Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

Jorge Goncalves
jorge.goncalves@unimelb.edu.au
University of Melbourne
Melbourne, Australia

Daniel Russo
daniel.russo@cs.aau.dk
Aalborg University
Aalborg, Denmark

Simo Hosio
simo.hosio@oulu.fi
University of Oulu
Oulu, Finland

Mikael B. Skov
dubois@cs.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

The uptake of artificial intelligence-based applications raises concerns about the fairness and transparency of AI behaviour. Consequently, the Computer Science community calls for the involvement of the general public in the design and evaluation of AI systems. Assessing the fairness of individual predictors is an essential step in the development of equitable algorithms. In this study, we evaluate the effect of two common visualisation techniques (text-based and scatterplot) and the display of the outcome information (*i.e.*, ground-truth) on the perceived fairness of predictors. Our results from an online crowdsourcing study (N = 80) show that the chosen visualisation technique significantly alters people's fairness perception and that the presented scenario, as well as the participant's gender and past education, influence perceived fairness. Based on these results we draw recommendations for future work that seeks to involve non-experts in AI fairness evaluations.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative and social computing**; *Empirical studies in collaborative and social computing.*

KEYWORDS

Artificial intelligence, fairness, transparency, crowdsourcing, machine learning, predictor selection, layperson, visualisation, AI, ML.

ACM Reference Format:
Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445365

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445365

1 INTRODUCTION

Technological developments have led to an increase in the deployment of artificial intelligence (AI) based decision-support solutions. Having moved beyond theoretical solutions or theories, such systems now have a profound effect on both individuals and society at large [25, 57]. The use of AI spans across a variety of application domains, including high-impact areas such as loan approval [16, 36], policing and law enforcement [29, 31], and hiring processes [14, 46]. Although AI technology promises more efficient decision technology, the use of automated decision systems also raises questions and risks concerning their fairness and equity. Notable examples of these include the COMPAS recidivism risk prediction system, which was found to have racial biases and was deployed in several U.S. states [2, 29], and an AI-powered recruitment tool previously used at Amazon that showed bias against women [14].

The rising concerns regarding the fairness of AI systems have led to the development of novel and interactive visualisation tools to probe machine learning (ML) models. These interpretability tools allow ML practitioners to answer questions such as, '*what if [...]?*' [62], '*what is a feature's effect on [...]?*' [33], and '*why?*'/'*why not?*' [38]. Simultaneously, Human-Computer Interaction (HCI) literature argues strongly for the involvement of stakeholders and the general public in the development of AI applications [22, 56, 66], motivated by *e.g.* the ability to capture stakeholder insights [66] and to safeguard end-user acceptance [56]. Involving stakeholders in the early stages of algorithm design, *i.e.* during predictor selection, has been identified as critical due to the ability to avoid biases in initial design choices [66]. The aforementioned algorithm probing tools are primarily designed with ML professionals in mind, and it is unclear whether such tools can be appropriated by those without knowledge of the jargon and technical underpinnings of algorithm development. The involvement of members from the general public introduces new considerations to the way in which we study algorithmic fairness. These considerations have not been sufficiently addressed in the literature. Although several studies have explored ways to capture people's opinions and perceptions of AI systems [22, 56, 65], these studies have largely ignored the effect of information presentation (*i.e.*, including/excluding certain information, visualisation techniques) and are typically limited to one application scenario (*e.g.*, recidivism). In this work we argue

that *how* information is presented is at least as important as *who* it is presented to, and that these factors are interdependent.

Motivated by the increasing need to capture people's perspectives on AI fairness, and the lack of knowledge of the factors that affect the perceptions of fairness among non-ML experts, we designed an online crowdsourcing study to systematically investigate fairness perceptions in an algorithmic decision-making setting. In our study, participants were asked to make a judgement as to whether predictors should be included in an AI-algorithm in a given scenario. We assess the effects of predictor type (*e.g.*, demographic), presentation technique (text-based vs scatterplot), the inclusion of ground truth data in the information presentation, and application scenario on the participants' judgement. Further, we ask participants to report their perceived fairness, demographics, and overall assessment of the presentation technique. We set out to answer the following research questions:

- How does the presentation of data points affect perceived algorithmic fairness?
- What is the effect of presenting / not presenting the outcome of the predicted variable on perceived algorithmic fairness?
- How do the demographic factors of education and gender affect perceived algorithmic fairness?

We find that our participants were less likely to include a predictor when it was presented in a scatterplot as opposed to summarised in writing. Participants were more likely to include variables in a recidivism-related scenario as compared to a lending scenario, regardless of the predictor type. A significant interaction effect between predictor type and visualisation technique highlights that demographic predictors and domain-specific predictors are affected differently by the presentation technique. Furthermore, our results highlight that educational attainment is a significant factor in fairness perception, with a higher completed education leading to a lower average fairness rating. Similarly, we find that women perceived the predictors as significantly less fair as compared to men. Our results show that fairness perceptions are indeed altered by how information is presented, and that significant differences in fairness perceptions between demographic groups exist. This raises important considerations for the future involvement of participants in the design and evaluation of AI applications, such as how we display (interactive) data to participants. Finally, we present recommendations for future studies on AI perceptions.

2 RELATED WORK

Both the HCI and the AI literature highlight an increasing effort and interest towards the active involvement of the general public and/or end-users in data collection [27], training [17], and the evaluation [60] of AI applications. We motivate our work by building on prior research on fairness perceptions, tools for inspection of ML models, and visualisation and interpretation research.

2.1 Fairness and the Public

A variety of disciplines, including Law, Social Sciences, and Computer Science, have proposed numerous definitions and perspectives on fairness. Examples include, but are not limited to;

- *Individual fairness*, which states that similar individuals should be treated similarly [15].
- *Demographic parity / Group fairness*, in which outcomes are equally distributed among each defined group [9, 51].
- *Equality of opportunity*, which states that individuals with an equal amount of talent and motivation should be offered the same prospective, regardless of their current position within the social system [24].
- *Treatment equality*, in which the ratio of false negatives and false positives is the same for each defined group [6].

Recent studies have explored how such theoretical definitions align with the perspectives of non-experts. Srivastava *et al.* studied fairness perceptions across two different application domains by presenting participants with the results of two hypothetical algorithms side-by-side [51]. Their results highlight that demographic parity is closest in line with people's perception of fairness. Similarly, Saxena *et al.* explored how three different definitions of fairness are perceived by the public in the context of loan allocation [48]. The definitions included are 'individual fairness' [15], 'never favour a worse individual over a better one' [32], and 'calibrated fairness' [39]. The results from Saxena *et al.* reveal a preference for calibrated fairness, in which resources (*e.g.*, money) are split between people in accordance with their 'quality'. Including additional demographic information on the loan applicant's race influenced participants' perceptions, which the authors attribute to affirmative action [48]. Binns *et al.* explore perceptions of justice in algorithmic decision-making across different explanation styles [7]. Their results show explanation styles primarily affect participants' perceptions when participants are presented with multiple different explanation styles. A case-based explanation, in which a representative example from the training data is presented, resulted in the most consistent negative impact on participants' justice perceptions [7].

