SUMMARY
We investigated unconscious bias embedded in job advertisements for the pre-thesis. With a focus on linguistic bias, we investigated how gender bias words impact potential job applicants. Although organizations aim to build diversity in the workplace, it is important to emphasize that a job advertisement where certain words consist of unconscious bias, has an impact on underrepresented groups [1, 2, 3]. The topic was inspired by the increasing demand for diversity during the recent rising social justice movement fueled by the Black Lives Matter protests [3, 4], which have put pressure on businesses to improve their diversity. Years of study and statistical analysis demonstrate that diversity improves the work environment, innovation, and creativity [1, 3, 5]

There is a growing interest in Human-Computer Interaction (HCI) in how design might perpetuate unconscious bias, which reflects values and norms related to each designer's identity. All human interactions are influenced by unconscious bias, which affects attitude and conduct. It has been established that focusing on the issue of unconscious bias results in more qualified candidates [6, 7, 8, 9].

Crowdsourcing is a powerful paradigm for human-powered problems solving on a wide level in areas such as picture classification, data input, object recognition, and editing. Crowdsourcing is founded on the premise that a huge number of people can give insight or value, even if some of them are incorrect [10, 11, 12, 13]. This technology is useful for solving complex tasks by delegating them to a wide pool of people. We researched studies that address the impact of interface design in a crowdsourcing context, to better comprehend the usability of the concept we intended to design.

Based on our research, we created ‘CrowdCorrector,' a concept that exercises the collective power of employees to detect biased terms in job adverts, and we used that prototype to provide us with insight into the effects of language usage. There was a coherent pattern of participants describing IT specialists as very demanding and intimidating in the findings in our pre-thesis. Although both advertisements feature soft skills, the soft skills were substantially more highlighted by participants in the nurse job advertisement when compared to the IT specialist.

The master thesis is heavily based on the findings on the pre-thesis, where existing literature and findings were taken into consideration. In this study, three conditions; Replacement, Suggestion, and Demographic, were researched in accordance with three job advertisements; Nurse, Business Consultant, IT Consultant in a WOz experiment [14]. Participants’ interactions were then linked as data to form a Thematic Code Analysis [15, 16], and then compared to The Five Burdens. As a result of the WOz experiment, it was visible that the replacement condition relies on the participants’ recall rather than recognition. It was also evident that participants enjoyed a certain impact in the suggestion condition, since they, on some level, had the opportunity to choose the wording of the job advertisement. In the demographic condition, there was a gap between our beliefs of fit for the participants' language versus their actual perception [17].

It gave us a better understanding of the areas that Crowdsourcing outperforms AI. Not only were we able to research the amount of gendered wording in job advertisements, how algorithmic prejudice is a burden to participants, and their preferences in the user interface. This study invokes a call in the HCI community to focus on the effect unconscious bias has on possible job seekers, when viewing a description of an advertisement.

Understanding UI in Crowdsourcing by Mitigating Unconscious Bias in Job Advertisement

Hind Morzogh
Aalborg University (AAU)
Aalborg, Denmark
hmorzo16@student.aau.dk

Ridwan Asad
Aalborg University (AAU)
Aalborg, Denmark
rmoham16@student.aau.dk

ABSTRACT

Social divisions regarding gender occupy an extensive role in inequality in workplaces. We believe that this inequality has a significant, albeit often unintended, effect. Job advertisements can often be very influential, and the choice of language can affect the candidate pool and has been shown to influence the likelihood that candidates will apply for a job. We examine the effect of gendered wording used in job advertisements for open positions (e.g., Nurse, IT Consultant, Business Consultant). A WOz technique [14] to conducted to examine nine participants' interactions and how each interaction has similarities, enables us to create a Thematic Analysis [15] that could relate to The Six Burdens [18].

AUTHORS KEYWORDS

Crowdsourcing; Unconcious bias; Job advertisments; The Six Burdens, HCI; AI.

1 INTRODUCTION

Despite progressive efforts, inequality between women and men in workplaces exists [4, 5]. Unconscious bias in the hiring process can have a significant impact on recruitment and candidate attraction [21]. In an experimental study [21], where fictitious resumes were sent for real open positions, 10.9% of male candidates were called in versus 7.7% of women. Social divisions regarding gender occupy an extensive role in inequality in workplaces. We believe that this inequality has a significant, albeit often unintended, impact on the recruitment and selection processes. This barrier can equally make it more difficult for recruiters to discover "the most qualified person" for the job.

The choice of language in job advertisements can affect the candidate pool and has been shown to influence the likelihood that candidates will apply for a job. According to statistics from a Hewlett Packard internal report [22], men feel they need to meet 60% of the characteristics of the qualifications when applying for a job, but women feel they need to meet 100%. Nonetheless, prior work shows that even when a job description lacks overt sexism or racism. The position is implicitly gendered through language and descriptions that specify the preferred gender of the ideal candidate. Likewise, in instances where hiring managers do not intend to discriminate against women, they may inadvertently reinforce gender stereotypes. An example of gender bias is that adjectives like ‘emotional’, ‘friendly’, and ‘caring’ are frequently associated with women [8, 9]. This can make it difficult for women to be perceived as leaders. However, it’s not solely women who are subject to gendered expectations. Men could also be less likely to apply for jobs that possessed feminine titles like “nanny” or “nurse [8, 9].

When someone’s gender is unknown, we tend to unconsciously associate them with a specific career field. Eagly’s theory [25] suggests that people will be more likely to assume an occupational role ascribed to their gender group. This unconscious bias steers us to consider others based on their gender instead of their skills or talents [25], [26], which ultimately limits the potential pool of job candidates. We believe that this gender stereotype is damaging because it limits how we perceive men and women, therefore also limiting the opportunities for both genders.

