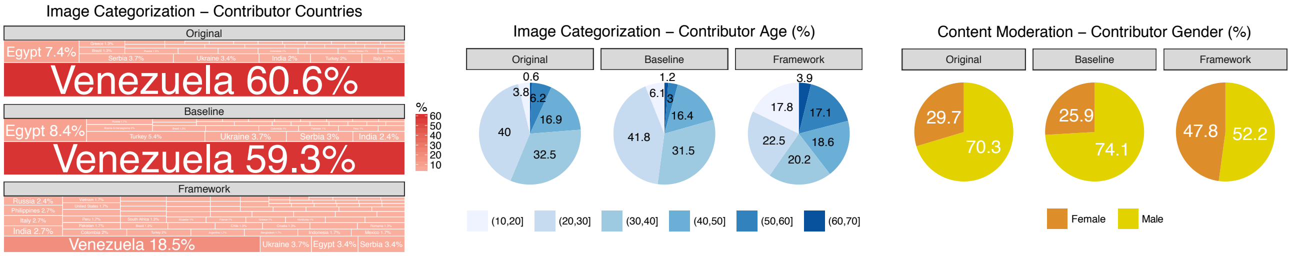

Figure 3: Left: Percentage of contributors from each country in the three conditions for the image categorization task. Middle: Percentage of contributors from each age group in the image categorization task. Right: Percentage of contributors from each gender in the three conditions for the content moderation task. Demographics in baseline and original tasks were biased by demographics of online active users of the platform.

who are online will complete the task very quickly. In the end, our data analysis had 297 contributors (96% of total in the finished original task) when analyzing the first launch of the image categorization task, and 184 (50% of total in original) contributors when analyzing the second. For the content moderation, the cap was at 206 contributors (36% of original). We also report the final state of the finished original tasks (e.g., distribution of demographics, average percent of minimum wage).

For the calculation of percentage to the minimum wage in our analysis, we first calculate the hourly pay of a contributor given by the pay per judgment divided by the average judgment duration in seconds, multiplying the result by 3,600 (number of seconds in an hour). Then, we divide this amount by the minimum hourly wage in the contributor's country. During the task assignment steps, this quantity was set to be maximized, with the reference average judgment duration being the average judgment duration of those who already worked on the task as a proxy for the expected pay, or the average judgment duration of the contributor in the past 30 days, if no one worked on the task before. We obtained minimum hourly wages by country from the International Labour Organization [25] and converted to U.S. Dollars using the *quantmod* [31] R package, which uses Yahoo Finance.

5 RESULTS

The framework was able to mitigate potentially unwanted demographic biases introduced by the labeling crowd while minimizing underpayment and keeping comparable contribution quality. We present the results of our study comparing the tasks launched originally, the tasks launched without our framework (i.e., the baseline condition), and the tasks launched with our framework. We emphasize once again the the tasks in the three conditions followed the exact same design (e.g., same data, pay, instructions, test units).

Demographic Biases

Our results show that the framework was successful in approximating the distribution of different demographics to the configuration for each task, effectively minimizing the likelihood that the distribution of any demographic was very skewed towards a certain subgroup. This is further supported when we look at the differences among the original task and the baseline task. In other words, despite being launched several months apart, the original and the baseline tasks yielded very similar results. Figure 3 shows differences in the demographics of contributors who worked on the tasks, which are described in more detail below.

Country of Origin. When using the framework in the image categorization task, contributors from 74 unique countries provided judgments, whereas this number was 33 in the baseline task and 39 in the original task. The country from which most contributors came from was the same in the three conditions: Venezuela. However, the percentage of contributors from Venezuela was 18.5% with the framework, compared to 59.3% in the baseline and 60.6% in the original. In the finished original task, contributors from 39 distinct countries provided judgments, with the top country (Venezuela) having 60.7% of contributors.

Gender. The distribution of contributors from each gender was closer to uniform in the framework task for image categorization, with 50.3% being male and 49.7% being female, whereas the distribution was 72.7% male and 27.3% female in the baseline task, and 68.3% male and 31.7% female in the original task. Even in the finished original task, 67.5% of contributors were male and 32.6% female.