Highly relevant to our work is a deception-based study by Wang *et al.*, in which the factors influencing perceived fairness of crowdworkers on an algorithm designed to determine their eligibility for a 'Master Qualification' was investigated [60]. Through a between-subjects design, the authors presented participants with different scenarios concerning the algorithm's development (*e.g.*, the openness of the algorithm's development) and performance across different demographics. Based on these scenarios, participants provided their expectation as to whether or not they would receive the master classification, after which a (random) decision result was presented. Participants subsequently rated the fairness of the algorithm. Wang *et al.*'s results highlight that participants perceive algorithms as more fair when the algorithm predicts in their favour – even when the algorithm shows severe biases against particular demographics [60]. This effect, known as outcome favourability, was found to be moderated by a number of demographic (*e.g.*, education) and scenario (*e.g.*, openness) variables.

Pierson collected fairness perceptions among university students prior to and following an hour-long lecture and discussion on algorithmic fairness [45]. Pierson's results show that this intervention changed the perspective of Computer Science students towards algorithmic fairness, and revealed gender differences in relation to the predictors that should be included in an algorithm. Grgić-Hlača *et al.* investigated factors influencing the fairness perceptions of recidivism prediction algorithms [22]. Their results identify eight

properties that predict people's perceptions towards a predictor (*e.g.*, perceived reliability, relevance). Expanding the work into collaborative interaction, Van Berkel *et al.* studied the fairness perception towards ML predictors through interactive discussions facilitated by a chatbot [56], allowing for private voting and public discussion. Their results highlight that being part of a more diverse group resulted in a stronger agreement with the overall majority vote. Similarly, Van Berkel *et al.* highlight the gap between theoretical perspectives on fairness and the public's perception of fairness as an important obstacle to overcome in the public acceptance of AI applications [56].

A common finding in the aforementioned studies is that fairness perceptions among the general public are influenced by a variety of factors. Obtaining a better understanding of these factors across a variety of stages in the development process of algorithms is crucial to develop tools and techniques which allow the general public to interact, inspect, and influence the development of algorithm-based applications. In this paper we focus on one of the initial steps in algorithm development, identifying the perceived fairness of potential algorithmic predictors.

2.2 Inspecting Machine Learning Models

A range of work has established the notion that ML models ought to be interpretable, specifically in order to identify and understand how the model behaves towards different groups [47]. This has led to the development of a variety of interactive applications that allow users to inspect ML models. For example, Wexler *et al.* introduce the 'What-If' tool, an interactive application which allows ML practitioners to "*probe, visualise, and analyze ML systems, with minimal coding*" [62]. The authors argue that no coding should be required to increase ones understanding of a model, instead allowing users to answers 'what-if' questions (*e.g.*, 'how would an increase in the value of age affect the model's decision?') using an interactive tool. Through an iterative design process with machine learning researchers and practitioners, Hohman *et al.* present a visual ML interpretation system named 'Gamut' [26]. Gamut presents an interactive interface with multiple coordinated views, *i.e.*, line charts on multiple features, waterfall charts to explain the cumulative impact of each feature on the final decision, and an interactive table which allows for sorting and filtering of the data.

Kaur *et al.* studied the use of two existing interpretability tools by ML practitioners through a contextual inquiry and survey [33]. Their results indicate that practitioners misuse interpretability tools due to a misalignment between the intended use of these tools and the participants understanding of their workings. This misuse is fuelled by an over-trust of participants in the visualisations: "*participants were excited about the visualizations and took them at face value instead of using them to dig deeper into issues with the dataset or model.*" [33].

We note that the majority of these interpretation tools are aimed at ML experts. Although HCI researchers have stressed the need for stakeholders to be involved in the development of algorithms [11, 56, 58, 64], a thorough understanding on how users would make use of such tools is still missing – resulting in the absence of adequate tooling for the general public. Liao *et al.*, in exploring the needs of practitioners in creating explainable AI applications, highlight that "*XAI [explainable AI] research still struggles with a lack of understanding of real-world user needs for AI transparency, and by far little consideration of what practitioners need to create explainable AI products*" [37]. In this paper, we explore factors affecting non-ML experts in their assessment of algorithmic predictors – a key step in algorithm development. As stated by Wexler *et al.*, "*A particular important future direction is to decrease the level of ML and data science expertise necessary to use this type of tool*" [62], in order to allow a wider public to participate in discussions surrounding the fairness of AI.

2.3 Visualisation and Interpretation

The HCI community has repeatedly shown that the way information is presented can significantly alter the perceptions of information consumers. Related to the topic of explanations, Wang *et al.* empirically compare the effectiveness of infographics and 'data comics' (a sequence of panels in which a story is conveyed through text and data visuals) [61]. Their results highlight that data comics were not only perceived as more engaging and enjoyable but also increased understanding among the readers. Sultanum *et al.* found that clinicians had difficulty obtaining meaningful information from extensive text-based patient records, and introduced a text-visualisation system which allowed clinicians to more efficiently navigate to the required information [52].

On the topic of decision-support algorithms, Cheng *et al.* studied the ability of different explanation interfaces to increase people's understanding of a university admissions decision system [11]. Through a factorial design, the authors explore the effect of interactive vs non-interactive explanations and white-box vs black-box explanations. Their results indicate that participants' *objective* understanding of the algorithm increased both in the interactive and white-box explanations. Interestingly, participants *self-reported* understanding did not increase in the white-box scenario, which the authors argue is the result of increased complexity and/or required cognitive load. Furthermore, an increased understanding did not result in an increase in the participants' trust towards the algorithm [11]. Shen *et al.* explored alternative designs of confusion matrices to support people's understanding of an algorithm's performance, ultimately identifying flow charts as the most useful due to their ability to point out the direction of reading [49].

Despite the literature highlighting that different visualisation techniques significantly impact perception and understanding among individuals, the effect of data presentation on the fairness assessment of algorithm predictors has not been previously studied. Here, we explore two information presentation techniques used in algorithmic assessment; a text-based visualisation and a scatterplot.

3 METHOD

In this paper, we set out to evaluate the effect of information presentation on participants' fairness perceptions and their decision to include or exclude a predictor for an algorithm. We present a mixed-model experimental design, with VISUALISATION TECHNIQUE (text-based or scatterplot) and DISPLAY OF OUTCOME VARIABLE (shown or not shown) as between-subjects factors, and SCENARIO (recidivism or lending) and PREDICTOR TYPE (demographic, validation, and

Figure 1: Overview of VISUALISATION TECHNIQUE and DISPLAY OF OUTCOME VARIABLE as shown to participants.

domain) as within-subjects factors. We first discuss the between-subject factors introduced in the study.

