Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study

NIELS VAN BERKEL, University College London, UK
JORGE GONCALVES, The University of Melbourne, Australia
DANULA HETTIACHCHI, The University of Melbourne, Australia
SENURI WIJENAYAKE, The University of Melbourne, Australia
RYAN M. KELLY, The University of Melbourne, Australia
VASSILIS KOSTAKOS, The University of Melbourne, Australia

The increased reliance on algorithmic decision-making in socially impactful processes has intensified the calls for algorithms that are unbiased and procedurally fair. Identifying fair predictors is an essential step in the construction of equitable algorithms, but the lack of ground-truth in fair predictor selection makes this a challenging task. In our study, we recruit 90 crowdworkers to judge the inclusion of various predictors for recidivism. We divide participants across three conditions with varying group composition. Our results show that participants were able to make informed decisions on predictor selection. We find that agreement with the majority vote is higher when participants are part of a more diverse group. The presented workflow, which provides a scalable and practical approach to reach a diverse audience, allows researchers to capture participants' perceptions of fairness in private while simultaneously allowing for structured participant discussion.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Collaborative and social computing**.

Additional Key Words and Phrases: Algorithmic decision making; intelligible models; artificial intelligence; bias; crime; modelling bias; perceived fairness; chatbots; crowdsourcing; fairness.

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 28 (November 2019), 21 pages. https://doi.org/10.1145/3359130

1 INTRODUCTION

Data analysis and inference is increasingly being used for decision making in high-stakes processes [6, 20, 55], such as an individual's likelihood to default on a loan or to commit a crime [20]. Recent work has raised the question whether such algorithms are fair [2, 20, 36, 56, 60], citing *inter alia* discriminatory side-effects [2, 26]. Although some techniques can quantify the fairness of an algorithm by applying a predetermined perspective of fairness [37], recent work has stressed the

Authors' addresses: Niels van Berkel, n.vanberkel@ucl.ac.uk, University College London, 66-72 Gower St, London, WC1E 6EA, UK; Jorge Goncalves, jorge.goncalves@unimelb.edu.au, The University of Melbourne, Parkville, VIC, 3010, Australia; Danula Hettiachchi, danula.hettiachchi@unimelb.edu.au, The University of Melbourne, Parkville, VIC, 3010, Australia; Senuri Wijenayake, wijenayakes@unimelb.edu.au, The University of Melbourne, Parkville, VIC, 3010, Australia; Ryan M. Kelly, ryan.kelly@unimelb.edu.au, The University of Melbourne, Parkville, VIC, 3010, Australia; Vassilis Kostakos, vassilis.kostakos@unimelb.edu.au, The University of Melbourne, Parkville, VIC, 3010, Australia.

2573-0142/2019/11-ART28 $15.00
https://doi.org/10.1145/3359130

importance of obtaining human perspectives on fairness in algorithm development [19]. While there are multiple steps in the development of an algorithm which can influence the fairness of the outcome, previous work has highlighted that the perceived fairness of predictors themselves can be utilised to inform fair algorithmic decision making [21]. However, developing a scalable and practical feature selection process capable of obtaining the opinion of a diverse population sample remains an open research question [20, 55, 60]. As a first step in overcoming this challenge, here we explore the perceptions of crowdworkers with regards to fair predictors in a recidivism prediction model.

A growing body of work has begun exploring the use of 'intelligible' models for prediction [8]. Unlike typical 'black box' models, intelligible models can be visualised and easily understood by humans, allowing for intervention if undesired rules are added to the model. Our work is inspired by the recent controversy surrounding *COMPAS*, an algorithm aimed at predicting the risk of recidivism among defendants and thought to have a racial bias [2, 13, 26]. Here we describe how we built an intelligible model using a publicly available dataset on recidivism using a variety of predictors. We then present the generated 'intelligible' plots to crowdworkers, and ask them to make a judgement as to whether the presented predictor ought to be used for predicting recidivism. Using graphs generated through intelligible models, our crowdworkers consider not only the name or description of a predictor, but are able to interpret the underlying data for each predictor.

Zhu et al. argue for Value Sensitive Algorithm Design, in which stakeholders' tacit knowledge and feedback are incorporated in the early stages of algorithm design [64]. This call comes amid an increased focus within the HCI community on explainable, intelligible, and accountable systems [1]. In our work, we explore a novel approach for this societally relevant research agenda: using the knowledge of the crowd to obtain insight on the perceived fairness of predictors for Machine Learning (ML) models. Previous work has indicated that the crowd may not always agree with domain experts, and has begun to explore how the crowd can contribute to societally complex tasks [20, 24, 41]. Motivated by the lack of ground truth in identifying potential biases introduced by predictor selection, we explore the wisdom of the crowd to discern fair predictors for ML models [20, 24]. We develop an automated and scalable system that can guide participant's decision making and discussion through an interactive bot. Our study evaluates three ways of organising crowdworkers: individually, in a group of three with an appointed group lead, and in an unstructured group of three without lead. We analyse and discuss how these conditions affected discussion and the swaying of crowdworker perceptions on fairness, and reflect on the use of bots as a research instrument in socially sensitive topics.

Our results indicate that crowdworkers are able to make an informed decision about the perceived fairness of using particular predictors for ML algorithms, and utilise the intelligible plots to inform their decision making. We find that members of a structured group are more likely to change their opinion than those in an unstructured group. Furthermore, we find that more diverse groups are in higher agreement with the overall majority vote. Our results inform future work into the perception and selection of fair ML predictors by demonstrating how subjective feedback can be gathered from the crowd. We present a structured and scalable method for the collection of sensitive opinions which can be extended for future studies on algorithmic fairness.

2 RELATED WORK

We motivate our work by drawing together insights from algorithmic decision making in HCI, interpretability and explainability of machine learning models, and the application of crowdsourcing solutions to complex issues.

2.1 Algorithmic Fairness in HCI

Recent work highlights the increasing ubiquity of algorithms in day-to-day life [1, 2, 26]. Although the quantification and discussion around fair and unfair assessment tests has been ongoing for

at least 50 years [25], the renewed impact of recent ML algorithms on our lives has given rise to increased concern surrounding the fairness of these algorithms. Fairness can be defined as "*the lack of discrimination or bias in decision making*" [37]. Research in Computer Science, HCI, Justice, and the Social Sciences has begun to explore techniques in which algorithmic fairness can be measured. In doing so, the literature has identified a diverse range of perspectives on how fairness can be operationalised and quantified in day-to-day practice, including:

- **Group fairness**. Each group identified in the dataset receives an equal fraction of a possible outcome (applies to both positive and negative outcomes) [7].
- **Individual fairness**. Individuals with similar characteristics should be treated similarly [14].
- **Equality of opportunity**. Individuals with an equal amount of talent and motivation should be offered the same prospective, regardless of their current position within the social system [22].

These different perspectives each adopt a unique notion on fairness, often with partial overlap. Hutchinson & Mitchell note that discussion on fairness was previously fuelled by practical needs of society, politics, and the law, but that recent work has been less tied to practical needs [25]. As such, the authors propose to increase the accessibility of fairness work for the general public by relating the discussion around fairness to ethical theories and value systems. Green & Chen analyse the use of a risk assessment tool with crowdworkers and reveal under-performance of risk assessments, an inability to evaluate either their own or the risk assessment tool's performance, and the display of 'disparate interaction' (use of risk assessment resulted in higher risk predictions for a disadvantaged group) [19]. These results highlight the importance of considering the sociotechnical context in which ML algorithms are deployed. End-users apply their own perspective and interpretation on the level of fairness of the presented suggestions, potentially allowing risk prediction systems to "*become a leverage point around which discrimination manifests*" [19]. Green & Chen's work concludes that risk assessment systems, such as the recidivism system analysed in our paper, must be "*grounded in rigorous evaluations of their real-world impacts instead of in their theoretical potential*" [19]. This paper focuses specifically on the human perception of algorithmic fairness, addressing the gap between the theoretical and practical perspective on fairness in algorithmic systems.