Likewise, the framework was also effective in the content moderation task, for which when using the framework, 47.8% of contributors were female and 52.2% male, whereas 74.1% of contributors were male and 25.9% of contributors were female in the baseline condition, and 70.3% being male and 29.7% being female in the original task. This was similar in the finished original task for content moderation, with 72.9% being male and 27.1% female.

Age. The distribution of contributor age in the framework task for image categorization was also closer to uniform, with the age group with the most contributors being between (20,30] years old with 22.5%, with other percentages being 20.2% (30,40], 18.6% (40,50], 17.8% (50,60], 17.1%, and 3.9% (60,70]. In the baseline condition, 41.8% of contributors were between 21-30 years old, with the same age group having 40% of contributors in the original task, in other words, the age distribution very skewed toward individuals in their 20s. Results were also skewed in the finished original task with 40.2% in the (20,30] years old range.

These results suggest that when tasks are launched in our platform without the intervention of the framework, the distribution of demographics of those who will work on the task are likely to be biased towards the demographics of the active users in the platform (e.g., Venezuela, 21-30 years old, male), and the framework mitigates this.

Pay Gap and Quality

In addition to optimizing the distribution of contributor demographics, the framework also attempted to minimize underpayment in two tasks, while maximizing historical accuracy of those recruited for all tasks. Our results show that the framework can minimize the pay gap of contributors working on the task, paying contributors closer to minimum wage in their country when the task pay is low, and making the task more profitable when the task pay is already good – without changing the pay of the task. Also, even when optimizing the distribution of demographics, the framework was still able to select high-quality contributors for the task, albeit a slight decrease on accuracy of ground-truth labels was observed.

Pay Gap. We optimized the task to minimize the pay gap in two use cases: image categorization (Launch 2) and audio transcription. Minimizing the pay gap may also be referred to as maximizing the percentage of the pay relative to the minimum wage in each contributor's country. Figure 4 shows the comparison of the pay gap among the conditions.

More specifically, for the image categorization task, we launched a second task with the framework optimizing for the same demographics as before, but this time around also adding the percent of the minimum hourly wage as a metric to be maximized. This new framework task resulted in the mean percentage to the minimum wage of 44.9% (Mdn = 22.7%, SD = 48.2%, Min = 1%, Max = 284%), with this figure being higher than the other three conditions: the first framework task had 28.1% as average percent to minimum wage (Mdn = 16.2%, SD = 35.8%, Min = 0.2%, Max = 237%), the baseline task at 28.3% (Mdn = 35.5%, Min = 1.1%, Max = 201%), and the original task having contributors being paid an average of 31.5% the minimum wage of their country (Mdn = 15.5%, SD = 43.7%, Min = 1%, Max = 220%). The finished original task had mean percent of minimum hourly wage at 25.7% (Mdn = 14%, SD = 36.6%, Min = 0.5%, Max = 219.7%).

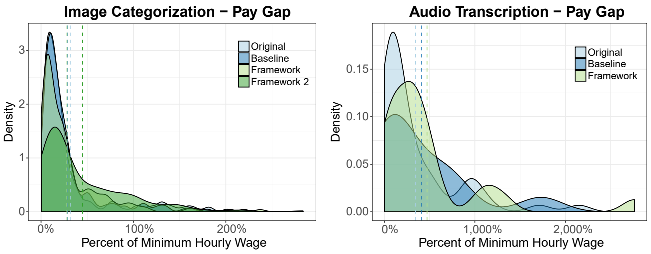

Figure 4: Density of percentage of minimum wage for image categorization (left) and audio transcription (right) tasks. Means are represented by vertical lines. When using the framework, contributors were compensated closer to minimum wage when pay is low (left) and completed more profitable work when pay is already good (right).

For the audio transcription task, we targeted only English-speaking countries. In addition to filtering by country, the framework configuration only had percent of minimum hourly wage and historical accuracy to be optimized. The average percentage of the minimum wage in the framework task was 471% (Mdn = 306%, SD = 603%, Min = 3%, Max = 2760%), while being 407% for the baseline condition (Mdn = 173%, SD = 521%, Min = 22.8%, Max = 1730%), and 346% in the original task (Mdn = 163%, SD = 462%, Min = 17.5%, Max = 2150%). Two reasons that percentages are high in the audio transcription task: (1) the task pays better and (2) most contributors were recruited from countries with a very low minimum hourly wage relative to the U.S. Dollar e.g., Ghana ($0.17), Egypt ($0.23), India ($0.25), Pakistan ($0.47), Kenya ($0.77).