- VISUALISATION TECHNIQUE. We analysed two visualisation techniques. The first technique was a text-based visualisation in which the data was summarised across three sub-groups (Figure 1-C & D), as used *e.g.* by Yu *et al.* [65]. The second technique was a scatterplot based visualisation (Figure 1-A & B). Scatterplots visualise the spread of data points across two variables and are frequently employed in algorithmic inspection applications [42, 62]. Further, scatterplots have been used in other HCI work, including studies involving university students [10] and crowdworkers [28]. Despite the frequent occurrence of scatterplots in ML inspection applications – including those presented to non-experts [65], literature shows that people are often unable to correctly interpret scatterplots and the (strength of the) relationship between the two variables displayed [12, 50]. It is therefore critical to understand whether the use of scatterplots in the context of algorithmic fairness is appropriate for novice users. Participants were able to navigate between different groups within the data (*e.g.*, race, length of loan duration – see Table 1) in both visualisation techniques.
- DISPLAY OF OUTCOME VARIABLE. Presenting the outcome variable may influence participants' decision making, as participants may, for instance, reason that a certain group over- or under-performs in comparison with other groups. We therefore presented either all data points in one overview (*i.e.*, no outcome variable – Figure 1-B & D), or presented the data as split into two views; one view containing all 'positive' outcomes (*e.g.*, did pay back loan) and one view containing all 'negative' outcomes (*e.g.*, did not pay back loan). The outcome variable was clearly indicated above each respective data presentation (Figure 1-A & C). Wang *et al.* identified that participants evaluate algorithms as less fair when they are biased against a certain demographic [60], but the effect of including the outcome variable during predictor assessment has not been previously assessed [56].

In addition to these between-subject manipulations, we introduced two within-subject variables. First, we included two distinct scenarios in order to evaluate whether the presented SCENARIO affects participant judgements; one scenario on recidivism and one scenario on loan approval. Second, for each of the two aforementioned scenarios we introduced a total of five algorithmic predictors – categorised into three PREDICTOR TYPES (demographic, domain-specific, and validation). The order of scenario presentation was counterbalanced, and the order of the predictors was randomised within each scenario. As such, participants first completed one scenario in full before they proceeded to the second scenario. We next detail the scenarios and their respective predictors.

3.1 Scenarios & Predictors

To support the external validity of the study, we made use of two existing real-world datasets. The first dataset is the COMPAS dataset on recidivism[1], as obtained by ProPublica [2] and used in various HCI studies [21, 56, 65]. This dataset contains information on criminal defendants of Broward County, Florida, including their demographic information, prior criminal activities, as well as ground-truth information on whether or not the defendant went on to commit another crime within two years (*i.e.*, the outcome variable). Following the same procedure as outlined by ProPublica [2], we excluded incomplete records, ordinary traffic offences (no jail time), and offenders who were released from a correctional facility less than two years ago.

The second dataset contains data on lending, as published by a peer-to-peer lending company[2], and is frequently used as a case study in machine learning research [30, 41]. The dataset contains information on the amount of money applied for, the status of the loan, as well as limited details on the loan applicant. The dataset covers the period 2007–2015 and consists of over two million loans. Our first step was to filter out loans that were still ongoing, leaving us with a binary outcome variable: loans that were either fully paid back or failed to be paid back (defaulted or written off).

Our selection is based on three criteria. First, an identical distribution of predictor types (demographic, domain-specific, and validation) between scenarios. Second, variables as similar as possible (*e.g.*, we could use 'sex' in both datasets and created 'age' for the lending dataset). Given the limited number of demographic variables in the lending dataset (*e.g.*, no information on the applicant's age), we fabricated the lenders' age by scaling the reported annual income of lenders (excluding outliers) to an age variable ranging from 18 to 86. Furthermore, we mapped the dataset's information on the lender's state to one of four regions as defined by the United States Census Bureau (*i.e.*, Northeast, Midwest, South, West). Third, a consistent presentation of the variables limited us to continuous variables for the predictor variable and categorical variables for the group filter. Therefore, not all predictors from the Lending dataset could be included. Following recommendations from earlier work [20, 34, 56], we generated fabricated predictors for both datasets to be used as 'explicitly verifiable' predictors. We present an overview of the variables selected in Table 1. For consistency in size between the two scenarios, we randomly sampled 5000 records of each dataset.

3.2 Participants & Procedure

Participants were recruited through Prolific Academic. To ensure sufficient response quality, we restricted participation to crowdworkers with an acceptance rate of 95% or above. Furthermore, we limited our study sample to participants with a US nationality. Although a wider selection of nationalities would allow for a more global perspective, this restriction aligns with the characteristics of the datasets. After accepting the task, participants were routed to an online website with an explanation of the task and a short introduction to decision-making algorithms by use of an example. Participants received a predetermined amount of money for the full completion of the task. Following the US minimum wage of $7.25 per hour at the time of our study and an expected completion time of 20 minutes (as based on our pilot data), we compensated each participant with $3.

To minimise type II errors, we defined the number of participants based on a power calculation using G*Power [18]. Given the exploratory nature of our investigation, we used medium-to-large effect sizes ($f^2 = 0.2$), an alpha level of 0.05, and a power of 0.8, in line with established methodological recommendations [23]. From our *a priori* multiple linear regression model with four predictors, the minimum sample size required is 65 participants. To be conservative (foreseeing dropouts) and to uphold reliability, we ended up recruiting 80 participants.

Upon accepting the task in the Prolific crowdsourcing platform, participants were assigned to one of four conditions – following the aforementioned study design. We limited participants from taking part in our experiment more than once, ensuring that each condition consists of twenty unique participants. Each participant completed the following four tasks:

(1) **Introduction**. The landing page explained the task at hand, including a fictional example of an algorithmic decision application. At this stage, participants provided their demographic information, which was used to inform the default options for the demographic predictors in task 3.
(2) **Background information**. Participants first read a high-level explanation of how algorithms come to a decision. Furthermore, the two scenarios included in the experiment (*i.e.*, recidivism and lending) were introduced. To assess whether participants understood and paid attention while reading this relatively short (~300 words) background information, we included two multiple-choice questions. Participants were allowed to continue with, and receive compensation for, the task regardless of the correctness of their answers.
(3) **Predictor assessment**. For the main task of the experiment, participants were asked to assess a total of ten 'predictor'–'group filter' pairs (see Table 1). For each predictor variable, participants could explore the data across a set of groups. The default display for the race and sex group filters was determined by the participant's demographic information (*e.g.*, a male participant will first see the data of male loan applicants). For the remaining group filter configurations, a consistent default group was selected. Participants were asked to assess and subsequently motivate whether a predictor variable should be included (binary yes/no decision) in an algorithm predicting either the risk of not paying back a loan or a convicts' chance of recidivism – as dependent on the current scenario. Furthermore, participants assessed the fairness of including this predictor in an algorithm for the given scenario (scale 0–100%). To recap, for each assessed predictor we kept track of the PREDICTOR TYPE, VISUALISATION TECHNIQUE, DISPLAY OF OUTCOME VARIABLE, and SCENARIO. This allowed us to later model the effect of these parameters on participants decision making.
(4) **Graph comprehension & completion**. Finally, we assessed the graph literacy of all participants using the Short Graph Literacy (SGL) scale [44]. The SGL is a validated questionnaire consisting of four questions – each including a visual graph.

[1]Data available at https://github.com/propublica/compas-analysis/.
[2]Data available at https://www.lendingclub.com/info/statistics.action.

Recidivism dataset		Lending dataset		Predictor type
Predictor variable	Group filter	Predictor variable	Group filter	
Age	Race	Age	US Region	Demographic
Age	Sex	Age	Sex	Demographic
Nr. prior convictions	Degree of charge	Loan amount	Loan duration	Domain-specific
Nr. prior convictions	Violent offence	Loan amount	Home ownership	Domain-specific
Height	Dominant hand	Weight	Vegetarian	Validation

Table 1: Overview of predictor variables and the respective group filter through which the display of the data is controlled. The last column shows the PREDICTOR TYPE of the predictor variables across the two scenarios.