The *perceived* fairness of algorithms, *i.e.* how end-users, or individuals otherwise involved, comprehend the fairness of decisions made by AI algorithms, has received increased academic interest [6, 20, 36, 43, 60]. Grgić-Hlača et al. investigate how individuals perceive fairness in algorithmic decision making in the area of recidivism [20]. Their work explores why people perceive machine learning predictors as (un)fair through scenario-based surveys, and identifies 8 properties which can be used to predict people's perception of a predictor (*e.g.*, perceived reliability, relevance). Lee et al. [36] interview various stakeholders related to a non-profit organisation and find that notions of fairness differ considerably between participants. Woodruff et al. conduct interviews and workshops with traditionally marginalised groups and propose best practices for technology companies: include fairness in product development, accommodate diverse perspectives in user studies, and co-develop with community groups [60]. These studies collectively raise concerns about the perceived fairness of algorithmic decision making, and point to a need to integrate social values into the design of algorithms.

2.2 Explainability of algorithms

To develop transparent algorithms, it is critical for an algorithm's inner workings to be explainable. Previous work points to HCI as a key contributor to this area: "*Given HCI's focus on technology that benefits people, we, as a community, should take the lead to ensure that new intelligent systems are transparent from the ground up*" [1]. Previous work has explored decision making involving individual devices or applications. For example, Auda et al. consider how users can create a rule-based model to manage their smartphone notification load and customise it to their personal needs [3]. Kulesza et al. and Stumpf et

al. demonstrate through a set of experiments and with the use of classical machine learning models the benefit of explainable interfaces to users [33, 47]. Lee & Baykal describe how algorithmic mediation is perceived in group decision making scenarios: in order for algorithmic mediation to be perceived as fair, algorithms should account for social and altruistic behaviours [35]. Finally, Dietvorst et al. show that allowing users to make changes to an algorithm increases user trust in its application [12].

Algorithmic explainability has also been investigated in the application of high-level decision-making. Binns et al. focus on the perception of justice in algorithmic decision making systems [6]. Participants were presented with four types of textual explanations (*e.g.*, demographic-based, sensitivity-based) of an algorithm's behaviour, but their results show no conclusive effect of explanation style. Rader et al. conduct a large-scale survey in which over 5000 participants consider different explanation styles (*e.g.*, 'why', 'what') on the workings of Facebook's News Feed ranking algorithm [43]. All explanation types increased participant's awareness. Our work presents the use of plots to provide participants with an understanding of individual predictors. Visualisations of the data and accompanying intelligible model provide a more concrete impression of predictors than possible by text alone.

2.3 Interpretable Models

Continuous developments in the field of ML have led to more powerful and accurate modelling techniques. However, these improvements have also resulted in a reduced interpretability – defined as "*the degree to which an observer can understand the cause of a decision*" [5, 40]. Gilpin et al. argue that models should not only be interpretable but also 'explainable'; "*summarise the reasons for [...] behaviour, gain the trust of users, or produce insights about the causes of decisions*" [17].

As an example of the importance of interpretable models, Caruna et al. study pneumonia using general additive modelling (GAM) [8]. Their initial model contained a surprising rule: asthmatic patients are less likely to die from pneumonia than other pneumonic patients. In collaboration with domain experts, the authors discovered that asthma patients receive more intense treatment as compared to other patients – effectively increasing their chances of recovery. This knowledge was not captured in the dataset, and therefore not included in the model. A black box model would not have revealed this surprising rule, potentially endangering asthmatic patients as the model would consider them to be 'low risk' as based on past clinical outcomes.

In this work we explore human decision making in relation to potentially controversial predictors. Rather than relying on domain experts to distinguish 'right' from 'wrong', as shown in [8], we investigate how crowdworkers perceive fairness of predictors, and how they come to a decision.

2.4 Crowdsourcing Opinions

HCI has an extensive record of collecting opinions through crowdsourcing, both from individuals and groups. Xu et al. introduce '*Voyant*', allowing crowdworkers to offer feedback on professionally created visual designs [61]. Designers found the feedback offered by non-experts "*useful for comparing the perceptions of the crowd to their own expectations*" [61]. A key element in collecting honest opinions from groups is 'psychological safety': the ability "*to show and employ one's self without fear of negative consequences to self image, status, or career*" [27]. Previous work has considered psychological safety in crowdsourcing environments. Salehi et al. present '*Huddler*', a system that allows crowdworkers to complete tasks with familiar team members over time [44]. Although this improved psychological safety, it also introduced delays in team formation and frustration among crowdworkers. Work on chatroom interactions shows that user anonymity allows for increased risk-taking with revelations [50], and results in people disclosing more information than they would in face-to-face interaction [57]. This work collectively demonstrates the potential value of crowdsourcing in gathering genuine and anonymous input.

A recent avenue in crowdsourcing examines how societally challenging tasks can be effectively and accurately completed by crowds. Hosio et al. present a crowdsourcing solution to identify suitable treatments for lower-back pain, a medical problem which lacks therapeutic consensus [24]. The authors create two knowledge bases, one filled by clinical experts and one by non-experts [24]. Their results show that while these groups did not always agree on the perceived characteristics of treatment solutions, both clinical experts and non-experts saw benefits in the observation of non-expert contributions. Kriplean et al. and Kim et al. present online platforms to engage users in political discussion by including contextual, fact-checking, and social information layers, facilitating users to take the perspective of others in dividing issues [28, 32]. Finally, we note emerging work on the use of crowdsourcing in algorithmic decision making. Noothigattu et al. present 'Moral Machine', a system which allows users to consider ethical decisions in the case of autonomous vehicles (*e.g.*, should a vehicle avoid collision with a pedestrian at the cost of harming its passengers) [41]. The authors propose that this voting data can be used to construct a model of societal preferences that can be invoked to resolve an ethical dilemma.

Our work builds on previous findings in the area of algorithmic fairness, which has focused primarily on collecting opinions of individuals through focus groups [55] and surveys [20]. Previous work utilising crowdsourcing has demonstrated the usefulness of input from non-experts in complex and domain specific issues [24, 41, 61], but this has not yet been applied to predictor assessment. Our work explores the use of an interactive bot, an interface paradigm which has recently seen increased popularity [31], to facilitate individuals and groups in discussing fair predictors. Our system can be readily extended for future studies on this topic.

3 METHOD

We evaluate a crowdsourcing approach to identify fair predictors for machine learning models. Using an interactive chat environment, we consider the effect of individual vs. small groups (three people) in this decision process, following prior work on social computing systems in groups [53, 63] and the use of focus groups to discuss algorithmic fairness [60]. We introduce the following three composition conditions:

- **Individual**. Every crowdworker judges the perceived fairness of a predictor individually.
- **Group - structured**. Working in groups of three, with a random group-member appointed as leader. The leader decides when to continue to the next sub-task.
- **Group - unstructured**. Working in groups of three with no appointed leader. Group members must collectively agree when to continue to the next sub-task.

Workers were recruited via Amazon Mechanical Turk (MTurk), with the advertisement offering a short explanation of the task. After accepting the task, participants were routed to a Slack (a popular workplace collaboration tool) group where they were greeted by an interactive bot. Our bespoke bot allowed us to create a structured and highly interactive, yet scalable, study deployment. The bot managed predetermined tasks, such as presenting instructions to participants and initialising discussion at predefined moments.

3.1 Dataset & Machine Learning Model

We use a public dataset containing information on criminal defendants of Broward County, Florida[1]. The dataset contains demographic information, previous criminal history of the defendant, various scores such as expected risk of recidivism, and whether or not the defendant went on to commit another crime within two years (ground-truth recidivism). The dataset is a combination of a public

[1]Dataset available at https://github.com/propublica/compas-analysis/.