Quality. In addition to optimizing for demographics and hourly pay, the framework attempted to maximize the historical accuracy of those who were selected for the task, serving as a proxy for quality or trust in the contributor. For both the image categorization and content moderation task, the historical accuracy of contributors was comparable and sometimes even higher when the framework was used. For example, in the framework task for image categorization, the average historical accuracy was 0.9, compared to 0.86 in the baseline condition and 0.89 in the original task. For the finished original task, the mean historical accuracy was 0.89.

Similarly, in the content moderation task, the accuracy in the framework task was higher at an average of 0.97, being 0.87 in the baseline task, and 0.87 in the original task. In the finished original task, the mean historical accuracy was 0.87.

The accuracy of contributors in the audio transcription task with the framework is unknown because none of the 24 contributors had historical accuracy on file, but this number was 0.81 for the baseline task, and 0.86 for the original task.

These results suggest that even when optimizing for human factors, the quality of contributions is comparable. In other words, by using our framework, we did not recruit more contributors who are historically "untrusted" or "inaccurate" (e.g., bots, random clickers).

On other hand, when comparing the percentage of incorrect judgments given to the test units (i.e., the ground-truth data), tasks where the framework was used showed a slight increase in incorrect judgments given to ground-truth data. More specifically, framework tasks had an average of 10% of incorrect judgments per test unit (Mdn = 8%, SD = 10%, Min = 0%, Max = 67%), while this number was 5% for the baseline (Mdn = 3%, SD = 8%, Min = 0%, Max = 55%), and 7% for the original (Mdn = 4%, SD = 8%, Min = 0%, Max = 38%). In other words, framework tasks had a 5% increase in incorrect responses given to ground-truth units provided by the original requesters compared to the baseline and 3% compared to the original tasks.

Distribution of Labels

In addition to more evenly distributed demographics among selected contributors, the resulting data (i.e., the labels) provided by contributors were also different when the framework was used. We looked at individual judgments provided by contributors on the same units (i.e., rows) of the data.

The image categorization task consisted of providing the skin tone and ethnicity for a profile photo of a person. The distribution of skin tones given by contributors in the framework task was different from the original and baseline condition, with fewer judgments given for the two extremes of skin tone. For example, while the original and baseline tasks resulted in 16.6% and 21% of judgments assigning 3 as the skin tone on a scale of 1 to 5, the framework resulted in 24%. Likewise, both the original and baseline tasks resulted in 15.3% of judgments assigning 5 as the skin tone, compared to 10.9% in the framework task. The distribution of ethnicity given by contributors was very similar, with the differences being within 1% among each category. This was likely due to the dataset being unbalanced, with more photos of White and Black/African American individuals.

In the content moderation task, for which the judgment is more subjective, the distribution of content deemed toxic in the framework task was very different from the other two tasks (i.e., the original and the baseline). In the framework task, which attempted to select contributors so that the final distribution of gender was closest to uniform, 40.3% of the judgments indicated that the content in the comments was a personal attack or deemed toxic, with 56.8% otherwise, and 2.9% unsure. Differently, the baseline task had 34.2% of the content marked as toxic, 65.3% otherwise, and 0.5% unsure, and the original task had 33.4% marked as toxic, with 64.8% marked otherwise, and 1.7% unsure. In other words, more content was marked as toxic by contributors when judgments were distributed more evenly among contributors of both genders. We do not claim causation in this result, but we do highlight how it can benefit scenarios where training data must be aligned with potential moderation scenarios where a model must not be biased by views of any one gender.

6 DISCUSSION

Our results show that our framework can mitigate biases in the resulting contributor sample while maintaining work quality and minimizing the pay gap when launching tasks aimed at labeling datasets for machine learning applications. This has important implications for the effectiveness of machine learning applications in the real world [34], especially when subjective opinions and judgments are involved in the process of data labeling. We discuss our results in more detail and directions for future work.