In this study, we examine the effect of gendered wording used in job advertisements for open positions. We designed a prototype “CrowdCorrector” to examine this effect and compare it to the five burdens existing AI technology has [18] by looking at three conditions; Replacement, Suggestion, and Demographic. Our study consisted of a semi-structured interview with nine participants where they were giving three job advertisements; Nurse, Business consultant and IT consultant. A WOz technique was conducted on the participants' interactions [14] and how it was linked to The Six Burdens, a study conducted by Park et. Al. [18] about the shortcomings of AI to understand if Crowdsourcing faced the same challenges as AI when reporting on gender bias in job advertisements. In our findings (view section 4), it was discovered that the condition replacement relies on the user’s recall rather than recognition. It was also noticed that the participants found most of the user interface elements in demographic useful, but could not identify where the annotated words were placed due to the UI. Ultimately, the user preferred the

suggestion condition because it gave them control to determine the context in which the job advertisement would have.

2 RELATED WORKS

Unconscious biases are often overlooked in organizations [2, 7, 8, 9], thus it becomes more critical to distinguish them as a stronger and more competent organization can be built. Accordingly, it is critical to recognize them with the goal that they can be eliminated, and our organizations can keep away from their hindering impacts. Building a concept that may recognize and mitigate unconscious bias, an element such as confidence when interacting is crucial [14, 15] because it assists individuals in overcoming risk and confusion.

Discrimination is difficult to prove since it is often contained in informal patterns of inequality in recruiting [14, 16, 17]. Furthermore, subjective experiences and perceptions of the hiring process, as well as the perception of discrimination, vary significantly amongst recruiters and job seekers. Obviously, the best candidate differs from one role to the next. Some professions and organizations, for example, follow a particular standard and level of education, while in others, it might be more necessary to have the "right attitude." [16, 18] There are also tacit, socially held ideas about the ideal worker, although they differ across industries. Covert discrimination is more malicious and systematic than subtle discrimination, which may be slight and/or unconscious.

The government Minister for Equalities, Lynne Featherstone, said in a recent speech at the launch of the Cranfield (2010) Female FTSE 100 reports that It's not that there aren't women out there [7]. Reporting has shown every year that there is an immense female talent pool. Many explorations have shown that the issue is caused by unconscious bias oblivious inclination – Where the higher you get in an organization, the more subjective the promotion measures become. In addition, when you permit a lot of subjectivity – Featherstone highlights that it is common to identify individuals employing staff who look and talk very much like them [7]. The problem is not only isolated to higher ranks in the organizations but accused in the recruitment process. She claims that sustainable culture change happens when you educate the organization, especially the middle managers who make numerous day-to-day choices [7] . These choices can be affected by unconscious biases and generalizing.

2.1 AI and Mitigating bias

Algorithmic decision-making is gaining popularity as a new source of HR recruiting and growth advice [3, 19, 20, 14]. The algorithms that power AI systems are described as mirrors that represent our research questions and data's unconscious biases. When an algorithm is improved on data that does not adequately reflect a population, or when the algorithm is constructed to optimize a single form of decision, prejudices may become rooted in the computer code. Bias also could arise when data scientists and software engineers fail to recognize the implications of their design decisions, as well as the larger social context in which the algorithmic decision-making results would be used [14, 15]. Johnson argues that the people who create the technology have the ability to influence how it functions, and that's much excessively powerful for any detailed demographic to control [12, 14]. A lack of diverse ideas and representation could exacerbate gender, race, and class divides.

Algorithmic decision-making in human resource management (HRM) is becoming more prevalent as a new source of information and advice, and it will be expected to expand in importance as organizations' digitalization accelerates [3, 16, 17, 21]. Automated decision-making and remote control, as well as standardization of routine workplace decisions, represent both examples of algorithmic decision-making. Algorithmic decision-making seems to be more rational and equitable than human decision-making at first glance. However, relying solely on algorithmic decision-making could lead to discrimination and unfair treatment.

Another study “Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens” by Park et al. [18] discuss how AI has currently been used to facilitate efficient decision-making in administrative and organizational contexts ranging from employment to dismissal. The article investigates the factors which exercise a negative influence on employees' views of AI when it comes to evaluating work performance. The study used scenario-based interviews to investigate this phenomenon. The modeled scenarios based on real-world use cases, Evaluation for Menial Work (Evaluation for Physical Work, Evaluation for Office Work, and Evaluation for Customer Service) were presented to the participants, in whom workers are supervised and evaluated by algorithms for job performance evaluations [18]. They found that participants experience six categories of burdens (emotional, mental, bias, manipulation, privacy, and social) as a result of AI's introduction to human resource management [18].

Bias burden

A total of 16 participants were worried that they would make mistakes or produce biased results due to insufficient data and algorithms. Notably, their fear was divided into two categories: (1) an inaccurate representation of reality and (2) algorithmic bias, in which the algorithm discriminates against minorities or causes biased results due to missing evidence. Some participants even found that their AI conducted racial bias towards the black participants.

Emotional burden

People frequently indicated a variety of negative feelings concerning algorithmic evaluations, which we classified into two sub-categories: "uncanny valley" and "inhumanity burdens." Six participants described their sentiments toward the AI, employing one of the phrases "odd, unusual, and alien" as an illustration of the uncanny valley [18].

Manipulation Burden

It was discovered that participants do not thoroughly comprehend AI's functions in terms of power, governance, and ownership. As a result, nine participants were concerned about the potential for manipulating AI processes [18].

Mental burden

Twelve participants expressed a cognitive burden as a result of having to guess, learn, and respond to unexpected AI. It was noted that the participants anthropomorphized AI and communicated with it like they would with humans. The participants expressed a significant number of mental burdens during this period because AI cannot communicate in the same way as humans do [18].

Privacy Burden

Fourteen participants raised privacy concerns regarding the collection and analysis of personal data such as CCTV, speech recordings, facial recognition, email scanning, etc. [18].

