Do I Look Like a Criminal? Examining how Race Presentation Impacts Human Judgement of Recidivism

Keri Mallari[†]
University of Washington
kmallari@uw.edu

Kori Inkpen
Microsoft Research
kori@microsoft.com

Paul Johns
Microsoft Research
pauljoh@microsoft.com

Sarah Tan
Cornell University
ht395@cornell.edu

Divya Ramesh[†]
University of Michigan
dramesh@umich.edu

Ece Kamar
Microsoft Research
eckamar@microsoft.com

ABSTRACT

Understanding how racial information impacts human decision making in online systems is critical in today's world. Prior work revealed that race information of criminal defendants, when presented as a text field, had no significant impact on users' judgements of recidivism [13]. We replicated and extended this work to explore how and when race information influences users' judgements, with respect to the saliency of presentation. Our results showed that adding photos to the race labels had a significant impact on recidivism predictions for users who identified as female, but not for those who identified as male. The race of the defendant also impacted these results, with black defendants being **less** likely to be predicted to recidivate compared to white defendants. These results have strong implications for how system-designers choose to display race information, and cautions researchers to be aware of gender and race effects when using Amazon Mechanical Turk workers.

Author Keywords

bias, recidivism, race, gender, crowd work, Mechanical Turk, legal, human-AI collaboration, human-centered AI

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI;** • **Applied computing** → **Law, social and behavioral sciences;**

INTRODUCTION

In today's online society, people are frequently asked to make judgements about others based on online profiles [12, 15, 2]. Whether it is rating Uber drivers, selecting applicants to interview for a job, or deciding whether someone gets approved or rejected for a loan, people are frequently asked to evaluate others.

In 2018, Dressel and Farid [13] compared human accuracy to the accuracy of a commercial AI system for recidivism prediction. They showed Amazon Mechanical Turk workers (whom we refer to as users in the rest of this paper) real defendant profiles and asked them to rate whether or not they thought that a defendant would commit a crime in the next two years (i.e. recidivate). Two different variants of the study were run, one where users were told the race of the defendant, and one where they were not. The authors found no significant differences in the results of these two variants and concluded that including race information had no impact on the accuracy or fairness of the users' decisions.

Concluding that defendant race has no significant effect on human judgements of recidivism is risky given the complexity of racial bias, especially on issues related to criminal justice [40, 29, 39, 41, 37]. However, Dressel and Farid's result is curious. Is it the case that race doesn't matter for predictions of recidivism? Probably not [47]. Is it the case that the users in Dressel and Farid's study were not biased? Probably not [49]. Is it the case that the Amazon Mechanical Turk workers were not engaged in the task such that race did not matter? Possibly, but given how common implicit biases are [44, 18, 25], it is unlikely that race would have no effect. These questions motivated us to investigate the validity and generalizability of Dressel and Farid's results and the implications it has on the design of evaluation systems, as well as studies utilizing Mechanical Turk workers.

Improving our understanding of biases related to human judgement in online systems is important given that the use of AI systems to assist human decision-making is growing, especially for critical domains such as medicine [5], criminal justice [14, 17], and social welfare [9]. Informed design of such systems and developing insights about potential biases they may introduce on decision-making

[†] Work done while interning at Microsoft Research.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376257

hinges on advancing our understanding of how sensitive attributes affect both sides of Human-AI systems.

In this paper we present three studies. First, a replication of Dressel and Farid's study [13] to confirm their experimental protocol and to validate the replicability of their results. Next, we describe Study 1, an extension of Dressel and Farid's work to examine the impact of adding photos to defendant profiles (which serves as a way to increase the prominence of racial cues). Finally, we describe Study 2, which isolated the impact of race from other defendant characteristics by controlling the race of defendant profiles. The key takeaways are:

- Showing race information had a significant impact on users' predictions of whether a defendant would recidivate. When both a photo and a race label were added to a defendant's profile, users were *less* likely to predict that they would recidivate [Study 1].
- Users' self-identified gender was a significant factor in recidivism predictions, with females being impacted when a photo and race label were provided but not males [Study 1 & 2].
- The race of the defendant also impacted female users' recidivism predictions, with black defendants being rated as *less* likely to recidivate compared to white defendants [Study 1 & 2].
- The two variants of our No Race condition produced different effects. In Study 1, when users were only shown text, defendants were rated higher in terms of recidivism than when a photo and race label were provided. In study 2, when an "Image Unavailable" icon was shown in the no race condition, and the trials were intermixed with other trials that included photos and race labels, recidivism predictions were lower.

RELATED WORK

A large body of work in HCI argues for careful considerations of issues of bias in the design of technology [16, 27, 20]. Biases, implicit or explicit, frequently limit opportunities for certain groups [51, 32]. Developing solutions to mitigate bias is currently a major challenge in AI [6, 1, 38, 35, 4]. Implicit gender and racial biases in technology is being studied by both AI and HCI researchers [34, 38, 35, 24, 36], and especially in the context of recidivism following ProPublica's report that COMPAS, a nationwide criminal risk assessment tool, contained racial disparities in its predictions [1, 46, 17, 11, 8, 43, 42].

Risk-Assessments and Judgement on Mechanical Turk

Recent studies on biases in human and algorithmic risk-assessment have utilized Amazon Mechanical Turk [13, 46, 17] as participants. However, such experiments have yielded mixed support for the claim of racial disparities in user judgements. For instance, Dressel and Farid, when comparing human and algorithmic accuracy for recidivism predictions, did not find race to be a factor influencing human judgement [13], while Green and Chen [17] found that users were more likely to predict higher recidivism risk for one racial group (i.e. black defendants), when aided with a risk assessment tool. Algorithmic aids have also been shown to bias workers' predictions of recidivism [19, 45].

Our work informs the design aspects of such systems and studies, in the context of human and algorithm assessments of recidivism.

Impact of Information Presentation on Human Judgement

Prior research in human decision making has shown that the modality in which information is presented can affect user performance and judgement [7, 21, 50, 3, 26, 23]. For instance, images were found to be more effective than text in presenting emergency and disaster information to users [7]; user perceptions of website quality were found to be influenced by subtle manipulations such as framing [21]; and perceived credibility of news websites were found to be affected by purely presentational factors [50].

Similar findings have been discovered about human judgement in the offline world. Newspaper photographs of political candidates can shape voters' perceptions of candidates and alter their likelihood of voting for them [3]. People's perceptions about gender distributions in careers in the real-world can shift by modifying representations of gender in career-related image search results online [26].