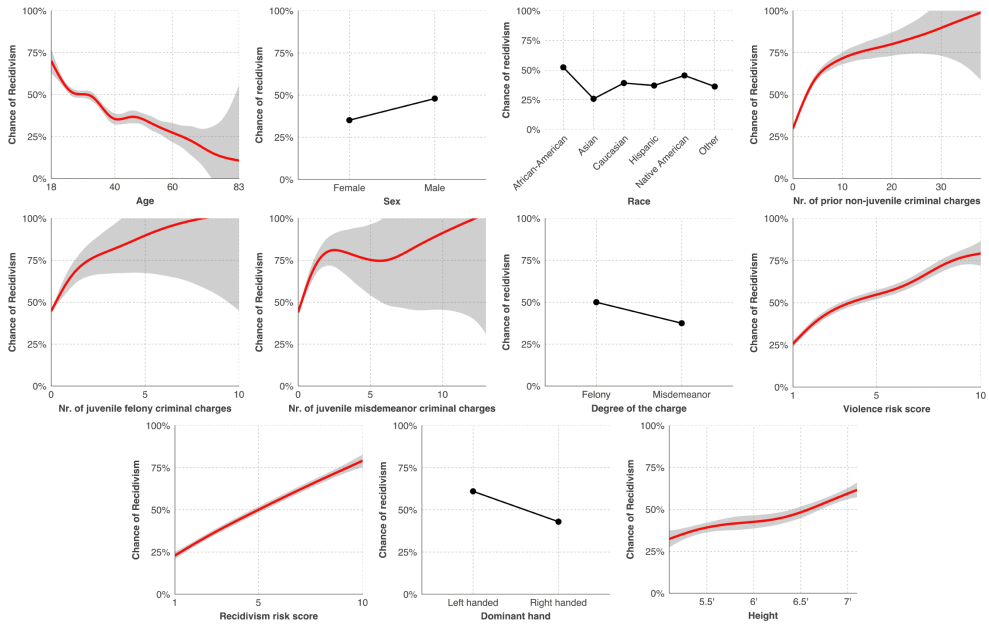

Fig. 1. The variables shown to participants. Grey areas demarcate confidence intervals. The variables 'Dominant Hand' and 'Height' were not part of the original dataset but included as explicitly verifiable predictors.

record requests on COMPAS scores (*i.e.*, risk scores) and public criminal records, as obtained by ProPublica [2].

After removing incomplete records, ordinary traffic offences (no jail time), and offenders who were released from a correctional facility less than two years ago (following [2]), the dataset contains information on 6172 defendants.

For each individual predictor we generate a plot based on the General Additive Model (GAM). GAM allows for more complex response shapes than offered through General Linear Models, while retaining a high level of intelligibility [23]. Using the GAM, we can generate a graph for each predictor to show its effect on the outcome variable (see Figure 1). We utilise the *mgcv*-package for R as described in [59], and fit all predictors individually using a smooth function. Our graphs considers only a single parameter at a time to force participants to reflect on the merits of individual predictors and to make the graphs less complex and more intuitive. We note that some of these parameters may have an interaction with each other, either a two-way interaction (*e.g.*, an amplified effect of age on recidivism among men but not among women), or more complicated effects such as a four-way interaction. Taking this into account would, however, further complicate the task, increasing the risk that our participants would be unable to comprehend the information shown to them. Therefore, we focus on predictors in isolation. We consider the following variables as predictors in our analysis:

- **Age**, ranges from 18–96, $M = 34.5$, $SD = 11.73$.
- **Sex**, following categorisation in the dataset as either male (4997) or female (1175).
- **Race**, following categorisation in the dataset: African-American (3175), Caucasian (2103), Hispanic (509), Asian (31), Native American (11), and Other (343).
- **Nr. of prior non-juvenile charges**, ranges from 0–38, $M = 3.25$, $SD = 4.74$.
- **Nr. of juvenile felony charges**, ranges from 0–20, $M = 0.06$, $SD = 0.46$.
- **Nr. of juvenile misdemeanour charges**, ranges from 0–13, $M = 0.09$, $SD = 0.50$.

- **Degree of the charge**, classified as felony (3970) or misdemeanour (2202).
- **Violence risk score**, ranges from 1–10 ($M = 3.64$, $SD = 2.49$), as generated by COMPAS.
- **Recidivism risk score**, ranges from 1–10 ($M = 4.42$, $SD = 2.84$), as generated by COMPAS.

The used dataset contains one more parameter, 'charge description', which was excluded due to highly unstructured and missing data. Furthermore, we generate two 'explicitly verifiable' predictors. These are two fabricated plots of variables which were not included in the original dataset;

- **Dominant hand**, showing a clear tendency towards higher recidivism for left-handed defendants.
- **Height**, showing a clear tendency towards higher recidivism for taller defendants.

See Fig. 1 for an overview of the plots as shown to our participants. In their design recommendations for crowdsourcing tasks, Kittur et al. [29] state the importance of including explicitly verifiable questions. Not only do these questions assess the motivation and capabilities of participants to answer questions correctly, they can be used to "*signal to users that their answers will be scrutinized, which may play a role in both reducing invalid responses and increasing time-on-task*" [29]. Given the complexity of our task, we wanted to ensure that our participants actually understand the task at hand. We followed design recommendations from Tufte [49] in the creation of our graphs to increase understandability for untrained readers (*e.g.*, consistent y-axis range from 0-100%).

3.2 Task Design

Upon accepting the task, participants were shown a detailed explanation of the study. Our explanation described the task of assessing decision-making models and included an example scenario on the chance of being involved in a car accident in relation to the driver's age. The example scenario consisted of explanatory writing as well as a plot identical in style to the ones used in the actual study. The different elements of the plot were described in detail, ensuring that all participants understood how to interpret the visuals that would be presented to them during the study. Subsequently, participants joined the Slack environment and our script automatically assigned the participant to a channel (*i.e.*, chatroom). A random pseudonym identified each participant (*e.g.*, 'Turker56'). After sufficient participants joined the channel (either 1 or 3, depending on the condition), the bot explained the task to participants. The bot offered a non-technical explanation of statistical models, and explained that we have developed a statistical model predicting recidivism across a number of predictors as based on an actual recidivism dataset. The participants' task was to indicate, for each predictor, whether they believe it should be included or excluded from the model. We fine-tuned the bot's script over multiple pilots to ensure it could be understood. Furthermore, we ensured that the script did not recommend parameters for removal, so as to avoid biasing participants.

We provide a visual summary of the interaction process in Figure 2. The bot presents each individual predictor, one by one and in randomised order, to the participants. For each predictor it shows a graph of the effect of the particular predictor on the chance of recidivism. At that point, participants are asked to indicate whether they believe the predictor is likely to introduce unwanted bias into the model and should therefore be removed. The voting outcome is not visible to other members of the channel (if in a group condition). An example of this task is shown in Figure 3.

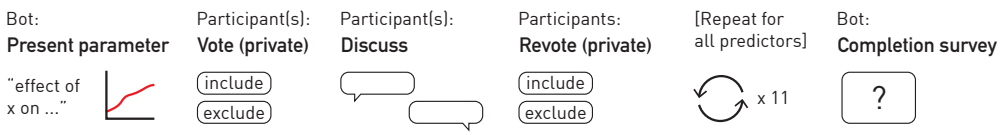

Fig. 2. Summary of task workflow.

After all participants cast their vote, the bot initialises a round of discussion in which the participants are asked to motivate their choice. Our intention is to initiate unstructured and interactive discussion between participants. For the individual worker condition, each participant simply types a motivation for their vote. The bot proceeds when participants indicate that they wish to proceed. For the 'structured group' condition, the (randomly appointed) group leader was responsible for progressing to the next sub-task – other group members were unable to proceed on their own. For the 'unstructured group' condition, a signal was required from all group members before the bot would proceed to the next sub-task.

For both group conditions, if the vote was not unanimous, we once again brought the predictor to a vote. We include this extra step to assess whether discussion is able to sway participant opinion. Regardless of the outcome of this second vote, participants then proceed to the next predictor. The individual condition never included a 'revote'. By collaborating within a small group, we encouraged participants to externalise their ideas and motivations. We recorded all group conversations for further analysis.