Moving Beyond Historical Accuracy

In our platform, contributors are leveled based on their historical accuracy e.g. Unleveled, L1, L2, and L3. This is also a common practice in other crowdsourcing platforms (e.g., AMT) where often acceptance rates are used in an attempt to obtain high-quality responses [1]. Our study shows that accuracy and acceptance rates alone are not appropriate indicators of training data quality and that the human behind the label must also be considered in order to mitigate issues of bias that can limit the performance of machine learning models used in the wild. In addition, in our study, contributors from any level were recruited so long as they were optimal as determined by the framework, without launching the tasks to any particular level. Therefore, our results indicate that it is possible to maintain work quality while mitigating biases.

Our framework was designed in a way that allows other goals to be easily incorporated. Its implementation will always attempt to make the best possible choice at a given time, making trade-offs as needed. One potential goal that could be introduced is the idea of skill ladders proposed by Bigham *et al.* [15]. For example, for a task that requires no special demand of cognitive abilities, such as image categorization of objects, one goal in the framework could be to select contributors so that the final distribution of contributor experience approximates a uniform distribution. This gives newcomers the opportunity to gain experience while mitigating scenarios where most labels are provided by the few most experienced contributors, which is commonplace.

Still on the idea of growth and engagement, contributors could be selected to maximize task novelty, that is, the percentage of contributors for which a task is very dissimilar from other tasks completed before, thus preventing contributors from getting stuck doing the same work for long periods of time. One caveat is that it is likely that contributors may spend more time doing an unfamiliar task, which can introduce conflicting situations when, for example, maximizing hourly pay is one of the other goals.

Among many possible improvements to our framework, one of them is adding the ability for it to automatically identify demographics that need to be optimized. For example, in identifying that the judgments about the same data units given by male and female contributors differ, it could learn that it is important to recruit with diversity in this case. This would help alleviate throughput issues and avoid the inclusion of unnecessary goals into the optimization process.

When implementing the framework, we argue that it must be done with transparency in mind so that not only requesters can make more informed decisions about biases and ethics, but also be made aware via soft paternalistic nudges [33] to encourage desirable behavior. For example, a requester launching a task within a U.S. timezone may be nudged that the price that they are setting for a task may be incompatible with the contributors available at the moment, which are in the U.S. Possible solutions may include suggesting to raise the pay or launching the task when the pay is more in line with online contributors' minimum wage. More importantly, when implemented, the framework could help requesters construct a recruitment plan based on historical data from the platform, automatically identifying potential biases and desired configurations to mitigate them under different use cases (e.g., audio collection, sentiment analysis), creating effective templates and defaults that minimize issues of bias and ethics by design. In exercising transparency in this manner, crowd work could be rehumanized, especially when used for machine learning purposes. Nonetheless, designers must be careful so that purposeful misuse such as excluding subgroups (e.g., gender) and unethical hiring (e.g., paying less than minimum wage) can be prevented. One idea is to enforce defaults such as always maximizing the percentage to the minimum hourly wage, and nudge requesters about exclusions and sample biases prior to launching a task.

Given that a large fraction of unpaid work is due to the time finding tasks [13], when our framework is implemented, contributors could be notified that they have a task for which they qualify based on the configuration set up by the requester – even if they are offline. For example, a contributor over 60 years old may receive an e-mail asking for their contribution because the task needs a perspective from that age group. This could reduce the effort spent by contributors to find good work (see [14, 18, 40]) and increase their motivation to work on the tasks [29].

Conflicts and Trade-offs

In any multi-objective optimization problem, conflicting states are likely to occur. This motivated our choice of applying Pareto-optimal selection so that these trade-offs could be accounted for. In turn, when conflicting states are present, labeling throughput can be affected. For example, consider a setup where the number of distinct countries is to be maximized while also minimizing the pay gap. In our platform, this can occur when most contributors online are from countries that would make the pay gap minimal, but selecting them would introduce contributor country bias.