Combinations of modalities and cognitive load also influence user judgements. Image + text information can cause information overload and distractions under high cognitive loads [7], but improve recall of information under low cognitive loads [23]. Expanding this line of literature, our work explores how the presentation of racial information influences human judgements of recidivism.

Gender and Race Interaction in Human Judgements

Racial disparities have been studied in domains such as health, education, crime rates and economic status [47, 48]. However, race is also a carrier of historical and cultural biases [28, 10]. Kleider et al. found associations between facial features and people's perception of criminal stereotypes [28, 29]. In addition, biases can have different effects across gender. In Madriz's study, women reported higher levels of fear of crime than men, and their perceptions of criminality and victimhood were affected by their own social position and race conveyed in images [31]. Similarly, Barrett et al. found that women were more likely to be influenced by positive pictures and men by negative ones when assessing political candidates [3]. Our work adds to this literature by exploring the influence of race and gender on human judgements of recidivism.

REPLICATION STUDY

Our goal was to investigate how the presentation of race information impacts human judgement of recidivism. To do so, we extended Dressel and Farid's study [13] by presenting race information in different forms to the user. Before extending Dressel and Farid's study, we first ran a replication study to confirm that our experimental protocol was consistent, and that we were able to replicate their results.

Dressel & Farid Experimental Protocol

Below is a brief summary of Dressel and Farid's experimental protocol. See their paper [13] for further details.

Dressel and Farid selected a subset of 1,000 defendants from the ProPublica COMPAS dataset of 7,214 pretrial defendants from Broward County, Florida [1]. These defendants were grouped into twenty groups of fifty defendants each. Twenty users were assigned to each group, for a total of four hundred users. Each user evaluated all fifty defendants in their assigned group.

Defendant profiles were presented to users in a descriptive paragraph, as shown below, comprising eight features as well as a description of the crime the defendant was charged with. Dressel and Farid ran two versions of this experiment, one that included defendant race information, and another without. Note that the race label is colored blue below for demonstration purposes, but was not blue in the actual presentation.

> The defendant is a [RACE] [SEX] aged [AGE]. They have been charged with: [CRIME CHARGE]. This crime is classified as a [CHARGE DEGREE]. They have been convicted of [NON-JUVENILE PRIOR COUNT] prior crimes. They have [JUVENILE-FELONY COUNT] juvenile felony charges and [JUVENILE-MISDEMEANOR COUNT] juvenile misdemeanor charges on their record.
>
> [CRIME CHARGE] : [CRIME DESCRIPTION]

Users were asked: "Do you think this defendant will commit a crime within 2 years?", to which they could respond "yes" or "no". Users then received two forms of feedback: an indication showing whether their response was correct for that defendant, and a running total showing their accuracy over all defendants evaluated so far.

Three catch trials were used to filter out users who were not paying close attention. The questions were formatted to look like the other questions but had easily identifiable answers. For example, "The state of California was the 31st state to join the Union. California's nickname is: The Golden State ... Does the state of California have a nickname?". If a user answered any of the catch-trials wrong, their data was removed from the dataset.

Replication Study Results

For our replication study, we repeated the no race experimental condition of Dressel and Farid's study, keeping the same set of defendants, the same catch-trials, and the same experimental protocol.

We report results in a slightly different format than Dressel and Farid. The numbers reported in their paper [13] are *defendant-level*, where twenty user judgements for each defendant were aggregated into a score using majority voting aggregation. They then computed performance measures such as accuracy, false positive rates, and false negative rates at the defendant level.

In our work we do not aggregate different user judgements on the same defendant; rather, we report *user-level* performance measures. This is core to our goal of ascertaining if a user's judgement of recidivism depends on how defendant race information is presented, and allows us to study how individual user characteristics and defendant characteristics may interact to impact user judgement.

For Table 1, we recalculated *user-level* performance measures on Dressel and Farid's data[1] using individual users' "yes" or "no" predictions of recidivism, and compare it to the defendant's ground truth recidivism outcome. We also calculated the same measures on our replication study data.

Comparing the results from our replication study to Dressel and Farid's results, Wilcoxon Signed-Ranks Tests revealed no significant differences for any of the measures (ACC: Z = -1.08, p = .281; FPR: Z = -1.03, p = .302; FNR: Z = -.392, p = .695). Given that we successfully replicated Dressel and Farid's study, we move on to extend their work.

Metric	Defendant Race	Dressel & Farid	Replication Study
ACC	Overall	62.3%	61.7%
	Black	62.2%	62.2%
	White	63.2%	61.6%
FPR	Overall	39.8%	40.3%
	Black	46.7%	45.3%
	White	31.1%	36.1%
FNR	Overall	35.4%	36.1%
	Black	31.1%	32.2%
	White	42.7%	42.3%

Table 1. Accuracy (ACC), false positive rates (FPR), and false negative rates (FNR) from our replication study compared to Dressel and Farid's study, for the No Race condition. No significant differences were found between Dressel and Farid's study and our results.

STUDY 1: DO RACE LABELS AND PHOTOS AFFECT JUDGEMENTS OF RECIDIVISM?

In this study, we examined whether the form in which race information is presented to a user influences their judgements of recidivism. In the real world, judges have access to additional information outside of the defendant's file. The judge may see the defendant, hear them speaking, or even ask them questions. Adding photos to defendant profiles may be more realistic than just providing textual data and could influence users' decisions on whether or not a defendant will recidivate.

Study 1 Experimental Protocol

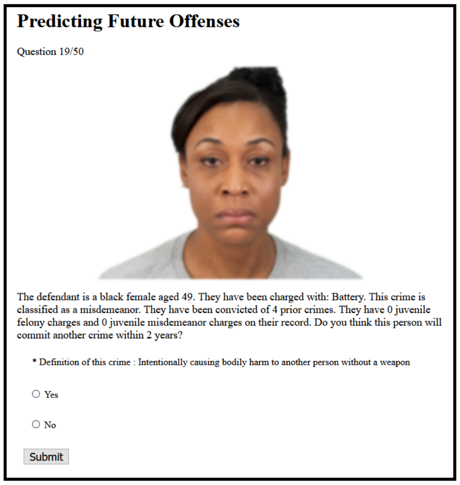
We designed a between-subjects study varying the presentation of race along two dimensions: Race Label (on/off)

[1] Dressel and Farid's data is available at www.cs.dartmouth.edu/farid/downloads/publications/scienceadvances17.

and Photo (on/off). This resulted in four experimental conditions: *No Race; Label Only; Photo Only; Label + Photo*. In their work, Dressel and Farid varied the Race Label (whether or not the defendant's race was listed in the text of their profile); this study extends their work by also varying whether or not a photo is displayed in the defendant's profile. Our study utilizes Dressel and Farid's experimental protocol (as described earlier) with a few minor modifications described below. Our study also underwent ethics, privacy, and IRB review.

