The Influences of Task Design on Crowdsourced Judgement: A Case Study of Recidivism Risk Evaluation

Xiaoni Duan
duan79@purdue.edu
Purdue University
USA

Chien-Ju Ho
chienju.ho@wustl.edu
Washington University in St. Louis
USA

Ming Yin
mingyin@purdue.edu
Purdue University
USA

ABSTRACT

Crowdsourcing is widely used to solicit judgement from people in diverse applications ranging from evaluating information quality to rating gig worker performance. To encourage the crowd to put in genuine effort in the judgement tasks, various ways to structure and organize these tasks have been explored, though the understandings of how these task design choices influence the crowd's judgement are still largely lacking. In this paper, using recidivism risk evaluation as an example, we conduct a randomized experiment to examine the effects of two common designs of crowdsourcing judgement tasks—encouraging the crowd to deliberate and providing feedback to the crowd—on the quality, strictness, and fairness of the crowd's recidivism risk judgements. Our results show that different designs of the judgement tasks significantly affect the strictness of the crowd's judgements. Moreover, task designs also have the potential to significantly influence how fairly the crowd judges defendants from different racial groups, on those cases where the crowd exhibits substantial in-group bias. Finally, we find that the impacts of task designs on the judgement also vary with the crowd workers' own characteristics, such as their cognitive reflection levels. Together, these results highlight the importance of obtaining a nuanced understanding on the relationship between task designs and properties of the crowdsourced judgements.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

KEYWORDS

Crowdsourcing, task design, quality, bias, fairness

ACM Reference Format:
Xiaoni Duan, Chien-Ju Ho, and Ming Yin. 2022. The Influences of Task Design on Crowdsourced Judgement: A Case Study of Recidivism Risk Evaluation . In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3485447.3512239

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512239

1 INTRODUCTION

In recent years, crowdsourcing has become a prevalent paradigm to collect judgements from the public—with many of them reflecting people's subjective opinions—to tap into the wisdom of the crowd. For example, the crowd is asked to rate the credibility of news content [2], to assess the performance of their peers in massive open online courses [15], and to make judgements about other *people* such as evaluating trustworthiness of freelancers [30]. To ensure that crowd workers make the effort to provide useful information in the crowdsourced judgement tasks, various attempts have been made with respect to the designs of these tasks, including varying the task interface and instructions [1, 24], changing the dimensionality and granularity of the judgement scale [44, 56], prompting workers to engage in thorough deliberation [60], and providing workers with arguments and feedback that are generated by their peers or algorithmic tools [18, 27, 59].

Intuitively, the designs of crowdsourced judgement tasks may affect the ways that crowd workers make their judgements in these tasks. While many existing studies have explored how the choices on task designs change the *quality* of crowdsourced judgement, more recently, an increasing amount of attention has been paid to the *bias* of the data obtained from crowdsourcing attempts. Yet, our knowledge of whether and how various task designs affect biases in the crowd's judgements is largely limited. Taking the crowdsourced judgement about *people* as an example, the "biases" in the judgements can be examined through at least two angles—the overall tendency for the crowd workers to favor one judgement over another (i.e., how "strict" the crowd is if one judgement is more favorable), and the extent to which the crowd workers judge people from different groups equally (i.e., how "fair" the crowd is). As the crowdsourced judgements about other people may both bring up real-world impact to those people (e.g., increase or decrease of job opportunities for freelancers) and largely impact the performance of downstream hybrid systems that utilize these judgements as inputs (e.g., an AI-driven freelancer recommending system), the need of deepening our understandings of how task designs influence the strictness and fairness of the crowd's judgements is pressing.

Therefore, in this paper, using recidivism risk assessment as a case study, we conduct an experimental study aiming to obtain a more comprehensive understanding of how the designs of crowdsourced judgement tasks affect not just the quality, but also the strictness and fairness of the judgements. We focus on two large categories of task designs—adding interventions to encourage deep deliberation from crowd workers, and providing different types of feedback to crowd workers in the task. Further, we are interested in understanding whether and how the impacts of task designs on the properties of the crowd's recidivism risk judgements, especially

with respect to the strictness and fairness of the judgements, are moderated by a variety of factors. Specifically, we ask:

- **RQ1**: How do task designs affect the recidivism judgements when the information that crowd workers receive in the tasks does/does not reinforce societal stereotypes in their mind (e.g., implicitly associate racial categories with recidivism risks)?
- **RQ2**: How do task designs affect the recidivism judgements on those cases where crowd workers are particularly vulnerable to their own biases (e.g., in-group bias)?
- **RQ3**: Do the influence of task designs on the recidivism judgements vary with crowd workers' own characteristics, such as their cognitive reflection levels?

To answer these questions, we conduct our experiments on two datasets that are either balanced or unbalanced with respect to the defendant's race and their true reoffending status to simulate two different judgement environments in which crowd workers' racial stereotypes in mind get or do not get reinforced. We also conduct disaggregated analysis of the experimental data, both for tasks that trigger human biases to different extent, and for crowd workers with different cognitive reflection levels. Our results show that while there is little evidence suggesting that the task designs that we have examined in this study significantly affect the quality of the crowd's recidivism judgements, they exhibit a strong impact on the *strictness* of the crowd's judgements, regardless of whether the judgement environment has reinforced crowd workers' racial stereotypes or not. Furthermore, on those tasks where crowd workers exhibit a high degree of in-group bias, task designs are shown to significantly affect how *fairly* crowd workers treat defendants of different racial groups—in particular, providing crowd workers with the feedback from a machine learning model that satisfies certain fairness constraints nudges them into making fairer judgements. Lastly, we also find that the impact of task designs on the crowd's recidivism judgements is generally stronger among crowd workers who have high cognitive reflection levels and thus tend to engage in slow and deliberative thinking. Together, these results highlight the importance of obtaining a nuanced understanding of how task designs influence crowdsourced judgements.

2 RELATED WORK

Crowdsourcing has been widely adopted to solicit judgements from humans, and various research efforts have been devoted to improve the quality of the judgements obtained from the crowd. The common approaches to improve the quality of crowdsourced judgements include post-hoc aggregating multiple noisy judgements into high-quality ones [12, 14, 16, 33, 34, 47, 52, 63, 68] and designing proper incentives [35–38, 50, 65]. Meanwhile, researchers also explore different ways to solicit high-quality judgements through improving the task design, such as changing the task interface [1, 24, 44, 56], engaging users in deliberation [54, 60], and providing feedback to users during decision making [18, 20, 27, 59].

Most recently, with the recognition of the important roles that crowdsourced data can play in the larger human-machine collaborative ecosystem [9, 32, 53, 67], a growing amount of attention has been paid on understanding *biases* in the crowdsourced data. For example, a few studies have examined the types of biases that may exist in crowdsourced datasets [22, 44, 64]. Different methods have been proposed to mitigate biases through raising people's awareness of biases [40] or accounting for biases during label aggregation or learning [26, 58]. Despite of these efforts, the question of how designs of tasks—which are decided by the requesters with or without consideration of data biases—influence biases (e.g., strictness and fairness) in the crowdsourced judgements, in addition to the quality of judgements, is generally under-explored. Answers to this question are likely nuanced, as how biased individuals are in their judgements may be dependent on both the characteristics of the individual and the judgement tasks.