Furthermore, we collected the participants' overall perceptions with regards to their decision confidence and perspectives on end-user involvement in algorithm development.

4 RESULTS

The average age of our participants was 29.05 years old (SD = 8.06), ranging between 18 and 60 years of age. We provide an overview of the gender, race, and educational attainment of our participant sample in Table 2, and augment this information with data from the 2019 US Census [54, 55]. Our participants' demographics are closely aligned with the US census. In terms of participants' race, we see an under-representation of African-Americans and an over-representation of Asians in comparison with the US population – this aligns with prior analysis of the crowdworker population [5].

First, we assessed participants' answers on the two questions intended to verify whether participants had read and understood the background information provided. For the first question ('*Algorithms calculate the risk of an individual by;*'), we found 14 participants that answered incorrectly. For the second question ('*A recidivist is someone who;*'), we found 13 participants who answered incorrectly. Five participants answered incorrectly on both questions, indicating a poor understanding, low participant effort, or automated efforts. These participants were therefore excluded from further analysis, leaving a total of 75 participants. Given this final sample size, we report a post-hoc power of 0.867.

Factor	*N*	% sample	US Census
Gender			
Women	39	48.8%	49.0%
Men	39	48.8%	51.0%
Non-binary	1	1.3%	-
Prefer not to disclose	1	1.3%	-
Race			
African-American	4	5.0%	13.4%
Asian	9	11.3%	5.9%
Caucasian	48	60.0%	60.4%
Hispanic	12	15.0%	18.3%
Native American	0	0%	1.5%
Other	6	7.5%	2.7%
Prefer not to disclose	1	1.3%	-
Educational attainment			
Primary education	2	2.5%	8.4%
Secondary edu. or high school	26	32.5%	28.3%
Bachelor or vocational degree	38	47.5%	31.1%
Master's degree	12	15.0%	10.2%
Doctorate or higher	2	2.5%	1.8%

Table 2: Overview of participants' demographic factors.

Average completion time for the ten primary questions was 18.91, 24.10, 22.57, and 18.77 minutes for the 'Scatterplot - Outcome', 'Scatterplot - No outcome', 'Text - Outcome', and 'Text - No outcome' conditions respectively. A two-way ANOVA, using type-III sums of squares to account for the slight imbalance in group size following the aforementioned participant removal, was conducted to examine the effects of VISUALISATION TECHNIQUE and DISPLAY OF OUTCOME VARIABLE on completion time. There was no statistically significant interaction between the effects of visualisation type and inclusion of outcome variable ($F(1,71) = 3.243$, $p = 0.08$). Simple main effects analysis also did not show a significant effect for visualisation type ($p = 0.13$) or inclusion of outcome variable ($p = 0.14$).

Participants' average graph literacy score, as assessed through the Short Graph Literacy scale (SGL) [44], was 2.77 (SD = 0.94, ranging from 1 to 4). This score is slightly higher than the average scores observed by the initial assessment of the SGL on a US population (score of 2.2, $N = 495$) [44].

4.1 Predictor Assessment

In order to identify the effects of PREDICTOR TYPE, VISUALISATION TECHNIQUE, DISPLAY OF OUTCOME VARIABLE, and SCENARIO on the likelihood that participants believed a predictor should be included in an algorithm, we constructed a binomial (include/exclude) generalised linear mixed model. The model is based on the collected participants' responses to include or exclude the predictor variables presented in Table 1. We limited our model to solely include model parameters which can be obtained without participant input, thereby choosing to exclude *e.g.* demographic factors and the participant's assessment on the fairness of including a predictor (as recorded from 0–100%), as these factors are not known *a priori*. For the categorisation of predictor types, we followed the categorisation shown in Table 1: demographic, domain-specific, and validation. In order to account for variability between participants, we set participant ID as the random factor of the model. The model is constructed using the *glmer* function in R package *lme4* [4].

	Include
Predictor type (Validation)	−3.92 (0.55)***
Predictor type (Domain)	−0.37 (0.35)
Visualisation technique (Text)	−0.82 (0.41)*
Display of outcome variable (Yes)	0.18 (0.41)
Scenario (Recidivism)	1.20 (0.34)***
Predictor type (Validation) : Visualisation technique (Text)	1.32 (0.61)*
Predictor type (Domain) : Visualisation technique (Text)	1.59 (0.44)***
Predictor type (Validation) : Display of outcome variable (Yes)	−0.10 (0.61)
Predictor type (Domain) : Display of outcome variable (Yes)	0.51 (0.43)
Scenario (Recidivism) : Visualisation technique (Text)	0.72 (0.41)
Scenario (Recidivism) : Display of outcome variable (Yes)	−0.93 (0.41)*

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table 3: The generalised linear model (binomial) of participants' decision to include or exclude a predictor. For each predictor we report coefficients, standard errors (in brackets), and significance indicators.

A likelihood ratio test as compared to the null model showed that our logistic regression model is statistically significant ($\chi^2(11) = 265.78$, $p < .001$) [8]. Our model explains 54.1% of the variance in predictor assessment ($\mathbb{R} = 0.73$, $\mathbb{R}^2 = 0.54$). To ensure the validity of the model, we checked for the existence of multicollinearity among the model's parameters. We found a variance inflation factor (VIF) of our parameters between 1.76 and 3.34, all below the often-used thresholds of five or ten to detect multicollinearity [23], indicating the validity of the model. The model outcomes are summarised in Table 3, with the following reference levels; PREDICTOR TYPE (Demographic), VISUALISATION TECHNIQUE (Scatter), DISPLAY OF OUTCOME VARIABLE (No), SCENARIO (Lending).

We discuss the significant predictors identified in our model (Table 3) in more detail to obtain a better understanding of their effect. We first present the main effects of PREDICTOR TYPE, VISUALISATION TECHNIQUE, and SCENARIO, after which the significant interaction effects are presented.

Figure 2-A highlights the significant difference between Validation predictors and the two remaining predictor types (Demographic and Domain), with participants being substantially less likely to include a Validation predictor. This highlights the Validation predictors' ability to function as an explicitly verifiable question. The main effect of the visualisation type is shown in Figure 2-B. Our results highlight that participants were significantly more likely to include a variable when presented in text as opposed to visualised in a scatterplot presentation. The third significant main effect, scenario, is visualised in Figure 2-C. These results show that participants were significantly less likely to vote for a parameter to be included in the model in the lending scenario. Furthermore, we found that the confidence intervals are substantially narrower for the recidivism scenario as compared to the lending scenario – indicating higher levels of agreement among participants for the recidivism scenario.