Social Burden

Fourteen participants were worried that AI in HRM could result in negative or unintended social changes in the workplace, such as unfair rivalry that undermines teamwork and relationships [18].

Fourteen participants were worried that AI in HRM could result in negative or unintended social changes in the workplace, such as unfair rivalry that undermines teamwork and relationships [18]. To define a potential job or social environment, some participants used the phrase "fake human relationship" to express their concern with the future.

Nineteen participants strongly agreed AI should clarify the reasons behind job success evaluation decisions. A participant asked, how could they accept the results without an idea why the AI executed those decisions. Another said that explaining why those tasks are a must-do, not a suggestion. Human-AI cooperation was preferred by 17 participants over human or machine-oriented assessment. We also discovered that people were aware of who takes the final decision in human-AI collaboration, and the outcome differed between participants [18]. Notably, they felt a bias burden in favor of an AI-based judgment and a coercion burden in favor of a human-based decision. Most participants desired to express their feelings (e.g., by compliments, motivation, or consolation) with "humans," rather than machines, particularly when they received negative outcomes. They felt AI empathy is fake and pointless [18].

2.2 Crowdsourcing

Crowdsourcing represents an effective tool for tapping into the collective insight of various demographic groups. Crowdsourcing is a new task allocation concept under which a task can be outsourced to a selected crowd rather than being handled by a single person [22, 25]. Crowdsourcing remains a thriving industry. Amazon's Mechanical Turk (MTurk) has over 800,000 tasks open for workers to complete; Upwork claims to have over twelve million workers and $1 billion in annual revenue, and 85% of the world's biggest corporations have used crowdsourcing in the last 10 years [22, 24]. MTurk is a labor marketplace that specializes in small piecework tasks and academic surveys . Tasks (moreover known as "personal intelligence tasks" or "HITs") are priced and published on an unregulated market by "requesters" [22, 23, 26]. Workers should go over the HITs and decide which ones they want to complete. The worker submits the job after the HIT is done, and the requester reviews it. The workers are compensated if the requester accepts the work; if the requester refuses it, the worker is unpaid [32].

3 METHOD

This section focuses on the methodology in which we applied to conduct and analyze the findings of the project. Using the within subjects method, we grouped the participants when testing the conditions, and used the order effect to understand how differently they perceive the conditions based on the order it was presented. We present three conditions; *Replacement*, *Suggestion* and *Demographic* which is researched through a WOz technique [14]. A thematic analysis [15] is then conducted based on the

WOz experiment and on a semi-structured interview. A short presentation of our participants' demographic is showcased in table 2.

3.1 Between-subjects study design

In HCI (Human Computer Interaction), when comparing several user interfaces in a single study, there are two methods of assigning the participants to these multiple conditions:

- "Within subjects"
- "Between subjects"

Which method to appropriately use, depends on the experiment that is conducted [27, 28]. *Within subject* involves each participant performing under all sets of conditions, whereas a *between subject* involves each of the participants only performing under one condition. For this experiment we found it appropriate to use within subject as all the participants will be comparing conditions (independent variable) [27, 29]. In this case, we wish to see how participants interact with "CrowdCorrector" by looking at three job advertisements with their own condition.

3.2 Wizard of Oz

Wizard of Oz (WOz) is a quick development design process intended to understand and improve the user experience. WOz technique necessitates the creation of a crude model of the finished product, known as a prototype [14]. The participants are presented with a prototype that displays biased words in advertisements, where the goal is to evaluate user perception of the different presentation techniques.

Participants are instructed to verbalize what they are doing when interacting with the different conditions, following the *thinking aloud* method. Thinking aloud is one of the most popular usability engineering methods [17]. When the participants verbalize their thoughts, it gives an overview of misconceptions they each might have, and therefore an understanding of their perception. Once participants have gone through testing the conditions, a debriefing will take place where we have prepared questions, based on a semi-structured interview [17].

3.3 Structure

The evaluation consists of three conditions. The first condition is a 'slide-show' that shows a preview of a word to take place instead and is merely a replacement (view illustration 1). The second condition is a drop-down menu with possible suggestions (view illustration 2). The third condition showcases the demographics of the users that have highlighted each word (view illustration 3).

Condition 1- Replacement

When the participant clicked on a word, we showcase a 'slide-show' with a replacement. When clicking on the replacement, we would replace the word with the chosen word, ultimately changing the wording of the job advertisement.

Illustration 1. Overview of our replacement feature, before and after review. The slide-show shows a short preview of the corrected word.

Condition 2 - Suggestion

When the participant clicks on a word, we showcase a drop-down menu with possible suggestions. When clicking on a suggestion, we would replace the word with the chosen suggestion, ultimately changing the wording of the job advertisement.

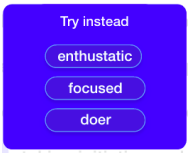

Illustration 2. An overview of the different suggestions of the condition presented.

Condition 3 - Demographic

When the participant clicks on the three dots, they are able to view demographics of the users who have highlighted each word.

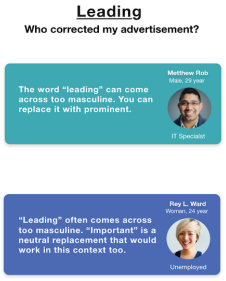

Illustration 3. An overview of the demographic interface of this condition. The corrector's age, gender, name and picture is displayed.

3.4 Order Effect

The essence of order effect is that the participants' performance may either improve or worsen because of the order of the test conditions. In some cases, the participants' performance will become worse if the tasks are predictable [27, 29]. When test conditions are assigned within-subjects and go through the same order of the tasks, they are most likely to be performance effects [28], caused by learning improvement that could lead to the participants performing better as the evaluation goes on. We could run into the risk that participants become familiar with the apparatus and

procedure. There might also be changes in attitude due to novelty effects, or it could happen due to fatigue [35].