Adding Photos

Defendant race labels are recorded in the ProPublica dataset [1], however photos of the actual defendants are not available. We therefore utilized images from a publicly available dataset: the Chicago Face Database (CFD) [30], which has been designed for use in scientific research. We matched each defendant with a photo corresponding to their age, race, and gender. Although the photo was not of the real defendant, the benefit of using these photos was that they were all captured in a consistent manner, and have similar visual characteristics. The CFD database was large enough that we were able to find representative photos for most of the defendants in our dataset. Some photos were reused for different defendants but we ensured that no user saw the same photo twice. We also applied a 1.75px blur to the photos so they looked less professional. Figure 1 shows a screenshot of a defendant profile. Users were not aware that the photos were not of the actual defendants until the end of the study when it was disclosed during a post-study debrief.



Predicting Future Offenses

Question 19/50

The defendant is a black female aged 49. They have been charged with: Battery. This crime is classified as a misdemeanor. They have been convicted of 4 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record. Do you think this person will commit another crime within 2 years?

* Definition of this crime : Intentionally causing bodily harm to another person without a weapon

○ Yes

○ No

Figure 1. Screenshot of the Label + Photo condition in Study 1. Users read the defendant's profile and were asked to predict if the defendant will commit a crime within two years.

Feedback and Payment Modifications

We made two modifications to Dressel and Farid's experimental protocol. First, we removed feedback telling users whether they were correct or not after evaluating each defendant, and we omitted the running accuracy score. These modifications were made because we were concerned that feedback would cause variation in user performance across the session [33, 22]. Additionally, this modification addresses Dressel and Farid's concern that users may lose motivation during the task if they feel they are not going to achieve a monetary bonus (given to users who exceed a certain overall accuracy).

Our second modification was increasing the payment to users to be $1.50 (from $1.00 paid by Dressel and Farid) for completion of the task, with a $5.00 bonus (same as Dressel and Farid) for reaching an overall accuracy of 65% or higher. This change was made to ensure that we were paying ethical wages [33]. On average, the task was expected to take 10 minutes to complete, so workers would earn $9 or $39 per hour[2], depending on whether they received the bonus.

Setup

The data for Study 1 was collected from 1,600 Amazon Mechanical Turk workers, four hundred per experimental condition. We used the same 1,000 defendants as Dressel and Farid, grouped into twenty sets of fifty defendants. Each set of defendants was rated by twenty users, for a total of four hundred users per condition (*No Race, Label Only, Photo Only, Label+Photo*).

Users ranged in age from 18 to over 74, with roughly even numbers of users who identified as females and males (50.7% females; 49.2% males). The majority of users self-identified as White (76.5%), followed by Black (7.3%), Asian (7.0%), and Hispanic (6.2%). Users also reported their education level, with 49.4% holding a Bachelor's degree or higher and 46.25% with education level below Bachelor's degree (with 4.35% choosing not to answer). An additional 236 users completed the task but failed one or more catch trials and were removed from the dataset.

Analyses

Since our primary goal was to explore whether the presentation of race impacted users' judgements regarding recidivism, our main dependent variable was whether the defendants were predicted to recidivate. We also calculated accuracy, false positive rates, and false negative rates. Non-parametric statistics were used to analyse our data given that the data was not continuous and violated assumptions of homogeneity and sphericity. Non-parametric Kruskall-Wallis H one-way ANOVA tests were used to examine differences in recidivism predictions across the experimental conditions (*No Race, Label Only, Photo Only, Label + Photo*) ($= .05$). Significant differences were followed-up using post-hoc pairwise Mann-Whitney U tests with Bonferroni corrections ($_{adjusted} = .008$).[3] Finally, to explore differences in recidivism predictions by group (e.g. user gender, defendant race, or defendant gender) or metric (e.g. accuracy, false positive rates), non-parametric Kruskall-Wallis H one-way ANOVA tests were applied within each group.

[2] If bonus is received, the estimated hourly rate is $39=$(1.50+5)*6.

[3] The adjusted alpha is .008=0.05/6, since there are six pairwise comparisons across four experimental conditions.

Study 1 Results

The percentage of defendants predicted to recidivate for each of the experimental conditions is shown in Table 2. A statistically significant difference was found across the four conditions ($H(3) = 20.93$, $p < 0.001$). Post-hoc pairwise comparisons revealed that users were significantly less likely to predict that defendants would recidivate when both race labels and photos were shown, as compared to the other conditions ($p < 0.008$). None of the other pairwise comparisons were statistically significant.

> **Result 1a:** Adding both photos and race labels to defendant profiles made users **less** likely to predict that a defendant would recidivate within two years.

No Race	Label Only	Photo Only	Label+Photo
54.7%	**55.1%**	**54.8%**	**53.1%***

*p < .008

Table 2. Study 1: Percentage of defendants predicted to recidivate across the four experimental conditions. Significant differences across conditions were found, with fewer defendants predicted to recidivate in the Label+Photo condition than the other three conditions.

User Gender

Prior research has shown that the gender of an Amazon Mechanical Turk worker can have a significant effect on results, particularly when gender or race is involved in the task [24, 36]. We therefore investigated whether users' self-identified gender impacted recidivism predictions for our experimental conditions. Table 3 shows the percentage of defendants predicted to recidivate in each condition, split by the gender of the user evaluating that defendant.

No significant differences across experimental conditions were found when the predictions were done by male users ($H(3) = 3.03$, $p = .388$). However, for female users, predictions made across different experimental conditions were significantly different ($H(3) = 21.78$, $p < 0.001$). Post-hoc pairwise comparisons revealed that female users were less likely to predict that a defendant would recidivate in the Label+Photo experimental condition than in the other three conditions ($p < .008$).

> **Result 1b:** Adding both photos and race labels made female users **less** likely to predict that a defendant would recidivate within two years but had no significant effect on male users' predictions.

Since no significant differences across experimental conditions were found for male users, subsequent analyses for Study 1 will focus on predictions made by female users.