Following the completion of all tasks, participants completed a short demographic questionnaire. The demographic questionnaire asked participants for their age, gender (text field), race (multiple choice – single answer, as operated in Table 2), and family income. We included family income as a question due to its strong association with education, criminality, and social capital [9] – and is therefore likely to affect thoughts on recidivism.

3.3 Participants & Reward

We recruited 90 participants through MTurk, and distributed them equally across the three conditions (see Table 1). We limited our tasks to crowdworkers from the United States. Although allowing crowdworkers from a wider range of countries would result in a more diverse sample, the fact that the recidivism dataset was collected in the U.S. makes the task more relatable to this population.

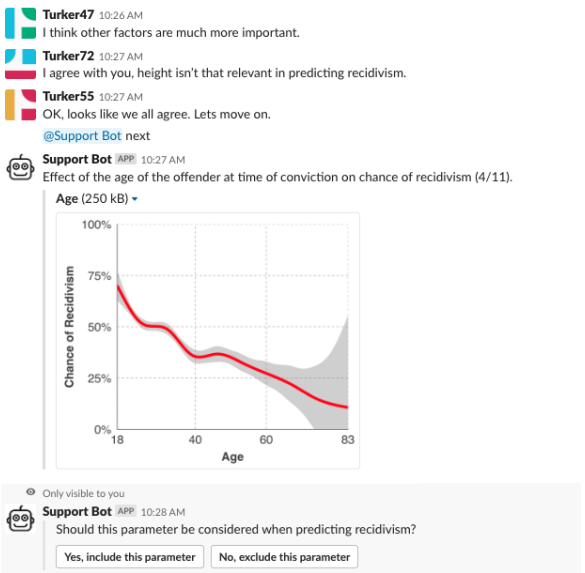

Fig. 3. Interaction between participants and 'Support Bot'.
Participant votes are not revealed to other participants. Trend line marked in red, confidence interval in grey.

Furthermore, we impose a minimum limit of 1000 completed tasks at a >95% acceptance rate for workers to be eligible – ensuring only serious workers would be recruited for the tasks.

Participants received a set amount of money for the completion of a task set (*i.e.*, answering all 11 predictors and the completion survey). The payment of workers involved in research projects is a heavily debated topic [45]. We followed the highest state-wide minimum wage in the US ($11.50/hour at time of our study). With an expected completion time of 20 minutes, we round up to a $4 compensation per task. Each participant could participate only once in our study.

Table 1. Number of participants per condition

Condition	N	Groups	Nr. of votes
Individual	30	-	330
Group - structured	30	10	330+
Group - unstructured	30	10	330+

Number of votes based on 11 predictors. Group conditions introduce an unknown number of re-votes.

4 RESULTS

We analysed participant voting behaviour across the three conditions, changes in voting following discussion, and the effect of group diversity on voting. Following this we present a qualitative analysis of chat transcripts using a general inductive approach [48], allowing us to derive themes as informed by the quantitative results. The qualitative analysis was completed independently by two authors, after which they agreed on the final themes in collaboration with a third author.

A total of 90 participants took part in the study (45 female, 45 male), equally distributed over the three conditions (Table 1). The age of participants ranged from 20 to 62 ($M = 35.01$, $SD = 9.84$). In terms of self-reported participant race, we note 70 White, 12 Black/African-American, 3 Asian, 3 Hispanic, and 2 Native American. These demographics are similar to the demographic distribution of the US population [52], as shown in Table 2. Average completion time of the study tasks was 23.7 minutes, with averages per condition at 19.5, 25.0, and 26.7 minutes for the Individual, Group - Structured, and Group - Unstructured condition respectively.

4.1 Voting Behaviour

Our participants cast 1212 votes, of which 222 were revotes triggered following a lack of consensus on the initial vote in group conditions. Overall, participants were in strong favour of including the defendant's number of prior charges as well as the defendant's age (respectively 93.3% and 86.7%), with dominant hand being the predictor least likely to be included (13.3%). The vote distribution for all predictors is shown in Figure 4. Shifts in voting patterns between initial vote and revote (applicable only to group conditions) are shown using arrows. Through a majority vote analysis, we found that the two explicitly verifiable predictors, 'Height' and 'Dominant hand', are most frequently identified for exclusion in all conditions. Furthermore, we found that the predictor 'Race' is identified for exclusion in both group conditions.

Although the overall pattern of inclusion of predictors is relatively similar across conditions (Figure 4), we note that voting patterns differ considerably for a number of variables. This can have significant consequences for borderline predictors (*e.g.*, 'Nr. juvenile misdemeanor charges', 'Race'). Since there is no direct ground truth against which to evaluate condition voting behaviour, we first calculated for each chat channel (30 Individual, 10 Group - Structured, 10 Group - Unstructured)

Table 2. Demographic information of our participants as compared to the 2017 U.S. Census [52]

Demographic attribute	Sample	U.S. Census
Male	50.0%	49.2%
Female	50.0%	50.8%
Age (median)	34	38
White	77.8%	76.6%
Black/African-American	13.3%	13.4%
Asian	3.3%	5.8%
Hispanic	3.3%	2.7%*
Native American	2.2%	1.5%

*Hispanic information is currently not recorded as a separate race, this is likely to change in future census collection [51]. Here we refer to the 'Two or more races' category of the census.

its agreement with the study-wide majority vote. As such, all members of a group would have to exclude 'Race', 'Height', and 'Dominant hand', and include all remaining predictors to obtain 100% agreement with the majority vote.

The average agreement among the 50 chat channels is 77.1% (SD = 16.6%). As the data is not normally distributed, we performed a Kruskal-Wallis rank sum test (non-parametric equivalent of one-way ANOVA) and found a significant effect of condition on agreement with the majority vote ($H(2) = 6.62$, $p = 0.037$). A post-hoc test using Wilcoxon rank-sum test with Bonferroni correction showed significant differences between the Individual and Group - Structured condition ($p = 0.012$, $r = 0.35$), but not between the Individual and Group - Unstructured condition ($p = 0.174$, $r = 0.19$) or Group - Structured and Group - Unstructured condition ($p = 0.440$, $r = 0.11$). Mean agreement with the majority vote was 72.7%, 86.4%, and 80.9% for the Individual, Group - Structured, and Group - Unstructured condition respectively. Our results indicate that those in the Group - Structured condition were significantly more in agreement with the majority vote than those in the Individual condition. We do not find a difference in time spent in-between predictor presentation and voting (considering initial vote) when including (M = 21.80 seconds) vs. excluding (M = 20.68 seconds) a predictor ($t(554.58) = 0.45$, $p = 0.66$). Similarly, we find no interaction between voting behaviour and reaction time for the revote scenario (relevant only for the Group conditions) ($t(217.82) = 0.38$, $p = 0.71$).

We analysed the possible effect of demographic features on voting behaviour (considering final votes). We were mostly interested in the effect of a participant's demographic feature on the corresponding predictor, *e.g.* does gender affect voting behaviour on including sex as a predictor? First, we found no significant effect of participant gender on considering sex as a suitable predictor for recidivism ($\chi^2(1, N = 990) = 0.81$, $p = 0.367$). Second, we find no significant effect of participant's self-identified race on including race as a predictor ($p = 0.522$, Fisher's exact test). Given the distribution of participant race we utilised a Fisher's exact test rather than a Chi-Square test of Independence. Third, we investigate the effect of participant age on voting behaviour on the age predictor. Using one-way ANOVA, we did not find a significant effect of participant age on including age as a predictor ($F(1,88) = 0.21$, $p = 0.643$). These results indicate that voting behaviour on demographic predictors was not affected by participant demographics.