Due to these risks, it is important to devise different tasks to reduce learning effects and also reduce boredom. The focus, in this case, is to evaluate the performance through conditions whilst minimizing confounding variables that would affect the accuracy of the experiment's results.

3.5 Latin Square Design

Due to the risk listed in the order effect section (view table 1), a systematic approach to variation is appropriate for this experiment. It will be conducted using the Latin Square Design [28] to administer the conditions in every possible sequence to different participants. A Latin square is a square grid where every element appears once in each row and each column. The rows represent the order in which test elements are administered to a participant, and a column represents the sequence of participants in the study [35].

Group	First task	Second task	Third task	Fourth task
i	Task A	Task B	Task C	Task D
ii	Task B	Task C	Task D	Task A
iii	Task C	Task D	Task A	Task B
iv	Task D	Task A	Task B	Task C

Table 1. Example of how to group participants and tasks they are given in an evaluation.

3.6 Qualitative Data Analysis

Qualitative researchers often take an inductive approach [2, 31], which means that they form a theory or investigate patterns of meaning based on the data that they have gathered. This entails moving from the specific to the general and is frequently referred to as a bottom-up approach [38]. A thematic analysis was conducted, which included relevant categories, topics, and relationships of the transcribed data [15].

3.7 Pilot Testing and Participants

To prevent mistakes and assure the quality of the experiment, a pilot testing takes place with a singular participant, that is in the same pool of the chosen participants. Pilot tests are meant to detect severe deficiencies in the test plan [17]. During this stage, we will know if the given instructions (if any) are incomprehensive or if there is a gap between the given tasks, in order to refine the procedure.

Nine participants were chosen based on their pre-existing experience seeing biased language in job advertisements. The participants were gathered from Facebook, LinkedIn, and through our personal network. It would be beneficial to include a participant from an HR department, who currently or previously, have worked with job advertisements. An overview of the participants' demographics can be seen in Table 2.

Participants	Gender	Age	Occupation
P1	Male	32	HR manager (YouSee)
P2	Female	22	HR consultant (YouSee)
P3	Female	23	HR consultant (YouSee)
P4	Male	29	Global Engineer (Microsoft) - HR experience
P5	Male	25	HR consultant (YouSee)
P6	Male	42	Product Designer (LEGO) - HR experience
P7	Female	33	UX Researcher (Vertica) - HR experience
P8	Female	29	Software Developer (ACTER) - HR experience
P9	Female	33	Product Owner (OJ Electronics) - HR experience

Table 2. Overview of our participants' demographic.

4 FINDINGS

Recall that we provided participants with three conditions: Replacement, Suggestion, and Demographic. AI is already deployed in the participants' workplace. To examine if one impacted them or more of the five burdens [18], with regard to dealing with the three conditions (view section 3), as they engage with the crowdsourcing concept. To investigate this, an analysis was conducted. The three occupations (nurse, IT Consultant, Business consultant) X three conditions(replacement, suggestion, demographic) were examined (view table 4) by adopting a hierarchical coding framework [15], to classify how each code relates to another. Furthermore, we involved the method within subjects [15] to eliminate the corresponding bias.

Condition	1 – Replacement	2 - Suggestion	3 - Demographic
User Interface	A pop-up menu with a substitute word appears.	A drop-down menu with possible choices appears.	The demographics of the users corrected each word.
Occupation	Nurse	IT Consultant	Business Consultant

Table 3. Overview of conditions.

Group	Task 1	Task 2	Task 3
Participants 1 - 4	1	2	3
Participants 4 -9	3	1	2

Table 4. Participants, order of conditions and fictional job advertisement.

4.1 Hierarchical Coding Frame

Following Braun and Clarke's [2, 33] thematic analysis approach to analyze our qualitative data. In order for us to do so, then the transcripts were analyzed thoroughly and withdrew categories that matched the statements, resulting in connecting themes. This was achieved by coding a set of the transcribed content and identifying themes throughout the process. This was an iterative process to discover participants' experience, views, and perceptions (e.g., benefits and concerns) of the three conditions. We have composed a point system that indicates the correlation between themes and statements, average link (+), Strong link (++), Weak link (-), and Poor link (- -). The following (illustration 4) is the themes extracted from the thematic coding analysis:

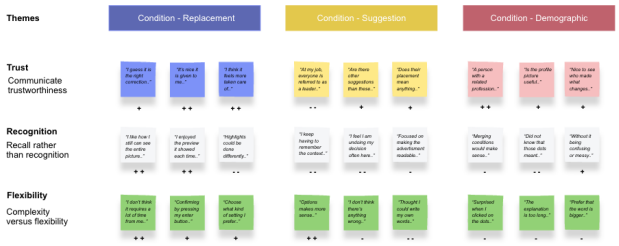

Illustration 4. Overview of our thematic analysis.

Trust

Trust in this context is how a user overcomes uncertainty when interacting with the three conditions. It consists of two different constructs: (1) system trust, e.g., how well the user trusts the system, and (2) interaction, e.g., how well the user's interaction fits with their intent.

Condition - Replacement

The study provided evidence that a few of the participants did not overcome uncertainty when engaging with the concept. When interacting with the concept in the condition with suggestions, there is a clear indication of the disconnect between participants' initial response from the system versus how the system responds. Meaning, that participants stated they were trusting the system, but their actions showed the opposite. When asked during the debriefing, several of the participants mentioned that they do trust its input, but they would prefer the ability to choose the sentence construction.

"I guess it is the right correction... but I feel like it's limited since I could not choose anything" - P2.

This statement was in contrast to a lesser number of participants who expressed they felt comfortable putting their trust in the corrections provided to them. This trust stems from their assumption that the system is designed to 'know' better with regard to neutral words.