Defendant Race

We explored whether the defendant's race had an impact on recidivism predictions. Given that there were very few Hispanic, Asian and Native American defendants in the dataset, profiles belonging to these race categories were grouped together, resulting in three race categories: White (n=377), Black (n=530) and Other (n=93).

User Gender	No Race	Label Only	Photo Only	Label+Photo
Male	52.4%	53.0%	53.3%	52.2%
Female	**56.6%**	**57.0%**	**56.5%**	**54.0%***

*p < .008

Table 3. Study 1: Percentage of defendants predicted to recidivate across the four experimental conditions, split by user-specified gender. Significant differences across conditions were found for female users, with fewer defendants predicted to recidivate in the Label+Photo condition than the other three conditions.

The percentage of defendants predicted by female users to recidivate in each experimental condition, by race, is shown in Table 4. No significant differences across the experimental conditions were found for white defendants ($H(3) = 2.81, p = .422$) and defendants of other races ($H(3) = 3.67, p = .300$), but a significant difference was found for black defendants ($H(3) = 23.21, p < 0.001$). Post-hoc pairwise comparisons revealed that female users were less likely to predict that a black defendant would recidivate in the Label + Photo condition than in the other three conditions ($p < .008$).

> **Result 1c:** Adding both photos and race labels only impacted black defendants, and not white defendants or defendants of other races. The photos and race labels made female users less likely to predict that a black defendant would recidivate within two years.

Defendant Race	No Race	Label Only	Photo Only	Label+Photo
White	49.3%	50.6%	49.6%	48.8%
Black	**63.3%**	**62.9%**	**62.7%**	**59.2%***
Other	48.0%	49.2%	48.5%	45.0%

*p < .008

Table 4. Study 1: Percentage of defendants predicted by female users to recidivate across the four experimental conditions, split by defendant race. The "Other" race category included Hispanic, Asian, and Native American defendants. Significant differences across conditions were found for black defendants, with fewer defendants predicted to recidivate in the Label+Photo condition than the other three conditions.

Defendant Gender

We explored whether the gender of the defendant (as specified in the ProPublica dataset [1]) impacted recidivism predictions made by users who identified as female, across experimental conditions. Note however that there were fewer female defendants (n=197) in our dataset than male defendants (n=803), so our statistical power for detecting differences for female defendants was lower. The percentage of defendants predicted by female users to recidivate in each experimental condition, split by defendant gender, are shown in Table 5. A significant difference in experimental condition was found for male defendants ($H(3) = 21.62, p < .001$), but not female defendants ($H(3) = 2.32, p = .509$). Post-hoc pairwise comparisons revealed that female users were less likely to predict that a male defendant would recidivate in the Label + Photo condition than in the other conditions ($p < .008$).

Result 1d: Adding both photos and race labels impacted male defendants, making female users less likely to predict that a male defendant would recidivate within two years.

Defendant Gender	No Race	Label Only	Photo Only	Label+Photo
Male	**59.7%**	**59.7%**	**59.4%**	**56.6%***
Female	43.8%	45.8%	44.7%	43.7%

*$p < .008$

Table 5. Study 1: Percentage of defendants predicted by female users to recidivate across the four experimental conditions, split by defendant gender. Significant differences across conditions were found for male defendants, with again fewer defendants predicted to recidivate in the Label+Photo condition than the other three conditions.

Accuracy, False Positives and False Negatives

Finally, we examined whether or not the different experimental conditions impacted accuracy, false positive or false negative rates for all users (see Table 6). There was no significant difference in accuracy across the conditions ($H(3) = 3.218, p = .359$), however there were significant differences in the false positive rates ($H(3) = 14.27, p = .003$) and false negative rates ($H(3) = 11.01, p = .012$) with the Label + Photo condition having a lower false positive rate, and a higher false negative rate. As expected, further analyses revealed that these differences were only observed for black defendants, and not for white defendants or defendants of other races.

Result 1e: Adding both photos and race labels had no impact on accuracy, but lowered false positive rates and increased false negative rates for black defendants.

Study 1 Discussion

Showing the race of a defendant using both a text label and a photo significantly impacted female users' assessments of black defendants, but had no impact on white or other race defendants. No significant differences in assessments of defendants were found for users who identified as male. Given that previous literature has noted bias against black defendants for recidivism [1, 40, 8], we expected that racial cues would have *increased* predictions of recidivism, however, our results were the opposite. Users were *less likely* to predict that a black defendant would recidivate within two years when a race label and photo were included in the defendant description.

Interestingly, neither the race label nor the photo on their own impacted the recidivism predictions. For the race label, we hypothesize that the saliency of this features is low. It is a small word in a block of text; it could be easily be overlooked by the user. While the photo is a more salient signal, it didn't make a difference on its own. We hypothesize that it can be difficult to infer race from a photo alone. As such, it was the combination of both the race label and the photo that had a significant effect on recidivism predictions.

Metric	Defendant Race	No Race	Label Only	Photo Only	Label+Photo
ACC	Overall	62.5%	62.6%	61.9%	62.6%
	Black	63.1%	63.9%	63.3%	63.7%
	White	62.4%	61.6%	60.5%	61.6%
FPR	Overall	42.5%	42.9%	43.2%	40.9%
	Black	**48.6%**	**47.7%**	**47.3%**	**44.7%***
	White	37.8%	38.7%	40.0%	38.2%
FNR	Overall	32.0%	31.4%	32.5%	33.5%
	Black	**28.3%**	**27.5%**	**28.7%**	**30.0%***
	White	37.4%	37.8%	38.8%	38.8%

*$p < .008$

Table 6. Study 1: Accuracy (ACC), false positive rates (FPR), and false negative rates (FNR) across the four experimental conditions, split by defendant race. Significant differences were found for black defendants with lower false positive rates and higher false negative rates in the Label+Photo condition than the other three conditions.

Our results revealed significant differences for male defendants but not female defendants, however, the number of female defendants in our dataset was low, and we did not systematically assign user gender to defendants. Our next study further explores the issue of defendant gender.

Examining the recidivism predictions in Table 4, we see that black defendants had an overall higher prediction of recidivism compared to the other races. This may be related to the defendant profiles, since ground truth recidivism was also higher among black defendants in the dataset compared to defendants of other races in the dataset – 57% of black defendants were reported to have recidivated compared to 37% of white defendants and 37% of defendants of other races. Hence, it is possible that the differences we observed after adding the race label and photos to defendant profiles were related to other defendant features rather than race (i.e. certain defendant profiles have higher ground truth recidivism rates). Hence, we designed our next study to isolate the effect of defendant race.