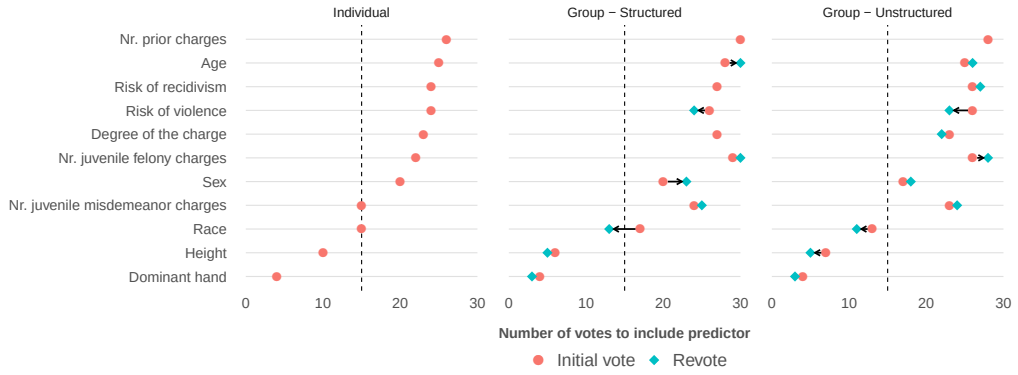

Fig. 4. Participant voting pattern for each predictor. More than 15 votes in favour are considered a majority.

4.2 Swaying Votes

We separately analysed the 222 revotes. We triggered a revote following discussion in cases where the initial vote was not unanimous. A total of 96 and 126 revotes were collected in the Structured and Unstructured group condition respectively. Following discussion with group members, participants changed their votes as compared to their original vote in 48 instances (21.6%, considering changes in either direction). A Chi-Square test with Yates' continuity correction revealed that the likelihood of a participant changing their vote differed significantly by condition, $\chi^2(1,N=222)=6.49$, $p=0.011$, with an odds ratio of 2.44. For the structured condition 30.2% of revotes resulted in a change of vote, compared to 15.1% in the unstructured condition.

Considering the effect of revoting on final voting outcome within groups, we identified 25 cases (33.8%) in which two group members convinced the remaining group member, 5 cases (6.8%) in which one group member convinced the other two members, 9 cases (12.2%) in which the majority opinion within the group shifted without reaching consensus, and 35 cases (47.3%) in which the majority vote remained the same but no consensus was reached.

We test for the effect of participant gender on a change in vote between the two group conditions. A Chi-Square test with Yates' continuity correction revealed that changes in vote significantly differed by gender for the structured conditions ($\chi^2(1,N=96)=4.03$, $p=0.045$, odds ratio 2.87), but not for the unstructured condition ($\chi^2(1,N=126)=0.58$, $p=0.447$, odds ratio 0.60). For the structured condition, women changed their vote in 39.6% of revotes, compared to men who changed their vote in 18.6% of revotes. For the unstructured condition, women and men changed their votes in 12.3% and 18.9% of revotes respectively. A Chi-Square test indicates no effect of the group leader's gender on female participants' likelihood to change vote ($\chi^2(1,N=53)=0.02$, $p=0.894$).

4.3 Group Composition

We investigated the effect of group composition on voting outcomes. We measured dissimilarity between group members using the survey results of participants' age, gender, race, and income. Given the categorical nature of some of these variables, we utilise Gower's distance [18] for dissimilarity calculation. Gower's distance applies a standardisation to each variable, and considers the distance between two units of one variable as the sum of all the variable-specific distances. For example, a group consisting of three women of identical race and similar age and income range will result in a low dissimilarity score. Dissimilarity scores in our sample range from 0.26 to 0.63, with an average dissimilarity of 0.50 (SD = 0.09). Figure 5 shows the average dissimilarity score of each of the 20

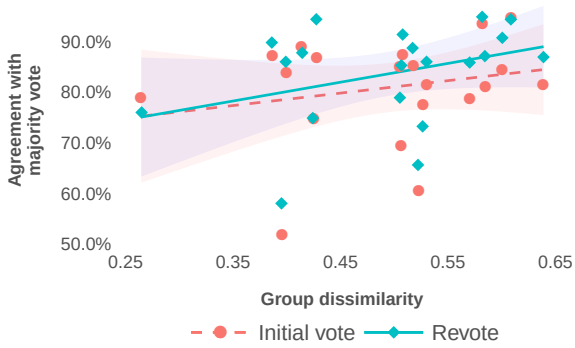

Fig. 5. Group diversity and agreement with majority vote. Coloured bands indicate 95% confidence intervals.

groups plotted against their agreement with the majority vote. We found that the votes of diverse groups tended to agree more with the majority vote. Furthermore, Figure 5 shows that this effect increases following discussion (*i.e.*, revotes).

4.4 Participant Discussion

We considered the results of our chat analysis in relation to the following themes: participant motivation for their voting behaviour, convincing group members to sway votes, and rationalisation for the 'risk of recidivism' and 'risk of violence' predictors. We include these themes for our inductive approach following our quantitative analysis of results. Transcripts were augmented with the participants voting behaviour and were analysed by two of the authors.

4.4.1 Motivating Voting Behaviour. We found that participants motivated their choices using a combination of arguments. These arguments consist of ethical considerations, technical concerns, and an appeal to (their interpretation of) the graphs. For example, P104 described his consideration on including juvenile charges in the prediction of recidivism, citing both ethical and technical concerns.

P104: *I feel like this is relevant, but I also feel very uncomfortable having crimes stick from juvy in any way. You're not supposed to be really culpable at that age.*

Similarly, P44 discussed his consideration of including sex as a predictor and compared it to previously assessed (demographic) features.

P44: *I think there is probably a big difference in the rate of which men and women offend. Hmm this one is hard. I know we excluded race and handedness, but I think sex is relevant.*

Participants also referred to the graphs to motivate their arguments, occasionally deflecting potential criticism on their voting behaviour to the 'authority' provided by the graph. For example, while discussing race as a predictor;

P16: *I really hate to include this one, but the chart doesn't lie.*

Similarly, participants refer to specific elements of the graphs to explain their voting behaviour. For example, several participants decided to exclude a parameter given the high confidence interval shown in the graph;

P160: *I'm mixed on this, it seems intuitively like it would be a useful parameter, but the amount of error is so high at points I feel uncomfortable including it.*

4.4.2 Convincing Group Members. Our results show differences in voting patterns between conditions (Figure 4). Although participant motivations were typically short statements, they were

sometimes sufficient to sway group members. It appears that offering an additional moment of reflection on an opposing viewpoint was sometimes sufficient to change the opinion of participants. For example, in the following discussion P151 convinces her group members to switch their vote:

P151: *I don't think race inherently makes someone more likely to re-offend and the data doesn't seem particularly convincing, so exclude in my opinion.*
P160: *Yes race really shouldn't be a factor.*
P188: *I did vote yes but once I look at it, race shouldn't be a factor.*
P160: *Yes me too!*

Other participants raised technical concerns to convince their group members. The following discussion concerns the defendant's 'degree of charge', in which P109 convinces the other group members to change their vote based on state differences in classifying charges:

P109: *A DUI* [driving under the influence] *in other states is a misdemeanor. So, say two males with identical backgrounds in two different states commit the same crime (DUI) one state its felony, other mis. Is one really more at risk?*
P103: *That's a good point.*
P104: *Ah. I didn't know that. That complicates things. That is a really good point, though, yea. Severity of crimes is measured in different ways depending on where you go. Now I'm a lot less comfortable about including this as is.*
P109: *I do get patterns of behavior and generally felonies are more severe. I can agree that it seems more likely a felon would re-offend, but I have to play devil's advocate haha.*
P104: *No, I think you're right. I think this data is relevant but it's far too general to be useful. We need a more zoomed-in, granular look at the trend to use it I think.*

4.4.3 Risk of Violence/Recidivism Scores. We investigated participant motivation in their decision making for the defendants' violence and recidivism risk scores. These predictors show high correlation with defendants' recidivism, but are arguably unclear in their origin. The description offered to participants did not provide any specifics and simply stated "*A numeric value* [...] *corresponding to the [recidivism/violence] risk score of the defendant*". As such, we are interested in the rationalisation and justification offered by participants. We observed that both proponents and opponents of these predictors wanted more information on these values:

P127: *I'm still curious where the 'risk scores' are coming from and how they are being calculated. In this case, I guess it would make sense to calculate based on previous amount of violence, but I still wish there was more information.*

Among those voting to exclude these predictors, participants commonly stated that the lack of information concerning the collection of risk scores affected their voting behaviour, expressing the need for an 'objective' score based on understandable and 'valid' parameters.