Thus, in this study, we focus on two common categories of tasks designs—engaging workers into deeper deliberations and providing workers with feedback, to explore the influences of task designs on properties of crowdsourced judgements. Previous studies have shown that these two categories of task designs both have the potential to improve the quality of the crowdsourced data. For example, Schaekermann et al. [54] found that deliberation can help crowd workers to increase their accuracy in both objective and subjective text classification tasks. In addition, a growing line of recent studies demonstrate that on a judgement task, providing feedback from experts or peer workers to crowd workers, either in the form of tips [11, 17, 69], summary statistics of their judgements on the task [49], or justification of their judgements [10, 18, 59], can all help enhance the quality of the crowdsourced judgements. The advance of AI technologies recently also makes it possible to provide AI-powered agents' recommendations on judgement tasks to crowd workers as the feedback, which is also shown to increase the quality of the crowd's judgement in many cases [27, 29, 45, 61]. Compared to these studies, a key difference of our work is that we go beyond quality and also look into how deliberation and feedback affects the biases in crowdsourced judgement.

We use recidivism risk evaluation as a case study. This task domain has recently attracted a great amount of attention due to its intrinsic difficulty and fairness implications, which also make it a suitable domain to examine the impacts of task design to the quality and biases in crowdsourced judgement. Among the literature in the human studies of recidivism risk evaluation [7, 19, 27–29, 39, 41, 46, 48], it is observed that the accuracy of human judgement is comparable to machine learning predictions when humans are provided with ground-truth feedback [19], but human accuracy becomes worse without such feedback [41]. Moreover, the quality of human judgement could vary depending on the distribution of the experimental data, e.g., the ratio of re-offended defendants [41], and whether they are provided with feedback from machine learning models [27, 28]. Our goal aligns with this line of work and aims to more comprehensively examine how different task designs impact the crowd's recidivism judgements.

3 STUDY DESIGN

In this work, we conducted a case study on human recidivism judgement to explore how task designs influence the quality, strictness, and fairness of the recidivism judgements that people make.

3.1 Tasks: Recidivism Risk Judgement

Participants were recruited to complete a set of recidivism risk judgement tasks in our study. Specifically, in each task, participants were asked to review the profile of a criminal defendant in a vignette format. Six features of the defendant were shown in the vignette, including the defendant's race, gender, age, the name and degree of the current criminal charge, and the number of prior non-juvenile criminal charge[1]. After reviewing the defendant's profile, participants were asked to estimate the risk for the defendant to reoffend within the next two years on a scale of 0% to 100% in intervals of 10% (i.e., 0–10%, 11–20%, etc.). Figure A1 in the Appendices shows an example of our task interface.

The defendant profiles that we showed to participants were taken from the COMPAS dataset [46], and we restricted our attention to only those defendants whose race is Caucasian (white) or African-American (black). Previously, Biswas et al. [7] identified a subset of 1,000 defendants from the original COMPAS dataset that was balanced in terms of the defendant's race and the true reoffending status, while Dressel et al. [19] constructed another subset of 1,000 defendants (only 907 of them are white or black defendants) from the COMPAS dataset that was unbalanced in terms of the defendant's race and the true reoffending status, with black defendants associated with higher probability of reoffending (see Table 1 for the summary statistics). Henceforth, we refer to the defendant datasets in Biswas et al. [7] and Dressel et al. [19] as the *balanced dataset* and *unbalanced dataset*, respectively. We conducted our experiment twice, with the defendant profiles sampled from the balanced dataset in Experiment 1 and from the unbalanced dataset in Experiment 2. Since earlier studies found that people tend to suffer from their implicit racial bias [51], we conjecture that crowd workers' racial steoreotypes (i.e., the implicit associations between racial group and recidivism risks) will not get reinforced in Experiment 1 but will get reinforced in Experiment 2. Thus, through these two experiments, we can explore how the influences of task designs on the crowd's recidivism judgements vary as the strength of the racial stereotypes in people's mind changes. Note that both Biswas et al. and Dressel et al. recruited human subjects to review the defendant profiles in their respective datasets and make binary predictions on whether the defendant would reoffend or not. These subjects' binary predictions were released as a part of their datasets, and we later utilized this information in designing some of our treatments (i.e., the Peer feedback treatment).

3.2 Experimental Treatments

We focused on two main approaches to structure the recidivism risk tasks—either adding interventions in the task that encourage crowd workers to deliberate about their judgements, or providing feedback to crowd workers to allow reflections on the judgements. Based on these two approaches, we created 7 treatments, each corresponding to a unique task design (see Figure A2 for the interfaces):

- **Control**: In each task, participants were asked to review the defendant's profile and then estimate the recidivism risk. This reflects the simplest design of recidivism judgement tasks.
- **Competing hypothesis**: In each task, after making her initial recidivism risk estimate on the defendant, the participant was asked to consider the "competing hypothesis"—if her initial risk estimate was above (below) 50%, she would be asked to consider whether a risk of <50% (>50%) was possible for this defendant. The participant was asked to select the features in the defendant's profile that may support the competing hypothesis, and she had the option of providing some reasons to explain why. After considering the competing hypothesis, the participant made her final risk estimate.
- **Counterfactual thinking**: In each task, after making her initial recidivism risk estimate, we provided the participant with the profile of a fictional defendant, who had exactly the same features as the original defendant in this task except for having the *opposite* race. The participant was asked to estimate the recidivism risk for the fictional defendant. Then, she made her final risk estimate for the original defendant in the task.
- **Rethink**: In each task, after making her initial recidivism risk estimate, the participant was asked to spend at least 15 seconds to evaluate the defendant's case in more depth. After doing so, the participant was asked to make her final risk estimate.
- **Peer feedback**: In each task, after making her initial recidivism risk estimate on the defendant, the participant was presented with feedback on previous workers' judgements on the same defendant. Specifically, based on historic human subjects' binary recidivism predictions on the current defendant, we informed participants of the majority prediction made by historic workers on this defendant, as well as the fraction of historic workers who supported this majority prediction (in Experiments 1 and 2, we considered human subjects recruited by Biswas et al. [7] and Dressel et al. [19] as the historic workers, respectively). After viewing this information, the participant was asked to make her final risk estimate for the defendant.
- **ML model feedback**: In each task, after making her initial recidivism risk estimate on the defendant, the participant was presented with a machine learning (ML) model's prediction on the risk of the defendant reoffending. Earlier research has found that when judging recidivism risks, people tend to have higher false positive rate on black defendants than white defendants [19]. In light of this, we trained a "fair" machine learning model—a logistic regression model with constraints on false positive rate parity (with respect to the defendant's race)—using the Fairlearn API [6][2]. We showed to the participant the model's predicted probability for the defendant in this task to reoffend. Then, the participant was asked to make her final risk estimate.
- **Ground truth feedback**: In each task, after making her recidivism risk estimate on the defendant, the participant was presented with the ground truth answer, i.e., whether the defendant actually reoffended within the next two years.

These experimental treatments covered a variety of ways to incorporate deliberation (e.g., competing hypothesis, counterfactual

[1] While we included defendants' race in the task, our purpose was not to advocate the inclusions of such sensitive information. Instead, we aim to understand the potential biases of human judgements when such information is present.