As observed in our experiment, there are trade-offs between *labeling with less bias* and *completing the task faster*. For example, because the distribution of the available workforce is inherently biased towards the active users of the platform, it is possible that during many steps the number of optimal contributors to be selected will be small, which in turn contributes to longer task completion times (i.e., takes longer to obtain all labels). For this reason, we created a neural network model to forecast the changes in the demographics 24 hours into the future. This model helps with the problem of throughput, by choosing launch windows that are in line with the desired configuration for a task. For example, the framework could schedule to launch a task at the time the number of distinct countries is the largest within the next 24 hours, if the desired arrangement for a task is to maximize contributors' number of distinct countries. Another potential solution to reduce throughput is to maximize the likelihood that a contributor will do a task when they are assigned, based on historical data, but one must be careful with biases in doing so, in case the majority of contributors who are more likely to work on it are from the same country.

Although minor, our findings point to a possible decrease in contribution quality when using our framework, as indicated by a 3-5% increase in the average percentage of incorrect responses given to ground-truth units in the tasks where our framework was used. This may have had to do with the fact that the framework tasks did not necessarily recruit the most active/experienced contributors because they would certainly bias the demographic distribution by being a from single country and/or gender. This increase in incorrect responses may translate into additional costs for task requesters and should be further explored in the future.

In experimenting with the initial demographics selected for our study, we came across a limitation where we did not have a reliable source of a demographic – the language of contributors. We had the language in which they use the platform as well as the language from their browsers, but we decided that this was not enough to be able to secure a selection criterion, therefore not using it in our study. This raises an interesting implication, which is, while it is possible to improve crowdsourcing by considering human-centric attributes such as demographics, it will also require platforms to collect more personal data, which may raise questions related to the privacy of contributors. In addition, contributors connected to Virtual Private Networks (VPNs) may mask their country, which may result in ineffective hiring.

Our choice of optimizing the selection process in real time – considering only online users – was so that we could assign tasks to both newcomers and active contributors. Given that contributors spend a considerable amount of time trying to find good tasks to work on, an alternative approach would also assign tasks to those who are offline, potentially sending them a notification that they have a good task waiting for them. Nonetheless, this would require careful thought, or exclusion of certain groups can occur (e.g., only recruiting active users in the last 30 days).

Limitations and Future Work

The personal data used in our experiments are voluntarily provided by contributors. Although about 65% of contributors do provide demographics, not all contributors do so, and there is no verification for it. This can lead to cases where the framework recruits contributors with missing data because it will not affect an optimal state. This may cause biases that are not possible to visualize once a contributor whose gender or age is unknown is recruited. Nevertheless, given the formative nature of our experiment we find that this is an acceptable limitation, given that the framework would be able to perform equally well if demographics were available for all contributors. We decided to consider contributors with missing data in order to increase throughput, otherwise we would not be able to collect enough data for our experiments.

Given that this was a research endeavor, our implementation was done causing the minimum disturbance possible to our platform. When our framework is incorporated into the platform in a more seamless way (i.e., part of the consolidated technology stack), the limitation of throughput will be greatly mitigated, for example, with shorter intervals between steps, giving it a quicker response time and targeting contributors who are actually online.

We are working on making our framework available for requesters to use on our platform, which will give them control and awareness of human-centric aspects in the process of manually labeling data for machine learning. We are also promoting a campaign in our platform to create contributor profiles before adding the framework as a platform feature.

7 CONCLUSION

The process of labeling data via crowdsourcing can promote dehumanization via unfair compensation, incompatible task assignments, and unintended amplification of human biases. To address these issues, we designed and evaluated a crowdsourcing framework, introducing more transparency and helping requesters achieve their labeling goals with human factors in mind. We conducted several crowdsourcing experiments on a popular crowdsourcing platform with 1,919 contributors (a.k.a. workers), collecting 160,345 judgments for labeling tasks related to machine learning use cases. We show how our framework can mitigate demographic biases in contributor samples and increase contributor hourly pay.

8 ACKNOWLEDGMENTS

We thank the Figure Eight team for their continuous support and feedback. We also thank anonymous reviewers for their guidance and help. We acknowledge Anthony Ikeda, Humayun Irshad, Phoebe Liu, Tong Liu, Qazaleh Mirsharif, Robert Munro, Rahul Parundekar, Jennifer Prendki, Amir Raffiee, Kiran Vajapey, James Wachira, and Joan Xiao.

REFERENCES

[1] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17, 2 (2013), 76–81.