P70: *I can't support it because it doesn't explain what it factors in.*
P44: *As long as the data used to determine this risk score is valid, I can see this being useful. Yes, not knowing how this score is determined makes this one a little tricky.*
P67: *True; it's unclear what the parameters of the risk score are. I'm undecided to be honest.*
P44: *Yes, at this point I would say no, with the available information.*

We note that a small number of participants made up their own interpretation of a predictor;

P34: *This is an assessment done by professionals to determine the likelihood of recidivism. Pretty scientific I think.*

Finally, we contrast the participants' statements with their *actual* voting behaviour. We were interested to see whether self-presentation would cause participants to express themselves differently than how they would cast their votes (a process hidden from the other team members). Our results do not indicate any discrepancy between participants' expressed opinion/voting behaviour and their actual behaviour. This might be explained by the anonymous nature of the study design, leading participants to openly share their viewpoints and reveal their (changes in) voting behaviour;

P188: *I did vote yes but once I look at it, race shouldn't be a factor.*

5 DISCUSSION

Conducting research on algorithmic fairness is challenging due to the subjective nature of 'fairness'. Our study does not set out to determine the degree of fairness of using particular predictors in a recidivism model. Rather, our work aims to understand how people subjectively assess the perceived fairness of a predictor, and we develop mechanisms to facilitate this process. Furthermore, as there is arguably no ground truth, an important measure is the 'social acceptability' of an algorithm. This is something that we can capture by adopting the notion of majority voting and considering the most popular opinion as the most socially acceptable one.

Therefore, in our study we set out to understand how people make decisions, identify suitable configurations, and develop automated mechanisms to expedite this process. There are two key elements of our work that we wish to highlight. First, the bot mechanism we have developed offers a number of benefits in collecting opinion data as required for fairness assessment. To support future research in this field, we have published the source code of our bot under the MIT license[2].

Second, our use of visual evidence (graphs), which reflect the measurable impact of predictors, has been key in the outcomes of this study. Our transcript analysis shows that these graphs helped participants to form opinions in light of available data and was actively used to motivate decision making. This is an important advance over previous work that solely considers the 'name' or description of the predictor [20], and relies solely on individuals' prior experience to make a judgement call. Additionally, this visual evidence allows workers to interrogate the dataset, and identify latent biases in the data itself that may or may not align with workers' own bias. For example, previous work has identified racial bias among police officers in stop-and-search [62], and therefore datasets may be biased in unpredictable ways. Our findings suggest that exposing participants to diverse opinions, and possibly conflicting biases, increases the likelihood that the opinion formed by a small group agrees with the rest of the population. Although this does not suggest that this opinion is 'correct', it does mean that it may be more broadly accepted. Dietvorst et al. [12] show that allowing users to modify an algorithm increases their trust in the algorithm. Similarly, feedback and modification of algorithms by the crowd can lead to a wider acceptance by a larger population.

To inject some measure of correctness in our findings, we introduce two fake 'verifiable' predictors (height, dominant hand), and show that participants are able to identify and exclude them. This outcome gives us confidence regarding the overall process, as it suggests that participants had a reasonable understanding of the task given that both of these predictors were largely excluded.

5.1 Algorithmic Decision Making

Previous work on algorithmic decision making has explored the use of surveys [43], face-to-face interviews [36, 55, 60], and systematic user studies [6]. We propose a method that differs in two key ways. First, we consider participants working individually, as well as in anonymous groups. This provides the opportunity for discussion and reflection, and therefore the chance for workers to change their mind. Second, to elicit their judgement of predictors and their effect on algorithmic fairness, we present participants with visual evidence: graphs generated by an intelligible model.

Our results indicate that participants in the Group conditions display a higher agreement with the majority vote, pointing to a positive effect of group discussion. This effect was enhanced in more diverse groups (Figure 5). The lack of diversity in teams constructing algorithms has been raised as a reason for algorithmic unfairness [60]. It can be argued that crowdsourcing is a means to obtain input from diverse crowds, and our work provides some evidence that more diverse groups tend to more closely align with the majority.

[2]Please see the paper's auxiliary materials or https://github.com/nielsvanberkel/SupportBot.

Besides reaching a more diverse crowd, our work demonstrates mechanisms to develop what Veale et al. call "*workable social and technical improvements*" to study politically charged settings [55]. An important advantage of our proposed mechanism is that it is automated and scalable, and can therefore be re-used extensively.

Previous work has explored how different types of textual explanations affect participant's perception of algorithmic justice [6]. The use of visual evidence (*i.e.*, plots) to explain the effect of a variable on our outcome variable (recidivism) is another type of explanation. Through these plots, we empower participants to decide on the trade-off between the potential 'added value' of the predictor and its consideration for inclusion. The graphs generated by our intelligible GAM were well understood by our participants, as evidenced by our transcript analysis and the exclusion of the 'verifiable' predictors. There are, however, alternative methods for displaying data, such as GA2M [8]. GA2M allows for analysis of the interaction between two predictors and one outcome variable. However, GA2M graphs are more complex to interpret and as this is an initial investigation of utilising ML biases in crowdsourcing, we decided to utilise the more established and easier to understand GAM. Another alternative is the use of *i.e.* decision trees (*e.g.*, Random Forest). We decided against the use of such models as they are generally less intelligible: a single factor can appear in multiple trees, making it difficult to assess the effect of an individual predictor.

The work presented here offers AI developers and researchers a practical tool to use in their algorithm development workflow. The perceived unfairness of specific predictors can be used to exclude predictors or identify biases in the dataset. These biases may not necessarily be apparent to those developing the AI, for example due to a lack of domain expertise, diverging social backgrounds, or personal predisposition. It is important to note that excluding one variable (*e.g.*, race) does not necessarily result in a more fair algorithm. A commonly used example in the US context is the relation between race and neighbourhood – in which a seemingly impartial variable (postal code) contains a racial indicator. However, recent work has explored novel algorithms to overcome this problem. Using the same recidivism dataset as used in this study, Friedler et al. show that fairness-enhancing interventions (*i.e.*, algorithms specifically developed to reduce bias) can be successfully instructed to account for protected attributes (*e.g.*, race) [15]. Hence, the methodology proposed in our paper can be used to identify the attributes to be assigned a protected status when employing such fairness-enhancing interventions.

5.2 Crowdsourcing Decision Making

Our work has explored the use of crowdsourcing not from the perspective of contributing new data, but instead on assessing existing model predictors [30]. In an analysis of 19 risk assessment instruments (including the COMPAS dataset used in our work), Desmarais & Singh find that *"[in] most cases, validity had only been examined in one or two studies conducted in the United States, and frequently, those investigations were completed by the same people who developed the instrument."* [11]. We argue that our approach has the potential to support the identification of biases in future machine learning endeavours by surveying the crowd about the perceived fairness of using certain information (*e.g.*, a person's race or age) in a decision making model.

This aligns with Kittur et al.'s vision on the 'Future of Crowd Work', in which crowds can be employed to guide algorithmic decision making [30]. In addition, Hutchinson et al.'s review of 50 years of assessment test fairness work identifies lessons for Machine Learning, pinpointing the public's concerns and perception of fairness as an area which requires careful attention [25]. In particular, the authors warn of a divide between the public perception and (abstract) technical definitions; "*If technical definitions of fairness stray too far from the public's perceptions of fairness, then the political will to use scientific contributions in advance of public policy may be difficult to obtain*" [25]. The combined geographic and demographic scale of online crowdsourcing markets (see Table 2) allows researchers to obtain perceptions on algorithmic fairness from a diverse audience.