"I doubt my own suggestion would be a hundred percent neutral, so it's nice it was given to me" - P6.

When reminded that the corrections came from representative end-users, the participant stated they felt more comfortable with the correctors.

"I think it feels more taken care of than it is by humans. I mean, it is specifically meant to be put in the right hands, right?" - P4

Condition - Suggestion

A larger number of the male participants in the given occupation, IT Consultant, associated more with the untouched advertisement, and therefore trusted it before it was corrected. This indicated that the majority of the male audience, related more to the masculine wording, and therefore felt no need for a change. This correlated to Eagly's theory [10], where people view an occupational role ascribed to their gender group. The statement confirmed the perception of belongingness.

"At my job, everyone is referred to as a leader. We do it to instill confidence in everyone and to make them realize that this company is also theirs. I didn't realize it was a gendered word." - P1.

Few of the statements correlated, although their concerns were directed in a slightly different direction. There were concerns regarding the quality of the corrections, ultimately affecting trust.

"Are there other suggestions than just these? I feel as if there are better words to use here" - P8.

Although the participants liked the suggestions, there was some confusion when it came to the hierarchy. This indicated there was an uncertainty as to which suggestion was highly favored compared to the others.

"I like that there are more choices, but does their placement mean that they are more useful than the others? Is there a voting system that determines that?" - P7.

Condition - Demographic

When users are able to get a visible picture of whom the correctors were, they felt as if their profession was valued.

"I think the corrections are more fitting if they're made by a person with a related profession… or similar" - P5.

Whilst a number of participants agreed (P2, P5), there were a few that expressed the lack of necessity for a profile picture.

"I don't know if the profile picture is useful? I am fine with looking at other information about the person correcting my advertisement" - P9.

However, it is important to emphasize that the majority of the participants enjoyed the demographic overview of the correctors.

"If I were looking for specific people, it'd be nice to see who made what changes, right?" - P4.

Recognition

Recognition in this context disguises between two types of memory retrieval; recognition versus recall. It focuses on how well the user recognizes the elements in the three conditions, such as familiarity versus recall, which requires the recollection of a memory.

Condition - Replacement

As seen in the previous condition, the replacement does have a blur, but it is at a lower opacity and the context is therefore visual without taking the focus of the task. This was heard with positive encouragement by the participants when they did not express confusion.

"I can't really see well, to begin with, so I like how I still can see the entire picture" - P6.

As viewed in illustration 1, the proposed word set to be the replacement is presented by sliding the cursor over highlighted words. This was a pleasant element according to the majority of the participants (P1, P8, P2) since it reduced the information that each user had to remember.

"This interface was very limited, but I enjoyed the preview it showed each time" - P8.

Some participants (P6, P8, P7) stated they would have preferred some explanatory text to what the highlights meant. Uncertainty of whether the word has been corrected was again expressed in this context, and the desire for a quick recognition was preferred.

"Maybe it is just me, but I feel as if the highlights could be done differently" - P6.

Condition - Suggestion

It was clear from several of the participants (P1, P4, P9) that they relied on their memory in the condition with replacement. When they clicked on a replacement, they would be met with a blurred-out background whilst the word was highlighted. Despite the fact that it was not directly stated by each participant, many of them had to undo an interaction to recollect the whole sentence.

"I keep having to remember what the context was." - P1.

Despite the fact that it was not directly stated by all participants, many of them had to undo an interaction to recollect the whole sentence.

"I feel like I am undoing my decision quite often here." - P4.

There was a mention of having to be careful when selecting a certain suggestion. This is again related to the overall context, but more specifically, whether the description makes sense when selecting certain suggestions.

"I was more confident in selecting the first two suggestions, but when I moved on to the rest it felt more difficult since I was more focused on making the advertisement readable" - P9.

Condition - Demographic

Participants expressed that despite the fact that this condition was not met with enthusiasm, several of them recommended that it be merged with the suggestion condition.

"Merging those conditions would make sense, maybe it is something to consider." - P1.

The three dots required some cognitive effort from the participants since it was a feature that they did not have seen before.

"I actually did not know that those dots meant, I don't think I have seen it before."- P4.

Despite their initial confusion at the three dots as a highlight for an incorrect word, they did like the overall design.

"I like how I am able to see every corrector without it being confusing or messy... I was actually expecting that." - P1.

Flexibility

Flexibility in this context focuses on how the three conditions cater to users' perception of different methods to accomplish the same task, and if it is fit to their preference. It is also looking at accelerators, e.g., shortcuts, that speed up their interactions.

Condition - Replacement

All participants were novice users of the prototype, but the initial idea was that there was no need for step-by-step guides since the prototype merely required a click on a certain word. This fast alternate method was especially used in the replacement condition, since the interaction consisted of a singular step, clicking on a corrected word. The replacement condition is less time-consuming to novice users.

"I don't think it requires a lot of time from me, other than me clicking the word" - P3.

Some users were used to confirming a selection by pressing the enter button. This shortcut was not implemented, however, it was specifically mentioned in the replacement condition.

"I am used to confirming by pressing my enter button... I know it is a minor issue, but I prefer it" - P7.

This issue was again brought up, mentioning the lack of personalization with regard to tailoring functionality for users.

"I am used to using applications where I can choose what kind of setting I prefer, but I don't think I can do the same here." - P1.

Condition - Suggestion

This condition had more flexibility since there were a variety of words to choose from. This gave the participants the flexibility to decide if the suggestion fit the context. This allowed them to customize the outcome of the job advertisement.

"Having several options makes more sense. I like to choose how I want the job advertisement to be in the end" - P2.

However, it did limit them to a set of suggestions that some participants did not agree with. There was some uncertainty of *how* flexible they are in their interactions, questioning if they are able to ignore a suggestion.

"Am I able to stick with a word? I don't think there's anything wrong with the word." - P5.