STUDY 2: ISOLATING EFFECT OF DEFENDANT RACE

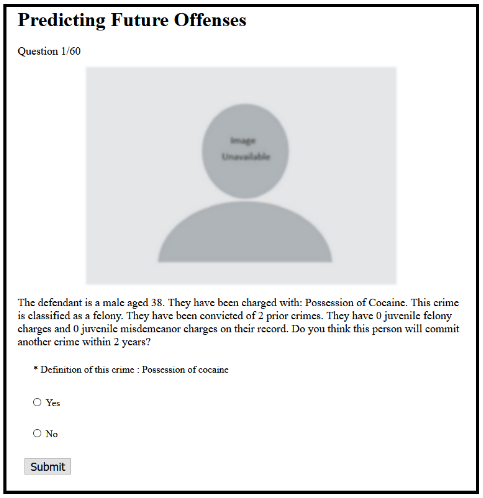
In this study we modified the experimental design from Study 1 to isolate the effect of defendant race and continue our investigation into whether additional racial information influences users' judgements of recidivism. In particular, we wanted to understand if adding race labels and photos has a differential impact on black versus white defendants, *independent of non-race defendant features*. Additionally, we increased the female defendants in our dataset to better understand if the impact of adding race labels and photos is different for male and female defendants.

Study 2 Experimental Protocol

We designed a within-subjects study where we controlled the race for each defendant profile, holding all other features constant. Each defendant profile was evaluated multiple times, once as a black defendant, once as a white defendant, and once with no race information provided. We then examined differences in the recidivism predictions across these three experimental conditions (*No Race, White Label+Photo, Black Label+Photo*). This important modification helped ensure that any differences observed were a result of the racial representation (the label and photo) and not related to characteristics of the defendant profile itself. We provide more details on our experimental design below.

Assignment of Defendant Race

Each experimental session involved sixty defendant profiles – twenty unique defendant profiles (ten male and ten female), presented in three trial clusters. In the first trial cluster, each defendant was randomly assigned to be either a black defendant, a white defendant, or have no race information. In the second trial cluster, one of the remaining representations was randomly chosen for each profile, and in the third cluster the final representation was chosen. The result of this within-subjects design was a presentation of sixty defendant profiles to each user, where the order was mixed in terms of the race of the defendant (Black, White, or No Race). For the no race instances, the race label was removed from the profile text and a blank photo icon was shown which stated "Image Unavailable" instead of a photo (see Figure 2).



Predicting Future Offenses

Question 1/60

The defendant is a male aged 38. They have been charged with: Possession of Cocaine. This crime is classified as a felony. They have been convicted of 2 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record. Do you think this person will commit another crime within 2 years?

* Definition of this crime : Possession of cocaine

○ Yes

○ No

Figure 2. Screenshot of the No Race condition in Study 2. Users read the defendant's profile and were asked to predict if the defendant will commit a crime within two years.

There was no time gap or delineating features between the different trial clusters, to give the appearance of sixty unique defendants. Given the similarities across profiles, and the length of the task, many users likely did not realize that the same profile was being presented multiple times. If however some users did recognize the profiles, this could reduce the likelihood of detecting differences between the conditions.

Balancing Defendant Gender

As pointed out before, Dressel and Farid's sample of 1,000 defendants has fewer female defendants (n=197) than male defendants (n=803), which reduces the statistical power for detecting differences for the smaller group of female defendants. To ensure that we had equal numbers of male and female defendants in this study, we selected additional defendants from the original ProPublica COMPAS dataset [1], resulting in eight hundred male and eight hundred female defendant profiles, split evenly across black and white defendants.

Setup

The data for Study 2 was collected from 1,600 Amazon Mechanical Turk workers (eight hundred who self-identified as female and eight hundred who self-identified as male), with equal numbers of females and males evaluating each defendant.

Similar to Study 1, users ranged in age from 18 to over 74. The majority of users self-identified as White (71.6%), followed by Black (10.2%), Asian (8.7%), and Hispanic (5.4%). Users were also asked to report their education level, with 54.4% holding a Bachelor's degree or higher and 45.0% with education level below Bachelor's Degree (and 0.6% choosing not to answer). An additional 635 users completed the task but failed at least one of the catch trials and were removed from the dataset.

Analyses

Similar to Study 1, our primary dependant measure was whether or not defendants were predicted to recidivate. We also analyzed accuracy, false positive rates, and false negative rates. Non-parametric statistics were used to analyse our data given that the it was not continuous and violated assumptions of homogeneity and sphericity. To account for the within-subjects design, non-parametric Friedman Tests were used to examine differences in recidivism predictions across the experimental conditions (*No Race, White Label+Photo, Black Label+Photo*) ($= .05$). Significant differences were followed-up using posthoc pairwise Wilcoxon Signed Rank Tests with Bonferroni corrections ($_{adjusted} = .017$).[4] Finally, to study if the difference in recidivism predictions across the three experimental conditions differed by group (e.g. user gender, defendant race, or defendant gender) or metric (e.g. accuracy, false positive rates), non-parametric Friedman tests were also applied, but within each group.

Study 2 Results

The percentage of defendants predicted to recidivate for each condition is shown in Table 7. Our results revealed a significant differences across experimental conditions ($\chi^2(2) = 43.95, p < 0.001$). All post-hoc pairwise comparisons were significant ($p < .017$). Similar to Study 1, when

[4] The adjusted alpha is .0017=0.05/3, since there are three pairwise comparisons across three experimental conditions.

a race label and a photo were shown to users, they were significantly less likely to predict that a black defendant would recidivate compared to a white defendant, even though the profiles were identical. In contrast to the results from Study 1, users were significantly less likely to predict that a defendant would recidivate when no race information was provided ($p < .017$).

Result 2a: Even when users rated the same defendant profile, they were **less** likely to predict that a defendant would recidivate when the defendant's race was presented as black instead of white.

No Race	White Label+Photo	Black Label+Photo
53.4%*	**55.0%***	**54.0%***

$^*p < .017$

Table 7. Study 2: Percentage of defendants predicted to recidivate across the three experimental conditions. All pairwise comparisons were significant with fewer defendants predicted to recidivate in the Black Label+Photo condition than the White Label+Photo condition, and even fewer defendants predicted to recidivate in the No Race condition.