Despite the fact that our results are promising, we note that the use of (untrained) crowdworkers does introduce several challenges. First, participants may not completely understand what they aim to achieve. This could lead to skewing of results when forcing participants to make a (binary) decision. This became apparent in the participant's discussion around the predictors 'risk of recidivism' and 'risk of violence', in which participants indicated they were unsure on the construction of the predictor. As such, we urge future deployments to offer participants the option to indicate their understanding of the current activity.

Second, a critical element in the discussion of algorithmic decision making is the use of a diverse participant sample [60]. Although the MTurk crowd is not representative of the general population (*e.g.*, above average education, lower average income [38]), previous work points to the overall diversity of crowdworkers on MTurk [38, 42]. Work from 2012 reports on an over-representation of Asians and under-representation of Blacks on MTurk [4]. This is not reflected in our sample and may be indicative of a shift in MTurk crowdworkers. In fact, the demographics of our sample are closely aligned to the larger U.S. population (Table 2). Alternatively, U.S. college students, a common source of study participants, consist of significantly more females and fewer non-Whites as compared to the overall population [16]. As such, we argue that the use of crowdworkers is an improvement over the recruitment of 'typical' college students. Levay et al. [38] report on techniques which can be used to further balance the study sample, for example by screening for political affiliation prior to the task. Furthermore, although the presented work analyses the topic of recidivism from a general perspective – specific topics may warrant the use of a specific or specialised crowd due to their unique experiences, insights, or likelihood to be either end-users or target-users of the developed AI technology.

5.3 Group Diversity

Our results indicate that more diverse groups, as based on participant's age, gender, race, and income, reached higher levels of agreement with the majority vote following discussion (Figure 5). Previous work on the relationship between group diversity and group performance has revealed inconsistent results [54, 58]. This discrepancy in the literature boils down to a difference in academic perspective. From a social categorisation perspective, people typically display increased trust and willingness to cooperate with 'ingroup' members over 'outgroup' members [54]. In contrast to this, research from a decision-making perspective has emphasised that group diversity can reveal differences in knowledge and perspective, leading to higher quality group work [54].

In our study, participants were unable to identify social category differences (*e.g.*, sex, age, race) or functional differences (*e.g.*, educational background) due to the study's anonymous setup. Based on the existing Psychology literature and the results from our study, we posit that the anonymity of our participants allowed diverse groups to obtain more novel perspectives while masking the unconscious biases typically encountered in a diverse group (social categorisation). Future work should explore more systematically which characteristics affect effective collaboration in an (anonymous) chatroom setting, and how this information can be utilised in task routing.

5.4 Bots as a Research Instrument

Researchers employ qualitative methods such as focus groups and interviews to elicit thoughts or opinions from participants. In this work we propose the use of interactive bots as a research instrument for the collection of both qualitative and quantitative data. Messaging bots, or 'botplicatons' [31], allow for increased productivity through streamlining, automating, and synchronising group efforts [31, 34].

Our study has enabled us to reflect on the use of bots as a research instrument. The 'interviewer effect', a widely reported interview bias, describes the distortion of participant responses based on interviewer characteristics. Previous work found that on race related questions, respondents "*report in ways that might be perceived more positively by persons of the interviewer's race or ethnicity*" [10].

Similarly, older interviewers obtain better cooperation from their interviewees (independent of interview experience) [46]. The use of a mixed and diverse group of interviewers has been recommended in socially sensitive topics, in which participants are prone to present socially desirable responses [10]. This diversity, however, cannot always be obtained in a realistic research setting.

We therefore argue that for sensitive and controversial topics, the use of bots as a primary facilitator can be beneficial. Our results show that participants are willing to reveal highly personal (criminal) experiences. The use of a bot as a research instrument ensures a structured and consistent interaction across participants, reducing experimenter bias. This behaviour scales to large samples and allows for novel interactions. For example, we were able to collect *private* voting information within the context of a group conversation – reducing the likelihood of conformity bias. Furthermore, as experimental conversations are highly structured around the task itself, we avoid the large gap between user expectation and experience often observed in the interaction with Conversational Agents [39]. Therefore, while bots may not be mature enough to have meaningful conversations, we find that given the constraints and scripted nature of our study, the bot was an asset rather than an obstacle.

Naturally, there are certain limitations that bots impose. First, it is typical for human interviewers to 'dive deeper' into relevant aspects brought up by the participant. Bots, on the other hand, are mostly restricted to pre-defined conversational patterns. This limits the potential richness of the interview data. Second, a bot is unable to pick up on (sudden) cues of the interviewee (*e.g.*, stress) and is therefore unable to appropriately steer the conversation when required. In a future iteration of our bot we will add mechanisms to ping and subsequently remove unresponsive participants.

5.5 Limitations & Future Work

This paper investigates the perceptions of fair predictors through the use of explainable models. To ensure that the generated model's graphs were understandable, we considered only a single parameter at a time. As such, we were unable to account for (hidden) interactions between parameters. An evaluation of the effect of interaction effects on perceived fairness among participants is a fruitful avenue for future work. However, we urge researchers to carefully consider the visual presentation of these interactions given the steeply increasing difficulty of interpreting multi-way interaction effects. Furthermore, we limit our sample to the U.S. crowdsourcing population given the characteristics of the dataset and to ensure a high level of English language skill. We expect that opinions on what constitutes a fair predictor differ between countries. Our results do therefore not generalise outside of our study sample. Future work should explore geographical/cultural differences in fair predictor selection, and consider collaboration between these groups.

Our results on group composition (*e.g.*, Figure 5) follow directly from the collected demographic information ('age', 'gender', 'race', and 'family income'). Future work could consider additional demographic factors or different ratios of these factors in their dissimilarity calculations. Furthermore, we note that while these factors are relevant in the context of recidivism, future work on algorithmic fairness should determine the sample's diversity through factors relevant to the application at hand. While our detailed explanation ensured participants had a shared understanding of the presented graphs, we did not control for domain expertise in statistics or visualisation. These differences in expertise may result in a confounding variable during group discussions. We recommend future work to follow our directive of including sufficient training materials with examples, and to consider the use of a short test to evaluate participants' ability to correctly interpret the presented visualisations.

The current work introduces a bot with neutral characteristics and low levels of expression. Future work can explore the effects of bots which mimic or oppose study participants' characteristics (*e.g.*, gender, social conformity), or use Natural Language Processing to allow for a more interactive conversation. Lastly, we note that coding our bot was far from trivial. Available bot toolkits are typically not developed with researchers in mind, requiring extensive customisation.

6 CONCLUSION

Emerging HCI research on algorithmic fairness has pointed to the importance of research methods that are scalable [20], practical [55], and able to ensure a diverse sample [60]. This work investigates the feasibility of utilising crowdsourcing for fair predictor assessment in machine learning. Our results indicate that our participants were able to make an informed decision using graphs, as indicated by the explicitly verifiable questions and analysed chat logs. We found higher performance following discussion among diverse groups as compared to uniform groups. Our study provides a look into the use of crowdworkers in discussions concerning the critical area of algorithmic fairness. It offers a structured and scalable approach, based on interactive (group) conversations led by a bot. Future work may expand this work into other areas of algorithmic fairness outside of predictor selection, and offer a more systematic assessment on the effect of group diversity.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. Retrieved June 14, 2018 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Jonas Auda, Dominik Weber, Alexandra Voit, and Stefan Schneegass. 2018. Understanding User Preferences Towards Rule-based Notification Deferral. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article LBW584, 6 pages. https://doi.org/10.1145/3170427.3188688

[4] Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20, 3 (2012), 351–368. https://doi.org/10.1093/pan/mpr057

[5] Philippe Besnard and Anthony Hunter. 2008. *Elements of Argumentation*. The MIT Press. 298 pages.

[6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 377, 14 pages. https://doi.org/10.1145/3173574.3173951

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (01 Sep 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[8] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[9] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. *Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States*. Working Paper 19843. National Bureau of Economic Research. https://doi.org/10.3386/w19843

[10] R. E. Davis, M. P. Couper, N. K. Janz, C. H. Caldwell, and K. Resnicow. 2010. Interviewer effects in public health surveys. *Health Education Research* 25, 1 (2010), 14–26. https://doi.org/10.1093/her/cyp046

[11] Sarah Desmarais and Jay Singh. 2013. Risk assessment instruments validated and implemented in correctional settings in the United States. *Lexington: Council of State Governments* (2013).