[2] For the 1,000 profiles in the balanced dataset, the ML model's false positive rates on black and white defendants are 0.286 and 0.297, respectively. For the 907 profiles in the unbalanced dataset, the ML model's false positive rates on black and white defendants are 0.228 and 0.224, respectively. While not directly tuned for false negative rate parity or accuracy parity, this model has similar levels of false negative rate and accuracy on black and white defendants in both datasets.

	Biswas et al. [7] (balanced)									Dressel et al. [19] (unbalanced)		
	All cases			Twin cases			Extreme cases			All cases		
	All	*Black*	*White*	*All*	*Black*	*White*	*All*	*Black*	*White*	*All*	*Black*	*White*
N	1,000	500	500	54	27	27	17	7	10	907	530	377
Reoffend=Yes	475 (47.5%)	238 (47.6%)	237 (47.4%)	23 (42.6%)	10 (37.0%)	13 (48.1%)	5 (29.4%)	3 (42.9%)	2 (20%)	442 (48.7%)	402 (57.0%)	140 (37.1%)
Reoffend=No	525 (52.5%)	262 (52.4%)	263 (52.6%)	31 (57.4%)	17 (63.0%)	14 (51.9%)	12 (70.6%)	4 (57.1%)	8 (80%)	465 (51.3%)	228 (43.0%)	237 (62.9%)

Table 1: Summary statistics for the two datasets of criminal defendant profiles that we used in our experiment.

thinking, rethink) and feedback (peer feedback, ML model feedback, ground truth feedback) into the designs of the recidivism risk judgement task, and many of these task designs were easily generalizable to other judgement tasks. Moreover, among all treatments, the Counterfactual thinking and ML model feedback treatments represent two task designs that requesters explicitly take judgement biases into consideration and attempt to nudge crowd workers towards fairer judgements through the designs.

3.3 Experimental Procedure

Before running Experiment 1, we pre-processed the balanced defendant dataset to identify some "special" defendant profiles. First, we identified 27 pairs of profiles (i.e., 54 profiles) such that within each pair, the two defendants had exactly the same values on all six features except for the race. We called the two defendants within a pair as "twins" (see the "Twin cases" column in Table 1). These twin cases offered us a unique perspective in understanding the fairness level of crowd workers' recidivism judgements (i.e., do crowd workers make the same judgements on the twins?).

Furthermore, for each profile in the balanced dataset, Biswas et al. recruited 20 subjects, including 10 black subjects and 10 white subjects, to make recidivism prediction. By analyzing these subjects' predictions on each profile, we identified 17 profiles on which the probability *difference* for black subjects and white subjects to make a positive prediction (i.e., predict the defendant will reoffend) is *at least* 0.5. We refer to these 17 profiles as the "extreme cases" (see the "Extreme cases" column in Table 1). Interestingly, on all extreme cases that involve a black (white) defendant, the probability for white subjects to provide a positive prediction is *higher* (*lower*) than that probability for black subjects by at least 0.5. This means that on extreme cases, historic subjects exhibited a strong *in-group bias*, so these extreme cases provided us an opportunity to understand how task designs influence crowd workers' judgements on those tasks where they are particularly vulnerable to their own bias.

After preprocessing the dataset, we recruited participants for our Experiment 1 by posting a HIT on Amazon Mechanical Turk (MTurk). The HIT was only open to U.S. workers, and each worker could take it only once. Upon arrival, participants were randomly assigned to one of the seven treatments. Participants firstly went through the instructions which explained both the tasks and the bonus rules. Then, they completed a demographic survey as well as a cognitive reflection test [25] before starting to work on a sequence of 32 recidivism risk judgement tasks. The defendant profiles in these tasks were sampled from the balanced dataset as follows: We first randomly sampled 20 profiles from the entire dataset while ensuring the race of the defendant and the ground-truth reoffending status of the defendant were balanced (i.e., 5 black reoffending, 5 black not reoffending, 5 white reoffending, and 5 white not reoffending). We refer to these 20 profiles as the "general cases." Next, we randomly sampled two pairs of profiles from all twin cases in the dataset, resulting in a set of 4 twin profiles in total. Finally, from the set of extreme cases in the dataset, we further sampled 8 extreme profiles while ensuring the balance of the defendant's race and true reoffending status among them. Once these 32 defendant profiles were selected, they were presented to participants in a random order in the HIT, and participants made recidivism judgement for each defendant following the procedure as defined by the treatment that they were assigned. We also included attention check questions (i.e., questions in which participants were instructed to select a pre-specified option) in the HIT to enable the filtering of inattentive participants later.

The procedure of Experiment 2 was very similar to that of Experiment 1. The only differences were: (1) in each HIT, we only included 20 profiles that were randomly sampled from the unbalanced defendant dataset *without* ensuring the balance of the defendant's race and true reoffending status; (2) we did not include any twin cases or extreme cases in the HIT[3]; (3) participants of Experiment 1 were excluded from taking part in this experiment.

Experiment 1 was conducted on July 12–16, 2021, and Experiment 2 was conducted on September 22–24 and 27–28, 2021. Both experiments were conducted between 9am–6pm ET on weekdays[4]. The base payments of our Experiment 1 HIT and Experiment 2 HIT were \$1.6 and \$1.0, respectively. To incentivize participants to carefully review the defendant profiles and make accurate recidivism risk judgements, in both experiments, we provided additional bonuses to participants based on the accuracy of their risk estimates. In particular, in each task, we computed the amount of bonus payment a participant could receive using a Brier score function: $[score = 1 - (prediction - outcome)^2]$, where $prediction \in \{0.05, \ldots, 0.95\}$ was the midpoint of the *final* risk interval that the participant selected in the task, while $outcome \in \{0, 1\}$ was the ground truth answer of the task. We then mapped the Brier score for each task to a maximum bonus payment of \$0.05.

3.4 Analysis Methods

3.4.1 Independent and Dependent Variables. Our independent variable is the experimental treatment that a participant was assigned to, while the main dependent variables we consider are the quality, strictness, and fairness of the participant's recidivism judgements. Specifically, we first transformed the final recidivism risk interval that a participant selected in a task into a binary prediction using 50% as the threshold (i.e., any interval that was above 50% was transformed to the positive prediction of predicting the defendant will reoffend)[5]. Based on this transformation, we measured the *quality* of a participant's judgements as the accuracy of her binary predictions, and the *strictness* of the participant's judgements was

[3] We can not identify extreme cases for the unbalanced dataset as Dressel et al. did not share the race information for each human subject in their study.
[4] All of our experiments were approved by the IRB at the authors' institution.
[5] We also analyzed the data using the raw recidivism risk estimates participants provided, and the results were qualitatively similar.

operationalized as the positive prediction rate (POS) of the participant (i.e., the probability for the participant to predict a defendant will reoffend)—Intuitively, higher accuracy implies judgements of higher quality, and higher POS indicates that crowd workers were less "lenient" (i.e., stricter) in their judgements. Finally, we quantified the level of *fairness* of a participant's recidivism judgements in treating defendants of different races[6] using a few metrics [4]:

- *Positive prediction rate difference* (ΔPOS): the participant's POS on black defendents, minus that on white defendants.
- *False positive rate difference* (ΔFPR): the participant's false positive rate on black defendants, minus that on white defendants.
- *False negative rate difference* (ΔFNR): the participant's false negative rate on black defendants, minus that on white defendants.
- *Twin cases difference* (ΔTwin): Within a pair of twin profiles, the participant's binary prediction on the black defendant, minus that on the white defendant.