User Gender

We again examined whether the self-identified gender of the user impacted recidivism predictions. The percentage of defendants predicted to recidivate in each condition, by user gender, are shown in Table 8. Similar to Study 1, there were no significant difference across conditions for male users ($\chi^2(2) = 4.81, p = 0.090$). However, for female users, the percentage of defendants predicted to recidivate differed significantly across conditions ($\chi^2(2) = 53.51, p < 0.001$). All post-hoc pairwise comparisons were significant ($p < .017$) with the No Race condition having the fewest predictions of recidivism. Similar to Study 1, the presented race of the defendant mattered to female users and they were less likely to predict that a defendant would recidivate when they were told it was a black defendant, compared to a white defendant.

Result 2b: Presenting a defendant's race as black instead of white made female users less likely to predict that the defendant would recidivate, but had no significant effect on male users' predictions.

User Gender	No Race	White Label+Photo	Black Label+Photo
Male	52.2%	52.9%	52.4%
Female	**54.8%***	**57.0%***	**55.7%***

$^*p < .017$

Table 8. Study 2: Percentage of defendants predicted to recidivate across the three experimental conditions, split by user-specified gender. Significant differences across conditions were found for female users, with fewer defendants predicted to recidivate in the Black Label+Photo condition than the White Label+Photo condition. All pairwise comparisons were significant.

Defendant Gender

We next explored whether the gender of the defendant (as specified in the ProPublica dataset [1]) impacted recidivism predictions. The percentage of female and male defendants predicted to recidivate by female users are shown in Table 9. Significant differences across conditions were found for both male and female defendants ($\chi^2(2) = 27.81, p < 0.001$, and ($\chi^2(2) = 30.77, p < 0.001$, respectively). Post-hoc pairwise comparisons revealed that for male defendants, there was no difference when they were portrayed as being white or black, but they were less likely to be predicted to recidivate by female users when no race information was provided. For female defendants, they were less likely to be predicted to recidivate by female users if they were portrayed as being black, or with no race information ($p < .017$).

Result 2c: Adding photos and race labels impacted both male and female defendants. For male defendants, the additional race information made female users more likely to predict that they would recidivate, irrespective of the portrayed race. For female defendants, they were more likely to be predicted to recidivate when they were portrayed as being white, compared to black, or when no race information was provided.

Defendant Gender	No Race	White Label+Photo	Black Label+Photo
Male	**60.8%***	**63.1%**	**62.4%**
Female	**48.7%**	**51.0%***	**49.0%**

$^*p < .017$

Table 9. Study 2: Percentage of defendants predicted by female users to recidivate across the three experimental conditions, split by defendant gender. Significant differences across conditions were found regardless of defendant reported gender; statistically significant pairwise comparisons are described above.

Accuracy, False Positives and False Negatives

We again examined whether or not the experimental conditions impacted accuracy, false positive and false negative rates for all users (see Table 10). Significant differences were found for all metrics (ACC: $\chi^2(2) = 10.16, p = 0.006$; FPR: $\chi^2(2) = 53.51, p < 0.001$; and FNR: $\chi^2(2) = 7.50, p = 0.023$). Post-hoc pairwise analyses revealed that when photos and race labels were provided, users had higher accuracy, lower false positive rates, and higher false negative rates for black defendants, compared to white defendants. Results for the no race condition were similar to the black defendants except the false positive rate was significantly lower.

Result 2d: Presenting a defendant's race as black instead of white increased accuracy, decreased false positive rates, and increased false negative rates.

Metric	No Race	White Label+Photo	Black Label+Photo
ACC	**60.8%**	**60.1%***	**60.4%**
FPR	**43.3%***	**45.3%***	**44.2%***
FNR	**34.2%**	**33.3%***	**34.0%**

*$p < .017$

Table 10. Study 2: Accuracy, false positive rates, and false negative rates across the three experimental conditions. Significant differences across conditions were found for all metrics. All pairwise comparisons were significant for the false positive rate metric. For accuracy and false negative rates, the White Label+Photo condition was significantly lower than the other conditions.

Study 2 Discussion

Similar to Study 1, when race labels and photos were added to defendant profiles, users were less likely to predict that black defendants would recidivate compared white defendants. This provides evidence that the observed differences are attributable to the presented race of the defendant, and not to other features inherent in the profiles. Additionally, Study 2 replicated the result that users who identified as female were sensitive to the race of the defendant, but not those who identified as male. The added female defendants in our dataset showed that these differences are applicable to both male and female defendants (not just males as suggested in Study 1).

One difference in the results between Study 1 and Study 2 was related to the No Race condition. In Study 1, users were **more** likely to predict that defendants in the No Race condition would recidivate, compared to the Label+Photo condition; however, in Study 2, users were **less** likely to predict that defendants in the No Race condition would recidivate, compared to the other conditions. There were two differences which may have contributed to this change. First, users in Study 1 were only exposed to one condition, while users in Study 2 were exposed to all three conditions, and the trials were intermixed. Second, the visual layout of the tasks were different in the two studies. In Study 1, the No Race condition only showed a block of text describing the defendant's profile. In Study 2 a placeholder "Image Unavailable" photo was added (see Figure 2). We hypothesize that if photos have a humanizing effect on the task (as suggested in Study 1), having trials with photos adjacent to the No Race trials may cause a "spillover" humanizing affect, making users more conscious about the people they were evaluating in the task. Alternatively, the "Image Unavailable" icon may have altered users perceptions of those defendants.

In Study 1 users were presented with fifty unique defendants, but in Study 2 users were presented with twenty unique defendants that they evaluated three times. We examined the level of agreement users had across the conditions. While a reasonable level of agreement is expected if users were being consistent in their predictions, we would expect some variability given the complexity of the task. Overall, only 0.3% of users (n=50) had perfect agreement across their predictions, and on average, the level of agreement was 75%. If users did notice the repetition of defendants, they would be more likely to repeat their predictions, making it more difficult to detect differences between conditions. Given that we still observed significant differences provides support for these results.

RESULTS SUMMARY

Although we hypothesized that adding defendant photos would impact users' judgements, we were surprised in the ways it manifested in our results.

Adding Photos Mattered

In contrast to Dressel and Farid's findings [13], showing race information had a significant impact on users' recidivism predictions but only when both a photo and race label were added to defendants' profiles. It is interesting that neither the label nor the photo on their own made a difference, but the combination of the two had a significant impact on users' judgement.