Our model highlights two significant interactions, the first one between PREDICTOR TYPE and VISUALISATION TECHNIQUE, and the second one between SCENARIO and DISPLAY OF OUTCOME VARIABLE. The interaction between PREDICTOR TYPE and VISUALISATION TECHNIQUE is shown in Figure 3-A. The significant interaction indicates that the chosen VISUALISATION TECHNIQUE affects participants' judgement differently between different PREDICTOR TYPES. For the Domain-specific predictors (*e.g.*, loan amount), a text-based visualisation resulted in a higher likelihood to include the variable, whereas the effect is reversed for the Demographic predictors. Although the DISPLAY OF OUTCOME VARIABLE is not a significant predictor of participants' assessment of predictor inclusion on its own,

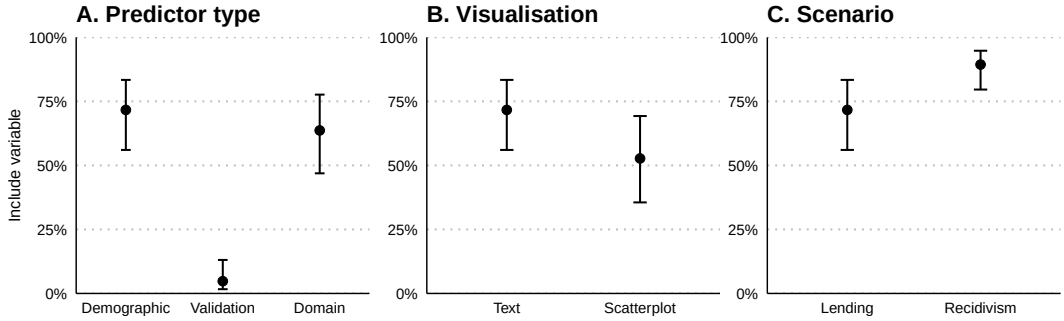

Figure 2: Visualisation of the significant main effects of our linear model (Table 3), using *ggeffects* [40].

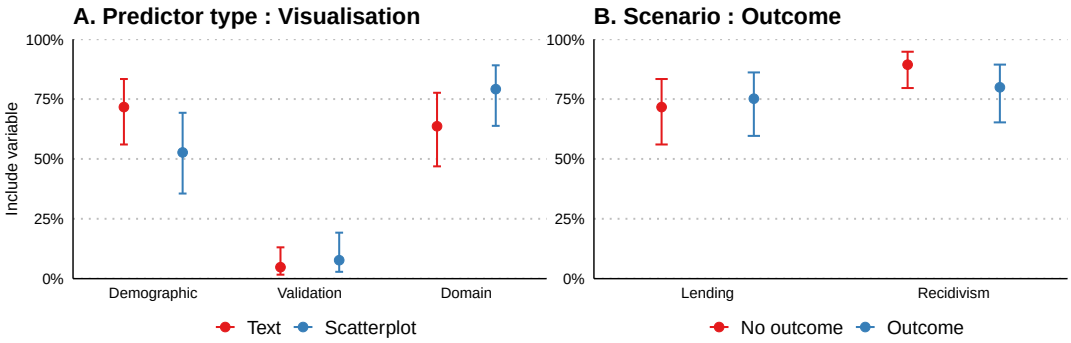

Figure 3: Visualisation of the significant interaction effects of our linear model (Table 3), using *ggeffects* [40].

the interaction between the DISPLAY OF OUTCOME VARIABLE and SCENARIO is significant (Table 3). As shown in Figure 3-B, displaying the outcome variable reduced the likelihood of a participant to include a parameter in the Recidivism scenario, whereas this effect is nonexistent for the Lending scenario.

4.2 Perceptions of Fairness

In addition to a binary assessment of including or excluding a predictor, as presented in the preceding section, we asked participants to assess the fairness of each algorithmic predictor on a scale from 0–100%. As expected, this perceived fairness is closely aligned to the participants' decision to include/exclude a predictor – with a mean perceived fairness of 25.0% (SD = 28.5) on the predictors participants voted to exclude, and 76.9% (SD = 23.1) on predictors voted to be included. A Mann-Whitney rank-sum test confirmed that the distributions of fairness perceptions (as rated from 0–100%) differed significantly between predictors voted to include versus predictors voted to exclude (U = 13765, $p < 0.001$).

Next, we explored the effect of the participants' educational attainment on their perceptions of fairness. Given the limited occurrence of participants with a highest level of education as 'Primary school' (2) or 'Doctorate or higher' (2) we excluded these four participants from our analysis. A Kruskal-Wallis rank-sum test showed a significant effect of attained education level on perceived fairness ($\chi^2(2) = 15.59$, $p < 0.001$), with the average level of perceived fairness decreasing with an increase in attained education. A pairwise comparison using Wilcoxon rank-sum test showed a significant difference between 'Secondary education or high school' and 'Bachelor or vocational degree' ($p = 0.006$) and between 'Secondary education or high school' and 'Master's degree' ($p < 0.001$). Figure 4 shows these differences in a density plot. A Kruskal-Wallis test between participant graph literacy score and their perceived fairness was significant ($\chi^2(3) = 7.843$, $p = 0.049$), with average fairness perceptions decreasing as graph literacy score increases (mean values of 0.65, 0.62, 0.56, and 0.55 for graph literacy scores 1 to 4 respectively).

Subsequently, we analysed the effect of gender on the participants' perceived fairness. Our sample consists of one non-binary participant, and one participant who did not wish to disclose their gender – we excluded these participants from our analysis due to their limited representation in our sample. Using a Mann-Whitney test, we found a significant difference between the perceived predictor fairness by women (mean = 52.9, SD = 34.3) and men (mean = 65.3,

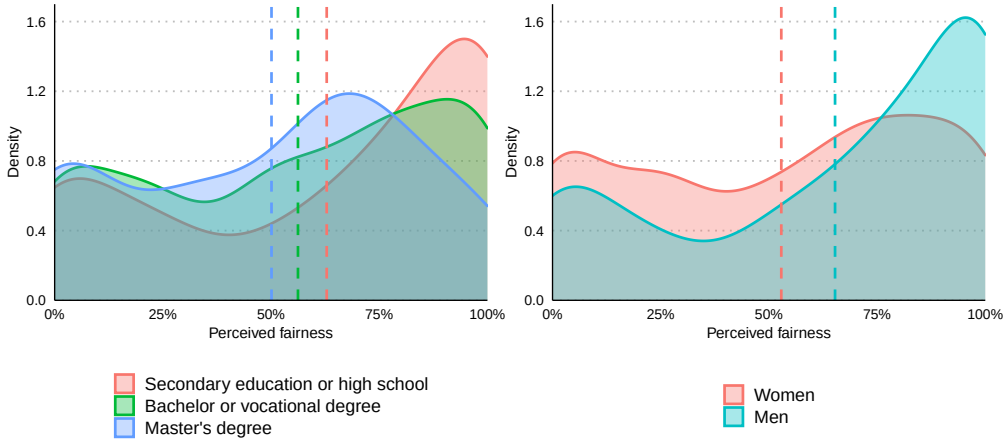

Figure 4: Fairness ratings by educational attainment (left) and sex (right). Vertical lines indicate mean group values.