Some participants wanted to see if they could type in a suggestion themselves. This preference is related to customization and indicates a need for tailoring content.

"I really thought I could write my own words instead of having to choose a suggestion." - P9.

Condition - Demographic

When transitioning between conditions, it was clear that there was some time to adjust to the new response the system gave as an output to the users. In terms of demographics, several participants were unsure if the corrections had been made. Although this factor relates to trust, it is also linked to flexibility and efficiency of use, since it might have required guidance in order to speed up their actions.

"I was surprised when I clicked on the dots, I thought the words needed to be corrected" - P2.

Several (P3, P2, P7) mentioned that it would be time-consuming to read their entire comment and add the correction.

"It's great to know who's fixing it, but the explanation is too long." - P7.

This condition did highlight the incorrect word as a title, which made it easier for the participants to figure out where in the text they were.

"I do prefer that the word is bigger and not on top of the text." - P3.

4.2 The Six Burdens

The evaluations were analyzed through the lens of the six burdens to better assess whether the concept addressed that issue raised by Park in the research [18]. It was clear that two of the six burdens were not brought up when interacting with the three conditions; social and privacy burdens. They will therefore not be included.

Bias Burden

Biased suggestions were a concern to multiple participants . This was due to the corrector's own bias, which made participants question whether or not they favored their own social group, causing them to discriminate harshly on certain words.

"I can't believe that these corrections are unbiased. If the majority of people solely chose one word to be 'unbiased' over another, I would believe it, but when I have different opinions of what it is supposed to be there.. It does not seem unbiased" - P1

When interacting with the demographic condition, P9 said they enjoyed seeing information about the correctors, however, he felt the need for age was unnecessary.

"I understand that there is a bias against minorities and women, but why is age a factor? How does knowing that make me trust their corrections more or less?" - P9

Emotional Burden

In the study by Park et al. the effect of algorithmic evaluations [18], has been referred to as the "uncanny valley" and "inhumanity burdens". Not only did conditions 2 and 3 provide participants the autonomy to make cognitive decisions based on presented information (ex choices), but they also understood that users were behind the decision-making, breaking down the alien feeling individuals have towards AI.

"It has a lot of potential; the user aspect takes it to the next level, and I can see a market for it. It's a good idea to have others check over your shoulder before sending it out." - P1

When the concept "resembled" the stereotyped elements of AI, there was a great dissatisfaction in the concept, which was strictly objective with no space for objections, as in condition 1, which left the participant with just one replacement and no option to dismiss it. Several of the participants felt that their ability to make a cognitive decision was taken and that their own expertise was not valued in the user interface.

Manipulation Burden

Condition 3 visualized the users who corrected the job advertisements position, thus highlighting their expertise.

The choices of participants seemed to be affected by the users' positions, creating a connection between position and influence. There seems to be a correlation with the theme *trust* , confirming its validity.

"I mean, choosing a correction is quite easy when you see that a person is a manager. I would believe they know more about the importance of keeping the company they work for neutral" - P4.

This was not only contradicted by the following statement, but made an inverse correlation with the theme of *trust.*

"I don't know if the profile picture is useful? I am fine with looking at other information about the person correcting my advertisement" - P3.

Several participants placed a strong emphasis on making a connection between the corrector's profession and their correction, rather than how their gender influenced their choice. This critique was internalized by the same participants who gave that critique. A participant specifically pointed out that they were not representative presented job advertisements.

"I don't know anything about the position of IT Consultant, but I feel as if a person who has not only studied IT would be a good fit, but also someone who looks like... they know what they are talking about" - P1.

Mental Burden

The participants also overcame a significant mental burden in condition 1 since the concept pushed one correct outcome of the job advertisement, even if they disagreed. This prompted the participants to ask us questions regarding the options and the evidence that supported them. This user interface omitted the logic for the selection of the word that was chosen to be a substitute, leading to misunderstanding and guessing its intentions. Participants highlighted that they felt that this condition limited their ability to reject the replacement.

"I wish I could ignore the suggestions, I feel like it is locking me in" - P2.

Participants valued their ability to make cognitive decisions, and they stated that condition 1 did not allow for this. Emphasizing in the design that those correcting, as important as they might seem, still should leave room for the participants to disagree with the suggestion was important to most of the participants.

For several of the participants, Condition 3 confused them due to a crucial element design of the interface. The three dots under the highlighted correction was new to most of them. It was not a feature that was recognizable for the participants. This correlates to the theme *recognition* since it was not recognizable for several participants (P1, P5, P9).

The participants suggested a red line when highlighting a mistake, a feature already popularized by software programs such as Grammerly.

"Are they dots there because they look pretty? I would prefer a red line. I would get that right away, and it would make me understand that something needs my attention." - P8.

5 DISCUSSION

Throughout this study, our goal is to understand how users perceive a crowdsourcing platform that neutralizes job advertisement through three interfaces, which we define as conditions. Our work is based on a call to tackle unconscious bias [9, 11, 14, 15]by viewing informal patterns of inequality in job advertisements. We explore the effect of gendered words (linguistic bias) and provide information on users' perception of three conditions (replacement, suggestion, demographic). For that reason, a perspicacious view of existing AI technology in employment advertising is crucial to gaining a better understanding of the unconscious biases that are driving the rise of AI in employment [19, 20].

5.1 Information Presentation

The perception of the three conditions was summarized into three themes in the thematic analysis [15], showing which was strongly preferred by the participants. We found that the replacement condition relies on the user's recall rather than recognition, since the blurred background made it difficult to view in which context the highlighted word was placed. Some participants had to either ask us for the context again or go back and look for it on their own, therefore spending more time in the same stage, reducing the overall flexibility of the concept.

Many of them associate the replacement condition with other software programs, such as Grammarly, prompting familiarizes. The participants valued that they had influence in how the outcome of the job advertisements would turn out, by choosing their own suggestion. A significant statement was the merging of two conditions; replacement and demographic.