We initially hypothesized that adding photos might induce racial bias and therefore would **increase** recidivism predictions for certain classes of defendants. However, in Study 1, adding labels and photos **reduced** recidivism predictions. This suggests that photos may have a humanizing effect on the task, making users less likely to predict that a defendant will recidivate.

No Race Conditions

It is unclear what caused the No Race condition to be rated higher (than the race conditions) in one study, and lower (than the race conditions) in another study. As previously explained, there were small differences in this condition between the two studies which may have impacted the results. Regardless, whether it was the visual presentation of the task (i.e. the "Image Unavailable" icon), or the fact that the race trials were intermixed, it suggests that human judgements are susceptible to even small design choices. Further research is needed to better understand the difference in these results.

Female vs Male Workers on Amazon Mechanical Turk

One of the strongest results from this work are the striking differences between users who identified as female versus male. In both studies, significant differences across experimental conditions were observed for female users but not males. This is consistent with other recent studies [36, 24], where female and male Mechanical Turk Workers reacted differently to issues related to race and gender. This is an important takeaway for researchers who use Amazon Mechanical Turk.

Differential Impact of Race

The results for black defendants were particularly interesting, especially given that there is well documented bias against black people in the criminal justice field [1, 40, 41, 39, 37], and other research on recidivism that suggests that users rate black defendants more harshly [28, 29, 10, 17]. The opposite was true in our studies, as the addition of race labels and photos reduced the likelihood that a black defendant would be predicted to recidivate. We hypothesized that users may have overcompensated to ensure they were not exhibiting a common societal bias.

Practical Significance

Many of the significant results uncovered in these studies only represent a small difference in the actual population percentages. Although this suggests that the results have low practical significance, when it comes to issue of racial and gender bias, making things better for even a small number of people is important. As such, we value any improvement we can make, even if the scope is small.

Study Limitations

Recidivism Predictions as a Mechanical Turk Task

We are not proposing that recidivism decisions should be done by Amazon Mechanical Turk workers. Recidivism is a complex criminal justice process that requires expert knowledge. However, in the context of this study, having Mechanical Turk workers assess the likelihood of someone committing a crime in the future is a reasonable lay-task. Juries are made up of everyday people who are asked to assess guilt or innocence. Also, as people go about their daily lives they often need to assess whether the people around them are law abiding citizens, or potential criminals. Previous work by Dressel and Farid provides an experimental framework to explore issues around bias that allows for different controls and extensions.

Defendant Photos

Our use of photos from the Chicago Face Database is both a benefit, and a limitation. Being able to use photos of real defendants would give us more accurate results; however, those photos could also be heavily biased as they would have been taken under widely varying conditions. Additionally, it is unethical to use photos from people who have not consented to their use. Getting access and consent to use photos from a large corpus of defendants would be difficult. In contrast, the CFD photos were collected with the intent to be used in scientific research.

User Race Interactions

Our results show interaction between race and gender, both for the defendants and for the users evaluating the defendants. There is a strong likelihood that users' race would also impact their judgments, and probably interact with the race and gender of the defendants. While we were able to collect a balanced sample of users who identified as female and male, approximately 75% of our users self-identified as white. Collecting a racially diverse set of users is challenging but should be pursued in future work.

Ground Truth

Although our results primarily focused on perceived criminality across conditions, we also reported on accuracy, false positive, and false negative rates. It is important to understand that these measures all compare human judgement to the recorded ground truth for the defendant; however, ground truth for recidivism is extremely noisy and heavily biased [41]. This raises concerns about the validity of ground truth recidivism data, and therefore measures of accuracy, false positive and false negative rates should be interpreted with caution.

Binary Gender

This study only reports results with male and female defendant profiles and Amazon Mechanical Turk workers who self-identified as male or female. While we recognize that there are many ways people self-identify their gender, the majority of our participants identified as either male or female, and the dataset used for defendant profiles only reported binary genders. We acknowledge the importance of studies that consider gender as a non-binary variable and hope that future studies will take this into consideration.

CONCLUSIONS

In conclusion, this work demonstrates that human judgement is impacted by how race information is presented, and sometimes in ways that contradict our expectations. By adding both race labels and photos to defendant profiles, users were made more aware of the defendant's race. However, instead of exacerbating racial biases, it made users **less** likely to view someone as a criminal.

This work also demonstrated an interaction between gender and race, as it was the female users in our study who were influenced by the additional race information. This resulted in the female users rating black defendants as less likely to recidivate compared to white defendants. Further research is needed to better understand why defendant race did not impact judgements of male users.

The results of our work suggest that system-designers need to develop ways to account for and mitigate human biases in development of future online systems. Since there was differential impact found based on interactions between the system users and the individuals they are asked to evaluate, a potential solution could involve personalizing tasks to different types of users to account for these interactions. Moreover, biases not only manifest in humans but also in algorithms, so system-designers should also consider Human + AI systems that complement each other and maximize their potential.

Overall, this paper extends the conversation started by Dressel and Farid and opens the door for more research investigating how representations influence human judgement. More understanding in this space is critical, especially given the growth of AI systems that need to work *with* users to help inform societal decisions.

ACKNOWLEDGEMENTS

We would like to thank Justin Cranshaw, Scott Counts, Gabrielle Quinn, and Alexa Brewer for their assistance with this project. We also want to thank the reviewers for their valuable feedback.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. How we analyzed the compas recidivism algorithm. (2016).

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm Accessed May 26, 2017.

[2] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *SIGCHI Conference on Human Factors in Computing Systems*. 965–974.

[3] Andrew W Barrett and Lowell W Barrington. 2005. Is a picture worth a thousand words? Newspaper photographs and voter evaluations of political candidates. *Harvard International Journal of Press/Politics* 10, 4 (2005), 98–113.

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

[5] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human Computer Interaction* 3, CSCW (2019), 1–24.

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[7] Yujia Cao, Mariët Theune, and Anton Nijholt. 2009. Modality effects on cognitive load and performance in high-load information presentation. In *International Conference on Intelligent User Interfaces*. 335–344.

[8] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[9] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.

[10] Travis L Dixon and Keith B Maddox. 2005. Skin tone, crime news, and social reality judgments: Priming the stereotype of the dark and dangerous black criminal. *Journal of Applied Social Psychology* 35, 8 (2005), 1555–1570.

[11] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel K.E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *International Conference on Intelligent User Interfaces*. 275–285.

[12] Judith Donath. 2007. Signals in social supernets. *Journal of Computer-Mediated Communication* 13, 1 (2007), 231–251.