[12] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64 (2018), 1155–1170. Issue 3. https://doi.org/10.1287/mnsc.2016.2643

[13] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018). https://doi.org/10.1126/sciadv.aao5580

[14] 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 (ITCS '12)*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[15] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 329–338.

https://doi.org/10.1145/3287560.3287589
[16] Antoine M. Garibaldi. 2014. The Expanding Gender and Racial Gap in American Higher Education. *The Journal of Negro Education* 83, 3 (2014), 371–384. https://doi.org/10.7709/jnegroeducation.83.3.0371
[17] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *CoRR* (2018). http://arxiv.org/abs/1806.00069
[18] John C. Gower. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27, 4 (1971), 857–871.
[19] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563
[20] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 903–912. https://doi.org/10.1145/3178876.3186138
[21] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA*.
[22] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 3323–3331. http://dl.acm.org/citation.cfm?id=3157382.3157469
[23] Trevor J. Hastie and Robert J. Tibshirani. 1990. *Generalized Additive Models*. Monographs on Statistics and Applied Probability, Vol. 43. Chapman & Hall, London. 352 pages.
[24] Simo Johannes Hosio, Jaro Karppinen, Esa-Pekka Takala, Jani Takatalo, Jorge Goncalves, Niels van Berkel, Shin'ichi Konomi, and Vassilis Kostakos. 2018. Crowdsourcing Treatments for Low Back Pain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 276, 12 pages. https://doi.org/10.1145/3173574.3173850
[25] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600
[26] Ellora Thadaney Israni. 2017. When an Algorithm Helps Send You to Prison. Retrieved June 2, 2018 from https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html
[27] William A. Kahn. 1990. Psychological Conditions of Personal Engagement and Disengagement at Work. *Academy of Management Journal* 33, 4 (1990), 692–724. https://doi.org/10.5465/256287
[28] Juho Kim, Eun-Young Ko, Jonghyuk Jung, Chang Won Lee, Nam Wook Kim, and Jihee Kim. 2015. Factful: Engaging Taxpayers in the Public Discussion of a Government Budget. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2843–2852. https://doi.org/10.1145/2702123.2702352
[29] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127
[30] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923
[31] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 555–565. https://doi.org/10.1145/3064663.3064672
[32] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting Reflective Public Thought with ConsiderIt. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 265–274. https://doi.org/10.1145/2145204.2145249
[33] Todd Kulesza, Weng-Keen Wong, Simone Stumpf, Stephen Perona, Rachel White, Margaret M. Burnett, Ian Oberst, and Andrew J. Ko. 2009. Fixing the Program My Computer Learned: Barriers for End Users, Challenges for the Machine. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*. ACM, New York, NY, USA, 187–196. https://doi.org/10.1145/1502650.1502678
[34] M. Lee, L.E. Frank, F. Beute, Y.A.W. de Kort, and W.A. IJsselsteijn. 2017. Bots mind the social-technical gap. In *Proceedings of 15th European Conference on Computer-Supported Cooperative Work, 28 August - 1 September 2017, Sheffield, United Kingdom (Reports of the European Society for Socially Embedded Technologies)*. European Society for Socially Embedded Technologies (EUSSET), 35–54. https://doi.org/10.18420/ecscw2017-14
[35] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on*

Computer Supported Cooperative Work and Social Computing (CSCW '17). ACM, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230

[36] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management That Allocates Donations to Non-Profit Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3365–3376. https://doi.org/10.1145/3025453.3025884

[37] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4 (01 Dec 2018), 611–627. https://doi.org/10.1007/s13347-017-0279-x

[38] Kevin E. Levay, Jeremy Freese, and James N. Druckman. 2016. The Demographic and Political Composition of Mechanical Turk Samples. *SAGE Open* 6, 1 (2016), 1–17. https://doi.org/10.1177/2158244016636433

[39] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[40] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artifical Intelligence* 267 (February 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[41] Ritesh Noothigattu, Snehalkumar (Neil) S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. 2017. A Voting-Based System for Ethical Decision Making. *CoRR* abs/1709.06692 (2017). http://arxiv.org/abs/1709.06692

[42] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188. https://doi.org/10.1177/0963721414531598

[43] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations As Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 103, 13 pages. https://doi.org/10.1145/3173574.3173677

[44] Niloufar Salehi, Andrew McCabe, Melissa Valentine, and Michael Bernstein. 2017. Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1700–1713. https://doi.org/10.1145/2998181.2998300

[45] M. Six Silberman, Lilly Irani, and Joel Ross. 2010. Ethics and Tactics of Professional Crowdwork. *XRDS* 17, 2 (2010), 39–43. https://doi.org/10.1145/1869086.1869100

[46] Eleanor Singer, Martin R. Frankel, and Marc B. Glassman. 1983. The Effect of Interviewer Characteristics and Expectations on Response. *The Public Opinion Quarterly* 47, 1 (1983), 68–83. http://www.jstor.org/stable/2748706

[47] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting Meaningfully with Machine Learning Systems: Three Experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662. https://doi.org/10.1016/j.ijhcs.2009.03.004

[48] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. https://doi.org/10.1177/1098214005283748

[49] Edward R. Tufte. 1986. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, CT, USA. 200 pages.

[50] Sherry Turkle. 1995. *Life on the Screen: Identity in the Age of the Internet.* Simon & Schuster Trade. 352 pages.

[51] U.S. Census Bureau. 2017. Press Kit: 2015 National Content Test. Retrieved August 27, 2018 from https://www.census.gov/newsroom/press-kits/2017/nct.html

[52] U.S. Census Bureau. 2018. Annual Estimates of the Resident Population by Sex, Race Alone or in Combination, and Hispanic Origin for the United States, States, and Counties. U.S. Census Bureau, Population Division. https://www.census.gov/newsroom/press-kits/2018/estimates-characteristics.html

[53] Melissa A. Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S. Bernstein. 2017. Flash Organizations: Crowdsourcing Complex Work by Structuring Crowds As Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3523–3537. https://doi.org/10.1145/3025453.3025811

[54] Daan van Knippenberg and Michaéla C. Schippers. 2007. Work Group Diversity. *Annual Review of Psychology* 58, 1 (2007), 515–541. https://doi.org/10.1146/annurev.psych.58.110405.085546 PMID: 16903805.

[55] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 440, 14 pages. https://doi.org/10.1145/3173574.3174014

[56] Anthony W. Flores, Kristin Bechtel, and Christopher Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.". 80 (2016).

[57] M. Whitty and J. Gavin. 2001. Age/Sex/Location: Uncovering the Social Cues in the Development of Online Relationships. *Cyberpsychology & Behavior* 4, 5 (2001), 623–630.

[58] Katherine Y. Williams and Charles A. O'Reilly. 1998. Demography and Diversity in Organizations: A Review of 40 Years of Research. *Research in Organizational Behavior* 20 (1998), 77–140.

[59] S.N Wood. 2017. *Generalized Additive Models: An Introduction with R* (2 ed.). Chapman and Hall/CRC.

[60] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 656, 14 pages. https://doi.org/10.1145/3173574.3174230

[61] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1433–1444. https://doi.org/10.1145/2531602.2531604

[62] Naomi Zack. 2015. *White Privilege and Black Rights: The Injustice of US Police Racial Profiling and Homicide*. Rowman & Littlefield.

[63] Sharon Zhou, Melissa Valentine, and Michael S. Bernstein. 2018. In Search of the Dream Team: Temporally Constrained Multi-Armed Bandits for Identifying Effective Team Structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 108, 13 pages. https://doi.org/10.1145/3173574.3173682

[64] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (Nov. 2018), 23 pages. https://doi.org/10.1145/3274463

Received April 2019; revised June 2019; accepted August 2019