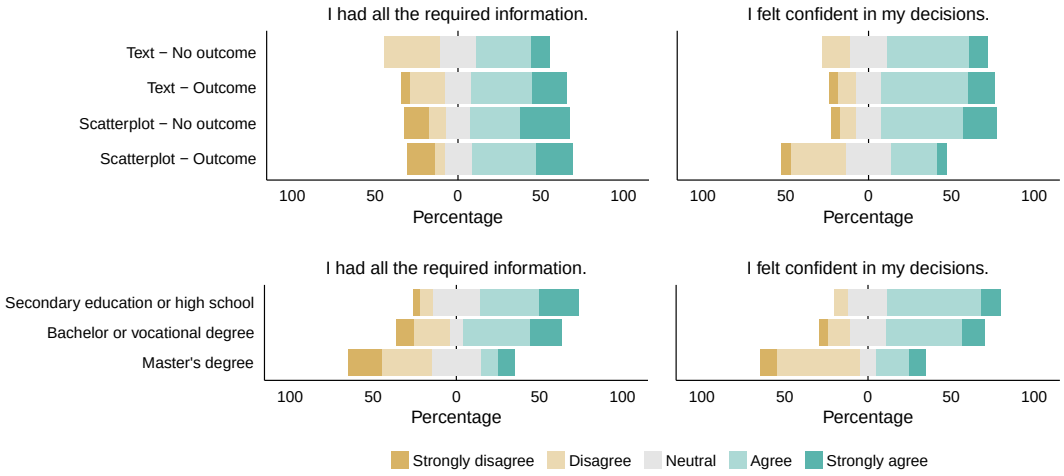

Figure 5: Overview of Likert-responses as split per condition (top) and educational attainment (bottom).

SD = 35.4), U = 13765, $p < 0.001$, with women reporting significantly lower perceived fairness levels. We did not find a significant difference between the women and men in our sample and their educational attainment ($\chi^2(4) = 2.547$, $p = 0.636$).

4.3 Information Demands and Confidence

Lastly, we present participants' responses to the Likert statements '*I had all the required information to assess the fairness of the algorithm.*' and '*I feel confident in my decisions*', as collected in the final stage of the study. Figure 5 presents participants' answers as split per condition and level of educational attainment. These results show that participants were largely divided over the fact whether all required information was presented, and that this division is more or less equal between conditions. Participants who responded negatively to the question commented primarily on a lack of background information presented. For example, one participant stated:

> "*What were people getting these loans for? Some loans and debt are relatively stable and others aren't. Also what else was included in the algorithm? Anything beyond the factors that were discussed? I might have had a different opinion of whether certain factors belonged in the algorithm if I knew what other factors were being taken into consideration.*" [P56]

Information was presented to participants per predictor, whereas several participants requested a wider overview of participant data:

> "*Hyperfocusing on just one aspect meant that I was missing the whole picture, and that meant certain groups were always going to be unfairly evaluated.*" [P60]

In terms of decision confidence, participants leaned towards self-assurance in all conditions, with the exception of the 'Scatterplot - Outcome' combination. Participants in this condition disproportionately commented on the difficulty of interpreting the presented information, such as:

> "*The fact that the graphs were incredibly difficult for me to decipher.*" [P13]

> "*I wasn't sure if I was reading the graphs entirely correctly.*" [P05]

One of the participants which expressed high confidence in their decisions stated to find confidence in their existing prejudices:

> "*Prior experience with individuals in some of the groups outlined.*" [P33]

In line with the aforementioned fairness perceptions (Figure 4), we found that those with higher educational attainment express a stronger need for additional information and generally have lower confidence in their decisions.

5 DISCUSSION

Through the presented study we aim to understand the effect of data presentation on people's fairness perceptions. As an individual's perception of fairness cannot be classified as inherently 'right' or 'wrong', our work does not aim to identify one optimal presentation technique which captures the fairest outcomes. Instead, our work is based on the call-to-action to the Computer Science community to actively involve the general public in the decision making process of new AI systems [13, 51, 56, 64]. Building on this call and the ongoing efforts towards creating interactive tools for inspecting algorithmic predictors and models [33, 38, 62], we explore the effect of data visualisation technique (text vs scatterplot) and providing information on the outcome variable (split data by outcome variable vs joint data presentation) on perceptions of ML predictors. From a wider perspective, a better understanding of the public's perceptions and interpretations of fairness are critical in supporting both an informed debate between the public and technologists, as well as giving a voice to non-ML experts in the design of future applications.

Following prior work on fairness perceptions [56], we introduce a 'verifiable' predictor pair in each scenario ('height–dominant hand' and 'weight–vegetarian') to assess participants' ability to identify and exclude these predictors. Our results show that these predictors are largely excluded (Figure 2-A) – providing confidence in the participants' understanding and ability to complete the given task of assessing algorithmic predictors.

5.1 Information Presentation

The presentation of predictors using a text-based visualisation technique, in which data was summarised into three categories, resulted in an increased likelihood of predictors being included as compared to a scatterplot-based visualisation technique – despite the underlying data being identical. The literature on risk management has a long history of studying the effects of information presentation on participant risk perceptions. For example, Gibson *et al.* found that a stacked bar graph improved participants' risk understanding as compared to textual information [19]. Similarly, Tait *et al.* found that pictographs were found to significantly increase health risk understanding as compared to text and tables [53], regardless of the participants' numerical literacy. Contrasting these findings to our results, in which a text-based presentation resulted in more variables being included (*i.e.*, higher perceived levels fairness), we hypothesise that the visual presentation included in the scatterplots more clearly communicated differences between groups (*e.g.*, race, loan duration) within predictors. Although our work was limited to two visualisation techniques currently used in the Human-AI literature, we consider this an opportunity for future work to explore visualisation techniques not yet widely considered in AI applications. For example, the use of pictographs, as studied by Tait *et al.* [53], might be more easily interpretable.

Our results also show that participants felt significantly less confident about their decisions in the condition in which data was visualised in a scatterplot while data was simultaneously split by outcome variable – this effect is not apparent in any other combination. Work by Cheng *et al.* highlights that white-box explanations do not increase participants *self-reported* level of understanding as compared to a black-box explanation [11] (although their *objective* understanding does increase), which the authors attribute to increased complexity and required cognitive load. Our results indicate that a similar effect appears in relation to participants' confidence, with our most complex design (Scatterplot - Outcome) showing lower levels of decision making confidence (Figure 5). We argue that this particular combination (Scatterplot - Outcome) may therefore exceed the capacity of study participants to fully comprehend the information shown.

5.2 Education and Gender

Our results highlight that participants with higher educational attainment significantly rated predictors as being less fair (Figure 4). Participants with a higher level of education also felt less confident in their decisions and were less likely to agree with the statement that they had all the required information to assess the fairness of the predictor. These results point in the direction of the well-known Dunning–Kruger effect [35]. In their seminal paper, Dunning and Kruger state that "*those with limited knowledge in a domain suffer a dual burden: Not only do they reach mistaken conclusions and make regrettable errors, but their incompetence robs them of the ability to realize it*" [35]. Therefore, future work which aims to involve a wider study population (in contrast to typically highly-educated AI developers) should take note of the fact that a large subset of their sample will not consider the fact that additional information might be required to draw conclusions. Researchers should, therefore, consider to explicitly indicate which information is missing when presenting information to participants.