For all the fairness measures listed above, a value that is closer to zero implies fairer judgements. In particular, when a participant's ΔPOS is zero, her recidivism judgements satisfy the fairness definition of *demographic parity* [8]. Further, the fairness definition of *equalized odds* is effectively the same as requiring both ΔFPR and ΔFNR to be zero [31, 66]. Finally, if ΔTwin for a participant is zero, the participant's recidivism judgements satisfy the notion of *individual fairness* [21], i.e., treating similar individuals similarly.

3.4.2 Statistical Methods. We compared the participant's accuracy across treatments to examine the influences of task designs on judgement quality, and compared the participant's POS across treatments to examine the influences of task designs on judgement strictness. Further, to understand the influences of task designs on the fairness of participants' recidivism judgements, we compared the values of ΔPOS, ΔFPR, ΔFNR for participants' recidivism judgements across different treatments; when applicable, we also compared the values of ΔTwin for participants' recidivism judgements on the twin cases across different treatments. In all of these analyses, given a dependent variable, we conducted one-way analysis of variance (ANOVA) [62] or Kruskal-Wallis test [43]—depending on whether the data is normally distributed—to determine whether there is a significant difference across treatments on that dependent variable. In the case that a significant difference was detected in a one-way ANOVA test, we conducted pairwise comparisons with the Tukey HSD tests to identify pairs of treatments that exhibit significant differences. Similarly, if a significant difference was detected in a Kruskal-Wallis test, we used Dunn's tests to identify pairwise significant comparisons, and we applied the Benjamini-Hochberg (BH) adjustment [5] to correct for multiple comparisons.

4 RESULTS

In our experiments, 949 and 910 participants took the HIT and passed the attention check questions for Experiment 1 and Experiment 2, respectively. In the following, we analyze the data collected from these valid participants to understand the influences of task designs on the crowd's recidivism judgements.

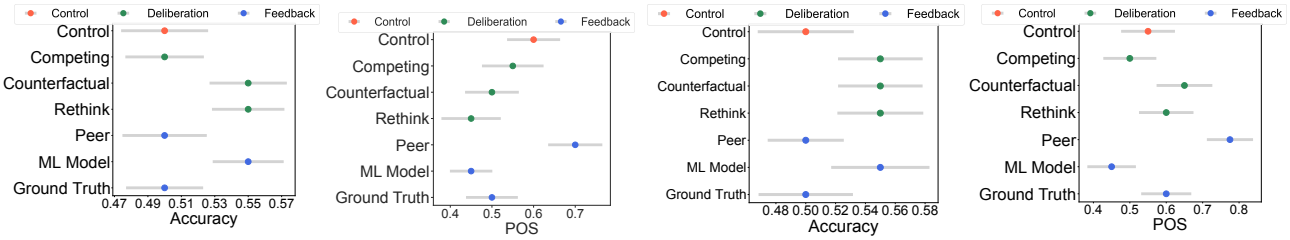
[6]We acknowledge that judgement fairness can be defined around defendants' other features beyond race, and real-world racial and ethnic identity is often not categorical.

4.1 RQ1: The Influences of Task Designs When the Strength of Stereotypes Vary

To begin with, we look into the general influences of task designs on the quality, strictness, and fairness of the crowd's recidivism judgements, when the judgement environment *does not reinforce* crowd workers' racial stereotypes (Experiment 1), as well as when it *does reinforce* crowd workers' racial stereotypes (Experiment 2).

4.1.1 Analysis of Experiment 1. First, we focus on analyzing the experimental data obtained from Experiment 1 to examine how different task designs influence the crowd's recidivism judgements when the strength of the racial stereotypes in people's mind is likely relatively weak. As participants' accuracy and positive prediction rate (POS) on the 20 *general cases* in Experiment 1 are not normally distributed, we present in Figures 1(a) and 1(b) their median values in different treatments. In terms of the quality of the crowd's recidvism judgements, we find that some task designs seem to slightly increase crowd workers' accuracy in their recidivism judgements, though results of the Kruskal-Wallis test suggest that these increases are not statistically significant ($p = 0.181$). Meanwhile, it is clear from Figure 1(b) that the designs of the recidivism judgement tasks significantly change the strictness of the crowd's judgements—compared to participants in the Control treatment, participants in all other treatments decrease their likelihood of making positive predictions on defendants, except for those participants in the Peer feedback treatment. A Kruskal-Wallis test confirms that the difference in participants' positive prediction rate on the 20 general cases across the seven treatments is significant ($p < 0.001$). Post-hoc Dunn's tests further suggest that participants in the Rethink and ML model feedback treatments were significantly *less* likely to make positive predictions compared to participants in the Control treatment (Control vs. Rethink: $p = 0.049$, Control vs. ML model feedback: $p = 0.044$). Moreover, for participants in the Peer feedback treatment, their positive prediction rates were shown to be significantly *higher* than participants in all other treatments ($p < 0.05$) except for those in the Control treatment, with the largest difference observed (when compared to the ML model feedback treatment) indicating a medium effect size of task designs on POS (i.e., Cohen's d=0.56).

Moving on to examine the influences of task designs on the fairness level of the crowd's recidivism judgements, Table 2 summarizes the average values of ΔPOS, ΔFPR, and ΔFNR on the 20 *general cases* in Experiment 1 for participants of different treatments. Since Experiment 1 included twin profiles, we also report in Table 2 the average values of ΔTwin in each treatment. According to the results in Table 2, when the defendant profiles were sampled from the balanced dataset and the recidivism tasks took the simplest design (i.e., the Control treatment), participants already exhibited a very high level of fairness in their recidivism judgements on defendants of different races, as the average values of all four fairness measures were close to zero. Moreover, the different designs of the recidivism judgement tasks do not seem to further affect the crowd's fairness in their judgements. In particular, the one-way analysis of variance (ANOVA) suggests that *none* of the differences on fairness measures across the seven treatments is significant at the level of $p = 0.05$.



(a) Balanced dataset: Accuracy (median) (b) Balanced dataset: POS (median) (c) Unbalanced dataset: Accuracy (median) (d) Unbalanced dataset: POS (median)

Figure 1: The influences of task designs on the quality and strictness of participants' recidivism judgements on the 20 general cases in Experiment 1 (a, b) and on all tasks in Experiment 2 (c, d). Error bars represent the 95% confidence interval.

Treatment	ΔPOS	ΔFPR	ΔFNR	ΔTwin
Control	0.013	0.003	−0.023	0.011
Competing hypothesis	0.036	0.051	−0.022	0.000
Counterfactual thinking	0.065	0.039	−0.092	0.061
Rethink	0.016	0.022	−0.010	0.015
Peer feedback	0.039	−0.002	−0.079	0.004
ML model feedback	0.038	−0.005	−0.082	−0.053
Ground truth feedback	0.036	0.026	−0.046	0.011
p-value (ANOVA)	0.406	0.523	0.076	0.199

Table 2: The effects of task designs on the fairness level of recidivism judgements in Experiment 1 (balanced dataset).