A crucial observation seen in the demographic condition, is the dots on top of each word which invoked uncertainty to the participants. During the evaluation, participants asked if the three dots were clickable, and some even looked past them, prompting them to direct us to the subject of the question. It was visible there was a gap between our beliefs of fit for the participants' language versus their actual perception. Surprisingly, the participants were disinterested in some elements of the characteristics, such as age or gender, and were instead focused on the professions of the

correctors. The participants emphasized their title would make them more fit to correct the advertisement.

5.2 AI vs. Crowdsourcing

Currently, employment advertising is edited by AI algorithms, which have an inherent bias that is seen in the amount of gendered terms in each job advertisement. It has been demonstrated that these facts are intimate to job applicants. Throughout this section, we will parallel the study *Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens* by Park et al. [18], with our findings to better understand the areas that crowdsourcing outperform AI.

- **Performance of AI**. AI has progressed to the point where it can now match, if not outperform, human performance on many cognition-related activities. AI has the potential to drastically reduce costs for industry and companies by lowering paid employment of humans. This has caused concern about job losses [31], [39], which may explain why participants in the Park et al. research [18] were worried of the AI technology. Several of the participants in that study alluded to a fear of losing their job if the AI misjudged them and their performance. Participants from a minority race also feared the consequences due to insufficient data collection. However, the amount of *threat* that has been associated with AI is not a factor in crowdsourcing as a technology deals with [3, 34].
- **Algorithmic prejudice**. In terms of algorithmic prejudice, five individuals felt burdened by the probable discriminations made by AI. The media has revealed the evident racial biases that AI algorithms can have, leading to discrimination against black people and women since it learned biased data from humans. Participating in such research has the potential to impact the participant's attitude, which, if it did, can lead to bias in trial outcomes.
- **Five burdens in crowdsourcing**. Crowdsourcing has not been politicized, making the participants not vary of the technology and quickly embracing the possibilities. This could explain why the privacy and social burden did not apply in any of the conditions, and the rest of the burden applied to a lesser degree [18].

5.3 Limitations

We recognize there are various limitations when interpreting our findings. First, in our approach to Latin Square Design, we should have given every job advertisement, every condition on separate occasions. Future work might assess the effect of this certain approach. It would've given us a better understanding of the participants who were given the same critiques on the different conditions, and not because they had a less favorable opinion of the job advertisement. Several of the participants did not only give criticism to the interactive elements in the three conditions but started to give overall criticism of the structure of the prototype until we directed them back to the task. We emphasized that the focus was the highlighted words, and not the overall, and the user experience. Secondly, it could have been beneficial to have implemented a statistical feature that could showcase a percentage of how many 'votes' each suggestion has received from correctors. This approach could have influenced the participants to choose a certain suggestion based on, e.g., a high percentage. In the condition replacement, the participants seemed very grounded in their own judgment, but it would have been interesting to observe if they had questioned it if a percentage were present. Lastly, we emphasize that cultural perception of bias was present and that had an influence on the overall data. Our findings are likely to vary when involving different participants from around the world. Our study consisted of Danish participants, which could have had an influence on how our English job advertisements were perceived. A word that could have inadvertently reinforced gender stereotypes, might have been interpreted differently since it was not in the user's native language. A French participant mentioned that their entire language was based on gendered words, and his relationship to gender bias did not feel problematic. For future work, it would be beneficial to take that aspect into consideration.

6 CONCLUSION

In this research, we examine the effect of gendered wording used in job advertisements for open positions. We designed a prototype CrowdCorrector to examine this effect and compare it to the five burdens. The prototype was tested through three conditions; replacement, suggestion and demographic condition as the interface design. Specifically, we looked at how CrowdCorrecter could be used as a tool viewing gendered words in three conditions, researching how well users' perceptions fit with the system's response. We also researched how crowdsourcing is used for dissecting and distributing linguistic tasks to an online worker community, where an organization presents a challenge to an online community in the context of a "peer-vetted" approach. In the replacement condition, the participants did not favor

their inability to have any control without disregarding the input of the concept. Also, the participants in the suggestion condition valued their ability to make cognitive decisions. Furthermore, the findings in the demographic condition showed the choices of participants seemed to be affected by the users' positions, creating a connection between position and influence. This clearly demonstrates that the participants prefer collaboration to autonomy.

ACKNOWLEDGMENTS

We would like to thank the participants who contributed to the interviews, and were patient with us during the technical difficulties. Their input has bought us valuable context and made us reflect on the course of the project. We would also like to thank our supervisor Niels, who has always given us the right guidelines on the topic of Crowdsourcing and unconscious bias.

REFERENCE

[1] K. B. Noland, Marcus, Moran Tyler, "Is Gender Diversity Profitable? Evidence from a Global Survey," 2016. .

[2] C. H. Neesha-ann Longdon, Dimitri Henry, "Diversity And Inclusion As A Social Imperative," *S&P global ratings*, Aug-.

[3] S. K. White, "How top tech companies are addressing diversity and inclusion," *CIO United STATES*, 2020. [Online]. Available: https://fortune.com/2020/06/16/pepsi-ceo-ramon-laguarta-black-lives-matter-diversity-and-inclusion-systemic-racism-in-business/.

[4] R. LAGUARTA, "PepsiCo CEO: 'Black Lives Matter, to our company and to me.' What the food and beverage giant will do next," *Fortune*, 2020. .

[5] W. J. Casper, J. H. Wayne, and J. G. Manegold, "Who Will We Recruit? Targeting Deep- and Surface-Level Diversity with Human Resource Policy Advertising," *Hum. Resour. Manage.*, vol. 52, no. 3, pp. 311–332, May 2013.