Furthermore, we found the gender difference in perceived fairness of predictors to be particularly noteworthy given the current state of the software industry. According to data from the US Bureau of Labor Statistics, women compose only 26% of the US computing workforce population in 2019 [43]. It is therefore probable that AI systems are primarily developed by men, while our results highlight that the fairness perspectives of women differ significantly from that of men. This is reflected in prior work in Economics, with *e.g.* Andreoni & Vesterlund finding empirical evidence for differences in altruistic behaviour between men and women under different circumstances [1]. We, therefore, align ourselves with earlier calls that expressed the need to collect opinions from a diverse participant sample when assessing perceptions of AI systems [45, 56, 66]. Pierson states; "*If demographics predict how we believe algorithms should behave, we need our algorithm designers to be more demographically representative if algorithms are to serve the will of the whole population*" [45]. Although we support such endeavours, an analysis by Wang *et al.* indicates that the gender gap in Computer Science research will not reach parity between men and women until 2137 under the current developments [59], if such parity will ever be reached at all. As such, we consider the input of a more diverse participant sample, as we are able to achieve through the applied method (Table 2), as a requirement when assessing the fairness of algorithmic systems. This aligns with prior work by Williams and her call for gender-neutral software [63].

5.3 Studying AI Perceptions

Correll argues that the HCI community has a moral duty to communicate the algorithmic decision-making process to those impacted by the algorithm, stating that current work on assessing and interpreting ML models is primarily focused on ML practitioners; "*An ethical concern is that we are therefore empowering the creators of ML models, but are not empowering the people affected by these models.*" [13]. Although our current study is focused solely on predictor selection, it highlights a number of recommendations for future research into the wider area of AI perceptions among non-ML experts. To support further research into this critical domain, we summarise our recommendations below;

- **Limit presentation complexity**. Our results show that participant confidence is reduced when presented with overly complex information. As such, we recommend against the direct use of machine learning terminology in studies involving participants from a non-ML background (*e.g.*, we find it unlikely that terms such as 'disparity' can be understood by the general public [65]). Prior work further points to a better understanding of risks using graphical visualisations rather than a text-based presentation [19, 53], opening the possibility for *e.g.* pictograph-based communication.
- **Tailor information presentation to target audience**. Following the above recommendation, we stress that information presentation should be adjusted to the target audience to ensure that the user can fully comprehend the information presented. As such, the same style of information presentation used by ML/AI

experts cannot simply be applied to novice users or members of the general public.

- **Validate results across scenarios/datasets**. While prior work has typically focused on one scenario or dataset at a time (see *e.g.*, [22, 56, 65]), our results show that the presented scenario has a significant effect on the participants' fairness assessments. Therefore, we recommend studies which aim to make a methodological contribution (such as ours) to validate findings across at least two distinct scenarios.
- **Include verifiable predictors**. Following earlier work and recommendations on including verifiable questions [20, 34, 56], we included two pairs of verifiable predictors. These predictors were excluded by a large majority of participants at a significantly higher rate than other types of predictors (Figure 2). By including verifiable predictors researchers can assess whether the task was sufficiently understood by participants.
- **Instruct and assess participants**. Although our goal was explicitly to widen data collection to non-ML experts, we decided to exclude five participants due to their incorrect answers on the background questions – following prior work by Yu *et al.* [65]. We argue that a basic understanding of the study concepts is required to provide reliable input, and it is therefore the responsibility of the researchers to carefully instruct participants on the problem at hand. While it is impossible to state whether the excluded participants were unable to comprehend our explanation, unmotivated to read the background text, or were perhaps automated survey completion bots, we consider all three as harmful to data collection. Therefore, we recommend excluding participants based on a short assessment of their understanding of the core concepts of the study following a succinct explanation.

We publicly release the source code of our study application to enable others to replicate and extend our work[3].

5.4 Limitations

We recognise several limitations in our work that should be considered when interpreting the presented results. First, our participants were presented with either a scatterplot or text-based presentation of the predictor variable. Numerous other visualisation techniques exist, and they are likely to result in a different effect on participants' fairness perceptions. Future work may assess the effect of other visualisation techniques. Second, the presented scenarios (*i.e.*, recidivism, lending) covered distinctly different topics with a disparate predictor data distribution. It is therefore unsurprising that participants assess the fairness of the individual predictors differently between scenarios, and the inclusion of additional scenarios would have likely revealed further differences. Although our results highlight similarities in the effects of predictor type and visualisation technique, we particularly stress the need for researchers to consider multiple scenarios when assessing human-AI assessment tools. Third, our study focused solely on predictor selection – one of the several steps in the development of an AI algorithm. We, therefore, call on the HCI research community to expand the study of methodological considerations of non-expert involvement to other phases of AI development. Finally, we note that not one 'public perception' of fairness exists, and that results are likely to differ when studying different populations. Our study was limited to American crowdworkers, and prior work shows that perceptions towards *e.g.* AI behaviour and fairness differ significantly between geographical and cultural clusters [3, 57].

[3] https://github.com/nielsvanberkel/Information-Presentation-Fairness

6 CONCLUSION

HCI research and practice builds on a longstanding tradition of end-user involvement. With an increasing focus on Human-AI interaction, we must ensure that the methods and tools used for assessing fairness perceptions provide researchers and practitioners with reliable insights. Through a mixed-model study design, we found that the frequently used scatterplot visualisation technique led to a significantly lower fairness perception as compared to a text-based visualisation. Presentation of the outcome variable (*i.e.*, splitting the presented dataset) did not result in a main effect on predictor assessment but did show an interaction effect with the presented scenario. Further, we found a significant difference in fairness assessments between the two scenarios, highlighting the need for method-based studies to consider more than one scenario when assessing the effect of *e.g.* data presentation on perceived fairness. Finally, gender and educational attainment significantly affected perceived fairness, with women and higher educated participants perceiving the presented predictors as less fair. This highlights the opportunities and needs for future work towards ensuring accessible and educational AI evaluation tools in order to reach a wide and diverse segment of the population.

ACKNOWLEDGMENTS

We would like to express our gratitude to our study participants and the reviewers of this manuscript. This research is partially funded by the GenZ strategic profiling theme at the University of Oulu, supported by the Academy of Finland (project number 318930).

REFERENCES

[1] James Andreoni and Lise Vesterlund. 2001. Which is the Fair Sex? Gender Differences in Altruism. *The Quarterly Journal of Economics* 116, 1 (2001), 293–312. https://doi.org/10.1162/003355301556419

[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 March 11, 2020 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (2018), 59. https://doi.org/10.1038/s41586-018-0637-6

[4] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[5] 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

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (2018), 1–42. https://doi.org/10.1177/0049124118782533

[7] 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)*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3173574.3173951

[8] Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 3 (2009), 127–135.

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

[10] Helen Chen, Sophie Engle, Alark Joshi, Eric D. Ragan, Beste F. Yuksel, and Lane Harrison. 2018. Using Animation to Alleviate Overdraw in Multiclass Scatterplot Matrices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3173574.3173991

[11] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300789

[12] William S Cleveland, Persi Diaconis, and Robert McGill. 1982. Variables on scatterplots look more highly correlated when the scales are increased. *Science* 216, 4550 (1982), 1138–1141. https://doi.org/10.1126/science.216.4550.1138

[13] Michael Correll. 2019. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3290605.3300418

[14] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women – Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[15] 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)*. Association for Computing Machinery, 214–226. https://doi.org/10.1145/2090236.2090255

[16] Sophie Emerson, Ruairí Kennedy, Luke O'Shea, and John O'Brien. 2019. Trends and applications of machine learning in quantitative finance. In *8th International Conference on Economics and Finance Research*.