4.1.2 Analysis of Experiment 2. We now focus on analyzing the experimental data obtained from Experiment 2. Recall that in Experiment 2, the defendant profiles in each HIT were sampled from the unbalanced dataset, and we did not place any constraints in ensuring the balance of defendant's race and true reoffending status among the 20 selected defendant profiles in a HIT. This means that in each Experiment 2 HIT, participants likely reviewed more profiles from black defendants, who were associated with a higher probability of having a ground-truth label of reoffending compared to white defendants. So, via analyzing Experiment 2 data, we aim to understand how different task designs influence the crowd's recidivism judgements when people's racial stereotypes *get reinforced*[7].

Figures 1(c) and 1(d) compare the median accuracy and positive prediction rate, respectively, for participants across the seven treatments of Experiment 2. Similar as that in Experiment 1, we find that task designs do not have clear impacts on participants' accuracy in different treatments ($p = 0.086$), but they significantly change how likely participants would predict a defendant to recidivate ($p < 0.001$). Yet, compared to what we've observed in Experiment 1, we notice that on the unbalanced dataset of Experiment 2, *fewer* task designs nudge participants into making more lenient judgements. For example, compared to the most basic task design in the Control treatment, the two task designs that encourages participants to consider the counterfactual defendant or spend more time evaluating the defendant profile in a task both lead to a slight increase in participants' positive prediction rate on the unbalanced dataset, which is different from their influences on the strictness of the crowd's judgements on the balanced dataset. Still, participants in the Peer feedback treatment were the strictest in their judgements, as they were significantly more likely to predict a defendant to recidivate than participants in all other treatments ($p < 0.05$), except for the Counterfactual thinking and Rethink treatments.

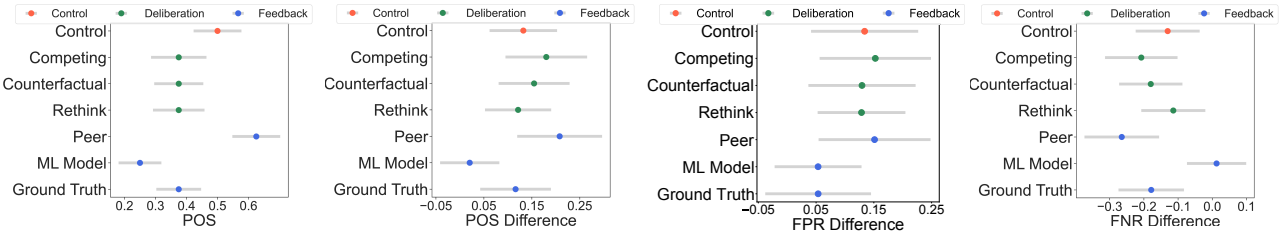
Lastly, we examine the effects of task designs on the fairness level of participants' recidivism judgements in Experiment 2, and we still find that the designs of the tasks do not significantly influence judgement fairness. For detailed results, see Table A1 in Appendices.

4.1.3 Summary. Taken together, our analysis of both experiments indicate that *in general*, the recidivism judgement task designs that we have considered in this experiment do not have clear influences on the quality and fairness of the crowd's recidivism judgements. However, the task designs do have substantial impacts on the *strictness* of the crowd's judgements—On both balanced and unbalanced datasets, showing historic human subjects' recidivism judgements to a crowd worker always leads to the strictest judgement, while showing the ML model's prediction tend to result in relatively lenient judgements. A closer look into the data suggests that compared to the recidivism judgement made by an average participant on a defendant profile in the Control treatment, historic human subjects are more likely to make a positive prediction while the ML model is more likely to make a negative prediction. Thus, the impacts of task designs on the strictness of participants' judgements could be caused by people's tendency to "match" the feedback received.

4.2 RQ2: The Influences of Task Designs on Extreme Cases

In Section 4.1, we have examined the influences of task designs on the crowd's recidivism judgements *in general* without differentiating the "difficulty" of the judgement task. In practice, however, crowd workers may find some recidivism judgement tasks to be easy as the defendant's profile contains clear "clues" in support of a certain judgement, while other tasks may be more difficult so that crowd makers are subject to their own biases to a higher extent when making judgements on them. Recall that in the balanced dataset, we identified a few such difficult tasks that can trigger high levels of in-group bias from humans (i.e., the "*extreme cases*"), and we included these tasks in our Experiment 1 HIT. Thus, in this subsection, we restrict our attention to analyze participants' judgements on the *extreme cases* in Experiment 1 to understand the influences of task designs on the crowd's recidivism judgements when the crowd is particularly vulnerable to their own biases.

[7]Though only those participants in the Ground truth feedback treatment could see the ground truth reoffending status for each defendant in their HITs, we suspect that participants in all treatments could "sense" the black defendants' higher reoffending probability through the defendant profiles.



(a) Extreme Case: POS (median) (b) Extreme Case: ΔPOS (mean) (c) Extreme Case: ΔFPR (mean) (d) Extreme Case: ΔFNR (mean)

Figure 2: The influences of task designs on the strictness and fairness levels of participants' recidivism judgements on the extreme cases in Experiment 1. Error bars represent the 95% confidence interval.

Specifically, we repeated the same set of analyses as that in Section 4.1 within the set of data obtained on extreme cases in Experiment 1. Consistent with our results in Section 4.1, we still find no evidence suggesting that task designs have any significant influences on the quality of the crowd's recidivism judgements on these extreme cases ($p = 0.690$), but we detect significant differences across the seven treatments with respect to how strict crowd workers were in making their judgements ($p < 0.001$; the largest difference was found between the ML MODEL FEEDBACK and PEER FEEDBACK treatments with a Cohen's d=0.80)—As shown in Figure 2(a), providing the ML model's prediction to crowd workers leads to a significantly lower positive prediction rate on the extreme cases than all other task designs ($p < 0.01$) except for the design that asks crowd workers to spend more time (i.e., the RETHINK treatment), while providing the historic subjects' judgement to crowd workers leads to a significantly higher positive prediction rate on the extreme cases than all other task designs ($p < 0.01$) except for the basic design in the CONTROL treatment. In addition, participants in the RETHINK treatment also made significantly more lenient judgements than those in the CONTROL treatment ($p < 0.05$).

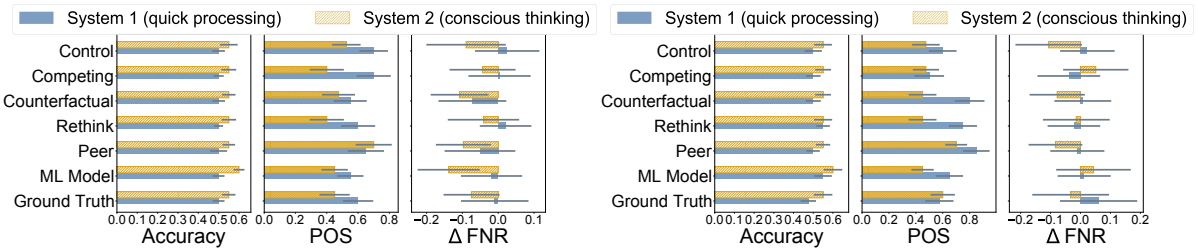
Most interestingly, we find that on the extreme cases, various task designs can significantly influence the fairness level of the crowd's recidivism judgements. More specifically, Figures 2(b)–2(d) compare the average values of the three fairness measures on the extreme cases across the seven treatments. Note that on the extreme cases, participants did show a considerable degree of unfairness in treating defendants of different races when the tasks took the most basic design in the CONTROL treatment—they were much more likely to make positive predictions on black defendants, leading to a higher FPR and a lower FNR on black defendants than white defendants. This is consistent with the fact that the majority of participants in our experiment self-reported to be Caucasian, while the selected extreme cases represent those tasks on which people might be more vulnerable to in-group bias in their recidivism judgements. Conducting one-way ANOVA on each of the three fairness measures, we find that there are significant differences in the average values of ΔPOS ($p < 0.001$) and ΔFNR ($p < 0.001$) across treatments. Results of the post-hoc pairwise comparisons show that these significant differences are mostly caused by participants who received feedback from the *fair* ML model—participants in the ML MODEL FEEDBACK treatment had ΔPOS and ΔFNR values that were significantly closer to zero than those in the PEER FEEDBACK, COMPETING HYPOTHESIS and COUNTERFACTUAL THINKING treatments ($p < 0.01$), and they also had closer-to-zero ΔFNR than participants in the GROUND TRUTH FEEDBACK treatment ($p = 0.002$). Again, the largest differences were detected between the ML MODEL FEEDBACK treatment and the PEER FEEDBACK treatment, which suggest the effects of task designs on the judgement fairness on the extreme cases are of medium size (d=0.57 for ΔPOS and d=0.64 for ΔFNR).