[6] S. Robinson, "Rethinking recruitment in policing in Australia: Can the continued use of masculinised recruitment tests and pass standards that limit the number of women be justified?," *Salus J.*, vol. 3, pp. 34–56, 2015.

[7] G. Beattie and P. Johnson, "Possible unconscious bias in recruitment and promotion and the need to promote equality," *Perspect. Policy Pract. High. Educ.*, vol. 16, no. 1, pp. 7–13, Jan. 2012.

[8] A. Köchling and M. C. Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development," *Bus. Res.*, vol. 13, no. 3, pp. 795–848, Nov. 2020.

[9] T. Koivunen, H. Ylöstalo, and K. Otonkorpi-Lehtoranta, "Informal Practices of Inequality in Recruitment in Finland," *Nord. J. Work. Life Stud.*, vol. 5, no. 3, p. 3, Oct. 2015.

[10] M. Lease, "On Quality Control and Machine Learning in Crowdsourcing."

[11] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide Web," *Commun. ACM*, vol. 54, no. 4, pp. 86–96, Apr. 2011.

[12] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality Control in Crowdsourcing Systems: Issues and Directions," *IEEE Internet Comput.*, vol. 17, no. 2, pp. 76–81, Mar. 2013.

[13] F. E. Geiger David, Rosemann Micheal, "Crowdsourcing Information Systems - A Systems Theory Perspective," 2011.

[14] A. Weiss, R. Bernhaupt, D. Schwaiger, M. Altmaninger, R. Buchner, and M. Tscheligi, "User experience evaluation with a Wizard of Oz approach: Technical and methodological considerations," in *2009 9th IEEE-RAS International Conference on Humanoid Robots*, 2009, pp. 303–308.

[15] M. Vaismoradi, J. Jones, H. Turunen, and S. Snelgrove, "Theme development in qualitative content analysis and thematic analysis," *J. Nurs. Educ. Pract.*, vol. 6, no. 5, Jan. 2016.

[16] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006.

[17] J. Nielsen, *Usability Engineering*. Elsevier, 1993.

[18] H. Park, D. Ahn, K. Hosanagar, and J. Lee, "Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.

[19] S. Tang, X. Zhang, J. Cryan, M. J. Metzger, H. Zheng, and B. Y. Zhao, "Gender Bias in the Job Market," *Proc. ACM Human-Computer Interact.*, vol. 1, no. CSCW, pp. 1–19, Dec. 2017.

[20] I. L. Organization, "Gender Inequality and Women in the US Labor Force," 2011.

[21] U. P. Fabra, "Women are 30 percent less likely to be considered for a hiring process than men," 2019.

[22] J. Zenger, “The Confidence Gap In Men And Women: Why It Matters And How To Overcome It,” *Forbes*, 2018. .

[23] C. J. Beukeboom and C. Burgers, “Linguistic Bias,” in *Oxford Research Encyclopedia of Communication*, Oxford University Press, 2017.

[24] C. J. Beukeboom, “Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies.,” 2014.

[25] A. H. Eagly and W. Wood, “Social Role Theory,” in *Handbook of Theories of Social Psychology*, 1 Oliver’s Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, pp. 458–476.

[26] T. O. S. University, “UNDERSTANDING IMPLICIT BIAS.”

[27] M. van Ryn and S. Saha, “Exploring Unconscious Bias in Disparities Research and Medical Education,” *JAMA*, vol. 306, no. 9, Sep. 2011.

[28] U. of Bristol, “Guidance to Unconscious Bias at Shortlisting and Interview.”

[29] W. Leung *et al.*, “Race, Gender and Beauty: The Effect of Information Provision on Online Hiring Biases,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–11.

[30] L. Yarger, F. Cobb Payton, and B. Neupane, “Algorithmic equity in the hiring of underrepresented IT job candidates,” *Online Inf. Rev.*, vol. 44, no. 2, pp. 383–395, Dec. 2019.

[31] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker, “How We’ve Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis,” *Proc. ACM Human-Computer Interact.*, vol. 4, no. CSCW1, pp. 1–35, May 2020.

[32] S. à Campo, V.-J. Khan, K. Papangelis, and P. Markopoulos, “Community heuristics for user interface evaluation of crowdsourcing platforms,” *Futur. Gener. Comput. Syst.*, vol. 95, pp. 775–789, Jun. 2019.

[33] J. Oppenlaender, N. Shireen, M. Mackeprang, H. Erhan, J. Goncalves, and S. Hosio, “Crowd-powered Interfaces for Creative Design Thinking,” in *Proceedings of the 2019 on Creativity and Cognition*, 2019, pp. 722–729.

[34] R. Budiu, “Between-Subjects vs. Within-Subjects Study Design,” *NNGroup*, 2018.

[35] I. S. MacKenzie, “Within-subjects vs. Between-subjects Designs: Which to Use?,” *Dept. Comput. Sci.*, 2013.

[36] G. Charness, U. Gneezy, and M. A. Kuhn, “Experimental methods: Between-subject and within-subject design,” *J. Econ. Behav. Organ.*, vol. 81, no. 1, pp. 1–8, Jan. 2012.

[37] A. Madill, “Interaction in the Semi-Structured Interview: A Comparative Analysis of the Use of and Response to Indirect Complaints,” *Qual. Res. Psychol.*, vol. 8, no. 4, pp. 333–353, Oct. 2011.

[38] E. Wynn and H. V. Hult, “Qualitative and Critical Research in Information Systems and Human Computer Interaction: Divergent and Convergent Paths,” *Found. Trends® Inf. Syst.*, vol. 3, no. 1–2, pp. 1–233, 2019.

[39] E. Awad, S. Dsouza, J.-F. Bonnefon, A. Shariff, and I. Rahwan, “Crowdsourcing moral machines,” *Commun. ACM*, vol. 63, no. 3, pp. 48–55, Feb. 2020.