[17] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–11. https://doi.org/10.1145/3313831.3376232

[18] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (2009), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

[19] Jacqueline MacDonald Gibson, Aimee Rowe, Eric R. Stone, and Wandi Bruine de Bruin. 2013. Communicating Quantitative Information About Unexploded Ordnance Risks to the Public. *Environmental Science & Technology* 47, 9 (2013), 4004–4013. https://doi.org/10.1021/es305254j

[20] Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. 2013. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 753–762. https://doi.org/10.1145/2493432.2493481

[21] 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, Article 178 (2019), 25 pages. https://doi.org/10.1145/3359280

[22] 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, 903–912. https://doi.org/10.1145/3178876.3186138

[23] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and RL Tatham. 2010. *Multivariate Data Analysis*. Pearson.

[24] 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., 3323–3331.

[25] Jeanne G Harris and Thomas H Davenport. 2005. Automated decision making comes of age. *MIT Sloan Management Review* 46, 4 (2005), 2–10.

[26] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3290605.3300809

[27] Simo Hosio, Andy Alorwu, Niels van Berkel, Miguel Bordallo, Mahalakshmy Seetharaman, Jonas Oppenlaender, and Jorge Goncalves. 2019. Fueling AI with Public Displays? A Feasibility Study of Collecting Biometrically Tagged Consensual Data on a University Campus. In *Proceedings of the ACM International Symposium on Pervasive Displays (PerDis'19)*. 14:1–14:7. https://doi.org/10.1145/3321335.3324943

[28] Jessica Hullman, Eytan Adar, and Priti Shah. 2011. The Impact of Social Information on Visual Judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, 1461–1470. https://doi.org/10.1145/1978942.1979157

[29] Ellora Thadaney Israni. 2017. When an Algorithm Helps Send You to Prison – New York Times. https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html

[30] Julapa Jagtiani and Catharine Lemieux. 2019. The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management* 48, 4 (2019), 1009–1029. https://doi.org/10.1111/fima.12295

[31] Elizabeth E Joh. 2017. Artificial intelligence and policing: First questions. *Seattle University Law Review* 41 (2017), 1139.

[32] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). 325–333.

[33] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3313831.3376219

[34] 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, 453–456. https://doi.org/10.1145/1357054.1357127

[35] J. Kruger and D. Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77, 6 (1999), 1121–1134.

[36] Jeff Levin. 2019. Three Ways AI Will Impact The Lending Industry – Forbes. https://www.forbes.com/sites/forbesrealestatecouncil/2019/10/30/three-ways-ai-will-impact-the-lending-industry/

[37] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–15. https://doi.org/10.1145/3313831.3376590

[38] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, 2119–2128. https://doi.org/10.1145/1518701.1519023

[39] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. 2017. Calibrated Fairness in Bandits. arXiv:cs.LG/1707.01875

[40] Daniel Lüdecke. 2018. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software* 3, 26 (2018), 772. https://doi.org/10.21105/joss.00772

[41] Milad Malekipirbazari and Vural Aksakalli. 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications* 42, 10 (2015), 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

[42] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. https://doi.org/10.1073/pnas.1900654116

[43] US Bureau of Labor Statistics. 2019. Labor Force Statistics from the Current Population Survey. https://www.bls.gov/cps/cpsaat11.htm

[44] Yasmina Okan, Eva Janssen, Mirta Galesic, and Erika A. Waters. 2019. Using the Short Graph Literacy Scale to Predict Precursors of Health Behavior Change. *Medical Decision Making* 39, 3 (2019), 183–195. https://doi.org/10.1177/0272989X19829728

[45] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. arXiv:cs.CY/1712.09124

[46] McKenzie Raub. 2018. Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Arkansas Law Review* 71 (2018), 529.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, 1135–1144. https://doi.org/10.1145/2939672.2939778

[48] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, 99–106. https://doi.org/10.1145/3306618.3314248

[49] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. In *Proceedings of the ACM 2020 Conference on Computer Supported Cooperative Work*

(CSCW '20). Association for Computing Machinery. https://doi.org/10.1145/3415224

[50] Varshita Sher, Karen G. Bemis, Ilaria Liccardi, and Min Chen. 2017. An Empirical Study on the Reliability of Perceiving Correlation Indices using Scatterplots. *Computer Graphics Forum* 36, 3 (2017), 61–72. https://doi.org/10.1111/cgf.13168

[51] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Association for Computing Machinery, 2459–2468. https://doi.org/10.1145/3292500.3330664

[52] Nicole Sultanum, Michael Brudno, Daniel Wigdor, and Fanny Chevalier. 2018. More Text Please! Understanding and Supporting the Use of Visualization for Clinical Text Overview. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3173574.3173996

[53] Alan R. Tait, Terri Voepel-Lewis, Brian J. Zikmund-Fisher, and Angela Fagerlin. 2010. The Effect of Format on Parents' Understanding of the Risks and Benefits of Clinical Research: A Comparison Between Text, Tables, and Graphics. *Journal of Health Communication* 15, 5 (2010), 487–501. https://doi.org/10.1080/10810730.2010.492560

[54] U.S. Census Bureau. 2019. Educational Attainment in the United States. https://www.census.gov/data/tables/2019/demo/educational-attainment/cps-detailed-tables.html.

[55] U.S. Census Bureau. 2019. Population Estimates by Age, Sex, Race and Hispanic Origin. https://www.census.gov/newsroom/press-kits/2020/population-estimates-detailed.html.

[56] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 28:1–28:21. https://doi.org/10.1145/3359130

[57] Niels van Berkel, Eleftherios Papachristos, Anastasia Giachanou, Simo Hosio, and Mikael B. Skov. 2020. A Systematic Assessment of National Artificial Intelligence Policies: Perspectives from the Nordics and Beyond. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI '20)*. Association for Computing Machinery, Article 10, 12 pages. https://doi.org/10.1145/3419249.3420106

[58] Niels van Berkel, Benjamin Tag, Jorge Goncalves, and Simo Hosio. 2020. Human-centred artificial intelligence: a contextual morality perspective. *Behaviour & Information Technology* 0, 0 (2020), 1–17. https://doi.org/10.1080/0144929X.2020.1818828

[59] Lucy Lu Wang, Gabriel Stanovsky, Luca Weihs, and Oren Etzioni. 2019. Gender trends in computer science authorship. arXiv:cs.DL/1906.07883

[60] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3313831.3376813

[61] Zezhong Wang, Shunming Wang, Matteo Farinella, Dave Murray-Rust, Nathalie Henry Riche, and Benjamin Bach. 2019. Comparing Effectiveness and Engagement of Data Comics and Infographics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300483

[62] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1, 56–65.

[63] Gayna Williams. 2014. Are You Sure Your Software is Gender-Neutral? *Interactions* 21, 1 (2014), 36–39. https://doi.org/10.1145/2524808

[64] 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, Article 656, 14 pages. https://doi.org/10.1145/3173574.3174230

[65] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. Association for Computing Machinery, 1245–1257. https://doi.org/10.1145/3357236.3395528

[66] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 194 (2018), 23 pages. https://doi.org/10.1145/3274463