In sum, our analysis on the extreme cases suggest that on the "difficult" tasks where people tend to exhibit a high degree of in-group bias, changing the designs of the tasks may significantly affect the strictness and fairness of the crowd's recidivism judgements.

4.3 RQ3: Individual Differences in The Impact of Task Designs on Recidivism Judgements

Finally, we look into the potential individual difference in the influences of task designs. In particular, the dual process theory in psychology suggests that biases in human decision making may be explained by the type of cognitive processes that people engage in when making decisions [13, 23, 42]—"System 1" processing is executed quickly without reflection, while "System 2" requires conscious thought and effort [25, 57]. Frederick [25] then designed the *cognitive reflection test* (CRT) to identify the cognitive processes that an individual tends to engage with, with higher CRT scores implying more frequent use of conscious processing. As participants completed the CRT in our experiment, we were able to compute each participant's CRT score (following the method in [55]) and use a median split to separate them into two groups—one group heavily utilizes System 1 processing (i.e., "quick processing participants") while the other group mostly engages in System 2 processing (i.e., "conscious thinking participants") in their decision making. We then explored whether the influences of task designs on the recidivism judgements vary between these two groups of participants.

Figures 3(a) and 3(b) compare the quality, strictness, and fairness (only ΔFNR is shown; see Figure A3 for additional plots) of recidivism judgements made on the balanced dataset (general cases) and the unbalanced dataset, respectively, across the two groups of participants. Overall, compared to participants who utilize quick processing more, conscious thinking participants are associated with making more accurate, less strict, and surprisingly, less fair recidivism judgements. More interestingly, we find that the task designs have *stronger* influences on the recidivism judgements made by conscious thinking participants than those made by quick processing participants. For example, when participants were asked to make recidivism judgements on the balanced dataset (i.e., Figure 3(a)), the designs of the judgement tasks are shown to *only* affect the positive prediction rate of conscious thinking participants ($p < 0.001$) but not quick processing participants ($p = 0.688$). Similarly, on the unbalanced dataset (i.e., Figure 3(b)), while the



(a) Balanced dataset (b) Unbalanced dataset

Figure 3: The influences of task designs on the quality, strictness, and fairness of System 1 (quick processing) and System 2 (conscious thinking) participants' recidivism judgements on the 20 general cases in Experiment 1 (a) and on all tasks in Experiment 2 (b). Median values are reported for accuracy and POS, while mean values are reported for ΔFNR. Error bars represent the 95% confidence interval.

task designs change the positive prediction rate for both quick processing participants ($p = 0.031$) and conscious thinking participants ($p = 0.004$), we also find some evidence indicating that the task designs influence the fairness level of the recidivism judgements (on ΔFNR) but only for conscious thinking participants ($p = 0.040$).

Furthermore, we restrict our attention to the extreme cases in Experiment 1 to understand whether quick processing participants and conscious thinking participants are influenced by the task designs in different ways on tasks that they might be vulnerable to biases. Again, we find that the significant influences of task designs on the strictness of participants' recidivism judgements are only observed among conscious thinking participants ($p < 0.001$) but not quick processing participsants. Similarly, we find that the task designs only significantly affect the fairness of participants' judgements on extreme cases for conscious thinking participants—providing the feedback from a fair ML model only leads to substantially fairer judgements on those cases where people are vulnerable to their in-group biases, if the participants tend to engage more in conscious thinking (see Table A2 in Appendices for details).

5 CONCLUSIONS AND DISCUSSIONS

We present an experimental study that examines the influence of task designs on the quality, strictness, and fairness of the crowd's recidivism risk judgements. We focus on two common categories of task designs–encouraging deliberation and providing feedback. Via two randomized experiments in which the racial stereotypes might be either weak or strong in crowd workers' minds, we show that the task design choices made by requesters can have substantial impacts on how biased the crowd's recidivism judgements are, in terms of both the crowd's overall tendency to make lenient judgements (i.e., strictness of judgements) and the extent to which the crowd treats defendants of different races equally (i.e., fairness of judgements) when they are vulnerable to their in-group biases. Moreover, we observe individual differences in the influences of task designs on crowdsourced recidivism judgements, with the judgements of those individuals who tend to engage more in slow and deliberative thinking affected by the task designs to a larger degree.

Our study has a few limitations. First, we'd like to emphasize the case study nature of this work—the primary goal of our study is to use recidivism risk evaluation as an example to investigate whether and to what extent task designs affect the properties of crowdsourced judgements, especially regarding how biased these judgements are. We believe the recidivism risk evaluation task is representative of a family of crowdsourced judgement tasks in which some judgement is generally perceived to be "preferable" (e.g., rate a job applicant as "qualified", determine a loan applicant as "credible"). We conjecture the findings of our study are more likely to generalize to this family of tasks. However, it's also possible that our results will not directly generalize to these tasks due to the uniqueness of the task domain. Therefore, more future studies should be conducted to examine the generalizability of our results in other domains thoroughly. Secondly, our *correlational* finding that workers who engaged more with conscious thinking produced more unfair judgements compared to workers who engaged more with quick processing is surprising, and we do not know why. It's possible that there exists a third explanatory factor, such as quick processing workers were simply less careful in reading the task information, but it may also relate to the deeper mechanisms underlying how stereotypes are formed and triggered [3]. Future studies should be conducted to understand this counter-intuitive result. We also acknowledge that due to the distributed nature of crowd work, we can not guarantee that workers have sufficiently engaged with the interventions that we included in different designs of the tasks. For instance, in the Rethink treatment, we included a 15-second timer on the interface to encourage subjects to spend more time evaluating the case in the task. Although subjects could not proceed to the next stage before the timer was up, it's possible for subjects to switch to another HIT during this 15-second period.

Despite the limitations, our findings suggest a few important implications. First, they highlight the importance of obtaining a nuanced understanding of how task designs influence various properties of crowdsourced judgements beyond quality, as some properties of the judgements can be highly sensitive to subtle changes in the task designs, even if requesters do not select the designs with the intention to influence those properties. In this sense, requesters should be aware of the possible unintended consequences of their task design choices. Deeper understandings of the relationships between task designs and properties of crowdsourced judgements can also inform better selection of task designs, or even enable personalized task designs given the observed individual differences. Another key lesson is that requesters should carefully select the feedback to present to crowd workers in judgement tasks as they may largely shape the crowd's judgements, which can imply both positive and negative outcomes (e.g., feedback from a fair ML model

leads to fairer judgements, feedback from biased peers leads to unfair judgements). We hope this work could open more discussions on obtaining deeper understandings of crowdsourced task designs.