[2] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 771–787.

[3] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202* (2013).

[4] Piyush Bansal, Carsten Eickhoff, and Thomas Hofmann. 2016. Active content-based crowdsourcing task selection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 529–538.

[5] Justin Cheng and Dan Cosley. 2013. How annotation styles influence content and preferences. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, 214–218.

[6] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[7] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. 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*. ACM, 367–374.

[8] Zhenhua Dong, Chuan Shi, Shilad Sen, Loren Terveen, and John Riedl. 2012. War versus inspirational in forrest gump: Cultural effects in tagging communities. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[9] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 162–170.

[10] Figure-Eight.com. 2018. Machine Learning, Training Data, and Artificial Intelligence Platform. Retrieved July 3, 2018 from http://www.figure-eight.com

[11] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems* 30, 4 (2015), 81–85.

[12] Ujwal Gadiraju, Patrick Siehndel, Besnik Fetahu, and Ricardo Kawase. 2015. Breaking bad: understanding behavior of crowd workers in categorization microtasks. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 33–38.

[13] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 449.

[14] Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 611–620.

[15] Nila Banerjee Jeffrey P. Bigham, Kristin Williams and John Zimmerman. 2017. Scopist: Building a Skill Ladder into Crowd Work. In *Proceedings of the Web for All Conference (W4A '17)*. ACM, New York, NY, USA, 10.

[16] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 1941–1944.

[17] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval* 16, 2 (2013), 138–178.

[18] Sarah Kessler. 2018. The Crazy Hacks One Woman Used to Make Money on Mechanical Turk. Retrieved September 18, 2018 from https://www.wired.com/story/the-crazy-hacks-one-woman-used-to-make-money-on-mechanical-turk/

[19] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*. Springer, 158–171.

[20] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[21] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1345–1354.

[22] MTurk.com. 2018. Amazon Mechanical Turk. Retrieved July 3, 2018 from http://www.mturk.com

[23] Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 1950–1961.

[24] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

[25] International Labour Organization. 2018. International Labour Organization - ILO Stat. Retrieved September 18, 2018 from https://www.ilo.org/ilostat/faces/wcnav_defaultSelection

[26] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 933–936.

[27] Sinno Jialin Pan, Qiang Yang, et al. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

[28] Pygmo. 2018. PyGMO - A Scientific Library for Massively Parallel Optimization. Retrieved September 18, 2018 from https://esa.github.io/pagmo2

[29] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *ICWSM* 11 (2011), 17–21.

[30] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.

[31] Jeffrey A. Ryan and Joshua M. Ulrich. 2018. *quantmod: Quantitative Financial Modelling Framework*. https://CRAN.R-project.org/package=quantmod R package version 0.4-13.

[32] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, et al. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 1621–1630.

[33] Jan Schnellenbach. 2012. Nudges and norms: On the political economy of soft paternalism. *European Journal of Political Economy* 28, 2 (2012), 266–277.

[34] Shilad Sen, Margaret E Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao Ken Wang, and Brent Hecht. 2015. Turkers, scholars, arafat and peace: Cultural communities and algorithmic gold standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 826–838.

[35] Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 275–284.

[36] Edwin Simpson and Stephen Roberts. 2015. Bayesian methods for intelligent task assignment in crowdsourcing systems. In *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*. Springer, 1–32.

[37] TaskRabbit.com. 2018. TaskRabbit connects you to safe and reliable help in your neighborhood. Retrieved July 3, 2018 from http://www.taskrabbit.com

[38] Jacob Thebault-Spieker, Daniel Kluver, Maximilian A Klein, Aaron Halfaker, Brent Hecht, Loren Terveen, and Joseph A Konstan. 2017. Simulation Experiments On (The Absence of) Ratings Bias in Reputation Systems. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 101.

[39] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*. Springer, 37–55.

[40] Kotaro Hara Toni Kaplan, Susumu Saito and Jeffrey P. Bigham. 2018. Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018)*.

[41] Mark E Whiting, Dilrukshi Gamage, Snehalkumar S Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, et al. 2016. Crowd guilds: Worker-led reputation and feedback on crowdsourcing platforms. *arXiv preprint arXiv:1611.01572* (2016).

[42] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).