[13] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580.

[14] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.

[15] Asle Fagerstrøm, Sanchit Pawar, Valdimar Sigurdsson, Gordon R Foxall, and Mirella Yani-de Soriano. 2017. That personal profile image might jeopardize your rental opportunity! On the relative impact of the seller's facial expressions upon buying behavior on Airbnb™. *Computers in Human Behavior* 72 (2017), 123–131.

[16] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *Transactions on Information Systems* 14, 3 (1996), 330–347.

[17] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Conference on Fairness, Accountability, and Transparency*. 90–99.

[18] Anthony G Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California Law Review* 94, 4 (2006), 945–967.

[19] Nina Grgic-Hlaca, Christoph Engel, and Krishna P. Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human Computer Interaction* 3, CSCW (2019).

[20] David Hankerson, Andrea R Marshall, Jennifer Booker, Houda El Mimouni, Imani Walker, and Jennifer A Rode. 2016. Does technology have race?. In *CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 473–486.

[21] Jan Hartmann, Antonella De Angeli, and Alistair Sutcliffe. 2008. Framing the user experience: Information biases on website quality judgement. In *SIGCHI Conference on Human Factors in Computing Systems*. 855–864.

[22] Alexander M. Holsinger, Christopher T. Lowenkamp, Edward J. Latessa, Ralph Serin, Thomas H. Cohen, Charles R. Robinson, Anthony W. Flores, and Scott W. VanBenschoten. 2018. A rejoinder to Dressel and Farid: New study finds computer algorithm is more accurate than humans at predicting arrest and as good as a group of 20 lay experts. *Federal Probation* 82, 2 (2018), 50–55.

[23] Weiyin Hong, James YL Thong, and Kar Yan Tam. 2004. Designing product listing pages on e-commerce websites: An examination of presentation mode and information format. *International Journal of Human-Computer Studies* 61, 4 (2004), 481–503.

[24] Farnaz Jahanbakhsh, Justin Cranshaw, Kori Inkpen, Scott Counts, and Walter Lasecki. 2020. Bias in social rating systems: The role of performance quality and gender. In *CHI Conference on Human Factors in Computing Systems*.

[25] Jerry Kang, Mark Bennett, Devon Carbado, and Pam Casey. 2011. Implicit bias in the courtroom. *UCLA Law Review* 59 (2011), 1124.

[26] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *CHI Conference on Human Factors in Computing Systems*. 3819–3828.

[27] William Kilbourne and Susan Weeks. 1997. A socio-economic perspective on gender bias in technology. *The Journal of Socio-Economics* 26, 3 (1997), 243–260.

[28] Heather M Kleider, Sarah E Cavrak, and Leslie R Knuycky. 2012. Looking like a criminal: Stereotypical black facial features promote face source memory error. *Memory & Cognition* 40, 8 (2012), 1200–1213.

[29] Heather M Kleider-Offutt, Alesha D Bond, and Shanna E.A. Hegerty. 2017. Black stereotypical features: When a face type can get you in trouble. *Current Directions in Psychological Science* 26, 1 (2017), 28–33.

[30] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47, 4 (2015), 1122–1135.

[31] Esther I Madriz. 1997. Images of criminals and victims: A study on women's fear and social control. *Gender & Society* 11, 3 (1997), 342–356.

[32] Jennifer Martin, Elspeth McKay, and Janki Shankar. 2006. Bias, misinformation and disinformation: Mental health, employment and human computer interaction. In *Informing Science and IT Education Joint Conference*.

[33] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (2012), 1–23.

[34] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

[35] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[36] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? The impact of representation criteria on human bias in hiring. In *AAAI Conference on Human Computation and Crowdsourcing*. 125–134.

[37] Jeffrey J Rachlinski, Sheri Lynn Johnson, Andrew J Wistrich, and Chris Guthrie. 2008. Does unconscious racial bias affect trial judges. *Notre Dame Law Review* 84 (2008), 1195.

[38] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *SIGCHI Conference on Human Factors in Computing Systems*. 315.

[39] Craig S Schwalbe, Mark W Fraser, Steven H Day, and Valerie Cooley. 2006. Classifying juvenile offenders according to risk of recidivism: Predictive validity, race/ethnicity, and gender. *Criminal Justice and Behavior* 33, 3 (2006), 305–324.

[40] Jennifer L Skeem and Christopher T Lowenkamp. 2016. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology* 54, 4 (2016), 680–712.

[41] Douglas A Smith, Christy A Visher, and Laura A Davidson. 1984. Equity and discretionary justice: The influence of race on police arrest decisions. *Journal of Criminal Law and Criminology* 75 (1984), 234.

[42] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018a. Investigating Human+Machine Complementarity for Recidivism Predictions. *arXiv preprint arXiv:1808.09123* (2018).

[43] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018b. Distill-and-compare: auditing black-box models using transparent model distillation. In *AAAI/ACM Conference on AI, Ethics, and Society*. 303–310.

[44] Bethany A Teachman and Kelly D Brownell. 2001. Implicit anti-fat bias among health professionals: is anyone immune? *International Journal of Obesity* 25, 10 (2001), 1525.

[45] Michelle Vaccaro and Jim Waldo. 2019. The effects of mixing machine learning and human judgment. *Commun. ACM* 62, 11 (2019), 104–110.

[46] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human Computer Interaction* 3, CSCW (2019).

[47] Michael M Wehrman. 2010. Race, concentrated disadvantage, and recidivism: A test of interaction effects. *Journal of Criminal Justice* 38, 4 (2010), 538–544.

[48] David R Williams. 1997. Race and health: basic questions, emerging directions. *Annals of epidemiology* 7, 5 (1997), 322–333.

[49] John Paul Wilson, Kurt Hugenberg, and Nicholas O Rule. 2017. Racial bias in judgments of physical size and formidability: From size to threat. *Journal of personality and social psychology* 113, 1 (2017), 59.

[50] Jacob O Wobbrock, Anya K Hsu, Marijn A Burger, and Michael J Magee. 2019. Isolating the effects of web page visual appearance on the perceived credibility of online news among college students. In *Conference on Hypertext and Social Media*. 191–200.

[51] Shannon N Zenk, Amy J Schulz, Barbara A Israel, Sherman A James, Shuming Bao, and Mark L Wilson. 2005. Neighborhood racial composition, neighborhood poverty, and the spatial accessibility of supermarkets in metropolitan Detroit. *American journal of public health* 95, 4 (2005), 660–667.