ACKNOWLEDGMENTS

This work is supported in part by the NSF FAI program in collaboration with Amazon under grant IIS-1939677 and IIS-2040800 and the Office of Naval Research Grant N00014-20-1-2240.

REFERENCES

[1] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3665–3674.

[2] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (2021), eabf4393. https://doi.org/10.1126/sciadv.abf4393 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abf4393

[3] Hal R Arkes. 1991. Costs and benefits of judgment errors: Implications for debiasing. *Psychological bulletin* 110, 3 (1991), 486.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *Nips tutorial* 1 (2017), 2017.

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[6] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[7] Arpita Biswas, Marta Kolczynska, Saana Rantanen, and Polina Rozenshtein. 2020. The Role of In-Group Bias and Balanced Data: A Comparison of Human and Machine Recidivism Risk Predictions. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 97–104.

[8] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[9] 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.

[10] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[11] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. 2018. Crowd coach: Peer coaching for crowd workers' skill growth. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–17.

[12] Sharath R. Cholleti, Sally A. Goldman, Avrim Blum, David G. Politte, and Steven Don. 2008. Veritas: Combining expert opinions without labeled data. In *Proceedings 20th IEEE international Conference on Tools with Artificial intelligence (ICTAI)*.

[13] Pat Croskerry, Geeta Singhal, and Sílvia Mamede. 2013. Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ quality & safety* 22, Suppl 2 (2013), ii58–ii64.

[14] A. P. Dawid and A. M. Skene. 1979. Maximum likeihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* 28 (1979), 20–28.

[15] Luca De Alfaro and Michael Shavlovsky. 2014. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM technical symposium on Computer science education*. 415–420.

[16] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39 (1977), 1–38.

[17] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1013–1022.

[18] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4.

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

[20] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. 2020. Does Exposure to Diverse Perspectives Mitigate Biases in Crowdwork? An Explorative Study. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 155–158.

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

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

[23] Jonathan St BT Evans and Keith Ed Frankish. 2009. *In two minds: Dual processes and beyond.* Oxford University Press.

[24] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*. 1–4.

[25] Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* 19, 4 (2005), 25–42.

[26] Meric Altug Gemalmaz and Ming Yin. 2021. Accounting for Confirmation Bias in Crowdsourced Label Aggregation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*.

[27] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[28] Ben Green and Yiling Chen. 2020. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *arXiv preprint arXiv:2012.05370* (2020).

[29] Nina Grgić-Hlača, 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), 1–25.

[30] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1914–1933.

[31] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[32] Ethan Haworth, Ted Grover, Justin Langston, Ankush Patel, Joseph West, and Alex C Williams. 2021. Classifying Reasonability in Retellings of Personal Events Shared on Social Media: A Preliminary Case Study with/r/AmITheAsshole. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 1075–1079.

[33] Chien-Ju Ho, Rafael Frongillo, and Yiling Chen. 2016. Eliciting categorical data for optimal aggregation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2450–2458.

[34] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on Machine Learning*. 534–542.

[35] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*.

[36] Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. 2016. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research* 55 (2016), 317–359.

[37] Chien-Ju Ho, Yu Zhang, Jennifer Wortman Vaughan, and Mihaela Van Der Schaar. 2012. Towards social norm design for crowdsourcing markets. In *Proceedings of the 4th Human Computation Workshop*.

[38] John Joseph Horton and Lydia B. Chilton. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce (EC)*.

[39] Xinlan Emily Hu, Mark E Whiting, and Michael S Bernstein. 2021. Can Online Juries Make Consistent, Repeatable Decisions?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[40] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 407.

[41] Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. 2020. The limits of human predictions of recidivism. *Science advances* 6, 7 (2020).

[42] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[43] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.

[44] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. *Advances in Information Retrieval* 12036 (2020), 207.

[45] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[46] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9, 1 (2016).

[47] Tianyi Li, Chandler J Manns, Chris North, and Kurt Luther. 2019. Dropping the baton? Understanding errors and bottlenecks in a crowdsourced sensemaking

pipeline. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[48] Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, and Ece Kamar. 2020. Do I Look Like a Criminal? Examining how Race Presentation Impacts Human Judgement of Recidivism. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[49] Lena Mamykina, Thomas N Smyth, Jill P Dimond, and Krzysztof Z Gajos. 2016. Learning from the crowd: Observational learning in crowdsourcing communities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2635–2644.

[50] Winter Mason and Duncane Watts. 2009. Financial Incentives and the "Performance of Crowds". In *Proceedings of the 1st Human Computation Workshop (HCOMP)*.

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

[52] Vikas Raykar, Shipeng Yu, Linda Zhao, Gerardo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11 (2010), 1297–1322.

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

[54] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[55] Aleksandr Sinayev and Ellen Peters. 2015. Cognitive reflection vs. calculation in decision making. *Frontiers in psychology* 6 (2015), 532.

[56] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management* 58, 6 (2021), 102710.

[57] Keith E Stanovich and Richard F West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences* 23, 5 (2000), 645–665.

[58] Wei Tang and Chien-Ju Ho. 2019. Bandit Learning with Biased Human Feedback. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*. 1324–1332.

[59] Wei Tang, Ming Yin, and Chien-Ju Ho. 2019. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*. 1794–1805.

[60] Carlos Toxtli, Angela Richmond-Fuller, and Saiph Savage. 2020. Reputation Agent: Prompting Fair Reviews in Gig Markets. In *Proceedings of The Web Conference 2020*. 1228–1240.

[61] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[62] Larry Wasserman. 2004. *All of statistics: a concise course in statistical inference*. Vol. 26. Springer.

[63] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems (NIPS)*.

[64] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 547–558.

[65] Ming Yin and Yiling Chen. 2016. Predicting crowd work quality under monetary interventions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 259–268.

[66] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[67] 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).

[68] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.

[69] Haiyi Zhu, Steven P Dow, Robert E Kraut, and Aniket Kittur. 2014. Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1445–1455.

A APPENDICES

A.1 Task Interfaces

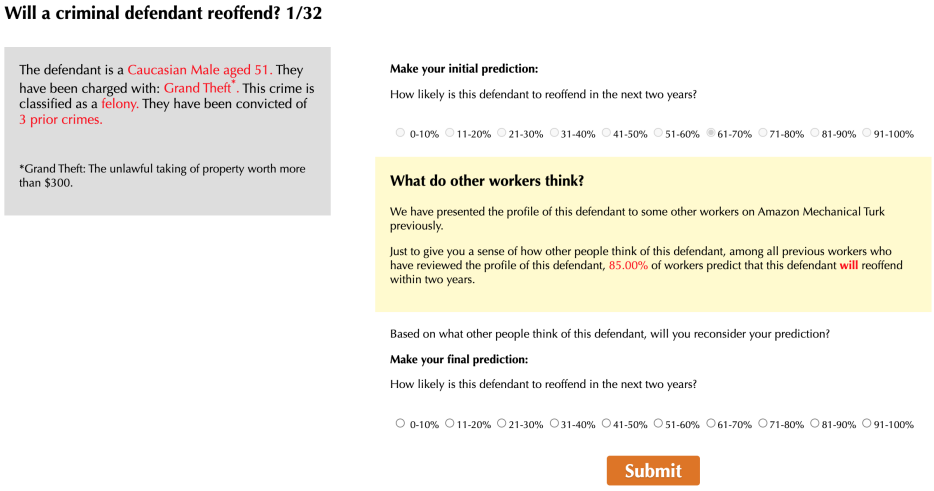
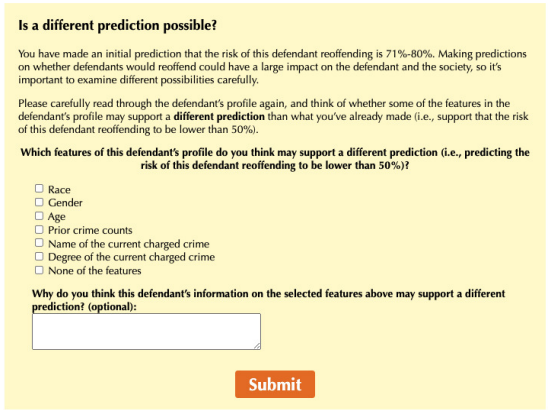
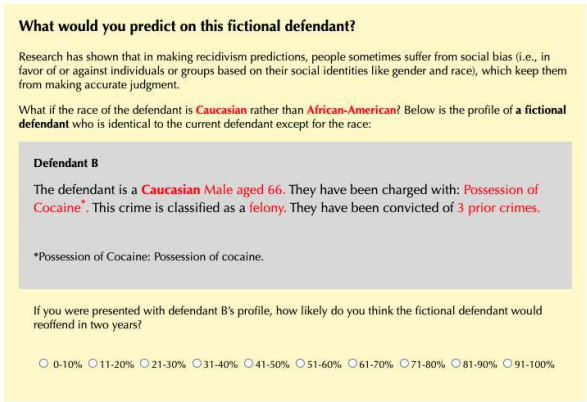


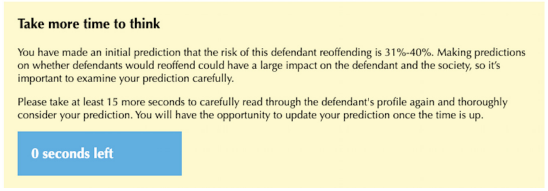
Figure A1: An example of the recidivism risk judgement task interface in the PEER FEEDBACK treatment. Left: the criminal defendant's profile. Top-Right: Participants were asked to make an initial recidivism risk estimate. Middle-Right: Participants received their treatments (shown in yellow background). In this case, we showed the participant the majority prediction given by historic human subjects on this defendant. Bottom-Right: Participants were asked to make a final risk estimate.



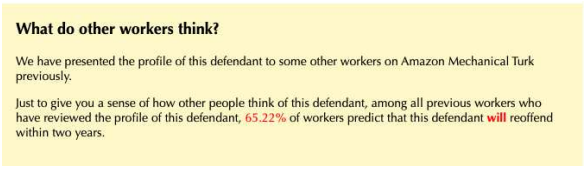
(a) Competing hypothesis



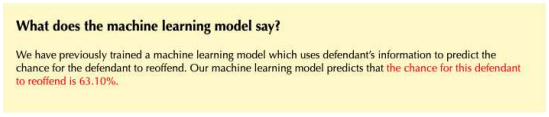
(b) Counterfactual thinking



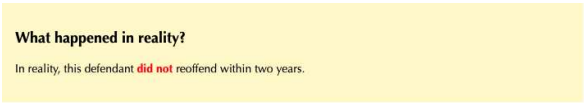
(c) Rethink



(d) Peer feedback



(e) ML model feedback



(f) Ground truth feedback

Figure A2: Interface of the deliberation or feedback components that subjects saw in different treatments of our study.

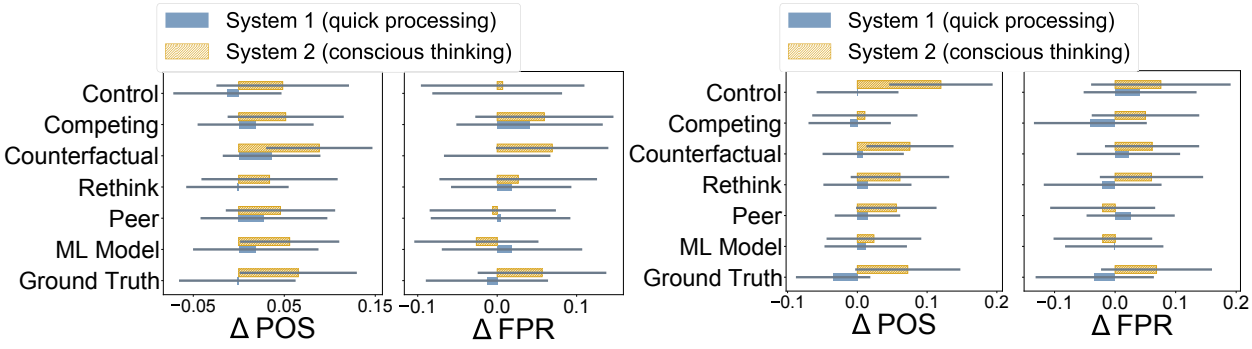
A.2 Additional Results

Treatment	ΔPOS	ΔFPR	ΔFNR
Control	0.051	0.056	−0.032
Competing hypothesis	0.001	0.009	0.010
Counterfactual thinking	0.041	0.041	−0.034
Rethink	0.038	0.020	−0.018
Peer feedback	0.036	0.001	−0.049
ML model feedback	0.018	−0.011	0.028
Ground truth feedback	0.021	0.020	0.012
p-value (ANOVA)	0.451	0.424	0.371

Table A1: The effects of task designs on the fairness level of participants' recidivism judgements in Experiment 2.

Treatment	ΔPOS		ΔFPR		ΔFNR	
	Sys1	Sys2	Sys1	Sys2	Sys1	Sys2
Control	0.050	0.246	0.069	0.224	−0.031	−0.267
Competing hypothesis	0.102	0.246	0.083	0.211	−0.120	−0.281
Counterfactual thinking	0.049	0.237	0.025	0.212	−0.074	−0.263
Rethink	0.040	0.202	0.044	0.213	−0.037	−0.191
Peer feedback	0.069	0.299	0.032	0.229	−0.106	−0.368
ML model feedback	−0.03	0.065	0.007	0.095	0.067	−0.036
Ground truth feedback	0.020	0.192	−0.016	0.109	−0.056	−0.276
p-value (ANOVA)	0.287	**0.001**	0.784	0.152	0.144	**<0.001**

Table A2: The effects of task designs on the fairness level of recidivism judgements in Experiment 1 (balanced dataset) extreme tasks for participants in different groups.



(a) Balanced dataset

(b) Unbalanced dataset

Figure A3: The influences of task designs on the average fairness levels (ΔPOS and ΔFPR) of System 1 (quick processing) and System 2 (conscious thinking) participants' recidivism judgements on the 20 general cases in Experiment 1 (a) and on all tasks in Experiment 2 (b). Error bars represent the 95% confidence interval.