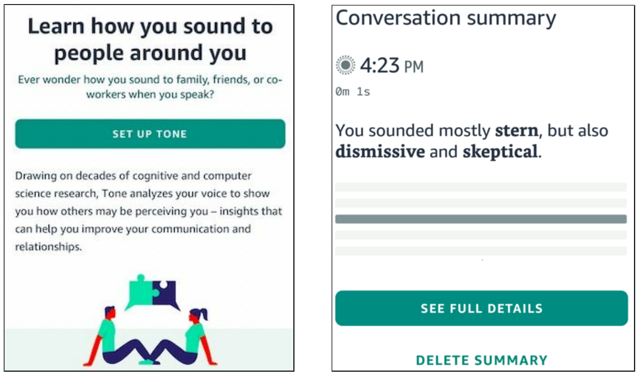
Gender Biases in Tone Analysis: A Case Study of a Commercial Wearable Device

ABSTRACT

In addition to being a health and fitness band, the *Amazon Halo* offers users information about how their voices sound, i.e., their ‘tones’. The Halo’s tone analysis capability leverages machine learning, which can lead to potentially biased inferences. We develop an auditing framework to evaluate the Amazon Halo’s tone analysis capabilities for gender biases. Our results show that Halo exhibits statistically significant gender biases, when the same emotion is conveyed by professional women and men actors through their recorded voices. For example, we find that over 75% of the words used by Halo to describe men’s emotions are positive whereas fewer than 50% of the words used by the Halo to describe women’s voices are positive. The Halo describes women as being ‘angry’, ‘disappointed’, ‘uncomfortable’, and ‘annoyed’ more often than men (adjectives with negative valence). The Halo describes men as being ‘knowledgeable’, ‘confident’, and ‘focused’ more often than women (adjectives with positive valence). Overall, our findings underscore that even commercially deployed ML models for day-to-day consumer use exhibit strong biases.

ACM Reference Format:
. 2023. Gender Biases in Tone Analysis: A Case Study of a Commercial Wearable Device. In *Proceedings of ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO ’23).* ACM, New York, NY, USA, 19 pages. https://doi.org/XXXXXXX.XXXXXXX



Learn how you sound to people around you

Ever wonder how you sound to family, friends, or co-workers when you speak?

Drawing on decades of cognitive and computer science research, Tone analyzes your voice to show you how others may be perceiving you – insights that can help you improve your communication and relationships.

Conversation summary

4:23 PM

0m 1s

You sounded mostly **stern**, but also **dismissive** and **skeptical**.

(a) A page seen during the Tone Analysis setup.

(b) An example of Amazon Halo’s Tone Analysis output.

Figure 1: Screenshots from the Amazon Halo’s smartphone companion app, taken August 18, 2022.

1 INTRODUCTION

As machine learning models become more prevalent in people’s daily lives, researchers have raised concerns about algorithmic fairness and the biases that these systems can perpetuate, e.g., [24, 27, 32, 34]. For instance, prior work has demonstrated how algorithmic decisions have resulted in racial [29] and gender inequities [18], and exacerbated marginalization of underrepresented communities [40].

Toward building a future in which machine learning algorithms are equitable and inclusive, it is important to understand how real system — including commercially deployed systems — manifest biases and inequities. In this paper, we focus on the *Amazon Halo*, a wearable device that, in addition to fitness tracking abilities, uses machine learning-based modeling to assess a user’s tone of voice. Amazon first introduced the Halo in December 2020 and sold the device through April 2023. As seen in Figure 1a and other promotional materials [1], Amazon marketed this ‘Tone Analysis’ feature in Halo as a way users can understand how they sound to people around them and use the output to help improve their speech. Figure 1b shows an example of one of the results the Halo shows to users. After analyzing user’s voices, Halo reports results by showing users three adjectives that it believes best describes how the user sounds, such as (as in Figure 1b) ‘stern,’ ‘dismissive,’ and ‘skeptical’ (their ‘tone’).

Motivated by prior research that explores gendered bias in machine learning-based algorithms, e.g., [12], as well as work demonstrating societal biases in how people perceive women and men’s voices, e.g., [25], we designed an auditing system to evaluate the Halo. We want to know whether there are gendered biases in the way the Tone Analysis reports men and women’s voices.[1] We capture this aspect of our investigation in research question 1 below:

- **RQ1:** Does the Amazon Halo exhibit gender biases in the tones it ascribes to the voices it hears?

We believe that RQ1 is important to ask for at least two reasons. First, if the Halo is perpetuating existing societal biases or creating new biases in the way it reports people’s voices, it could negatively impact people who are using it to evaluate the way they sound and might cause them to unnecessarily change their behavior.

Second, auditing the Halo for gendered bias gives us an opportunity to examine how large companies (in this case, Amazon), deploy machine learning-based models in commercially available products such as the Halo. This consideration is particularly relevant as companies continue to expand into the health and wellness industry [2, 4] and use potentially sensitive biometric information to create personalized content for users. Through scientific analyses such as ours, the research community can contribute to increasing

EAAMO ’23, Oct 30 – Nov 1, 2023, Boston, MA
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/23/10…$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1] While gender is a spectrum, foreshadowing our data analysis methodology and the existing, curated RAVDESS dataset [31], our experimental analyses focus on binary genders: women and men.

accountability and transparency to the commercial deployment of machine learning systems.

To examine the Halo and answer RQ1, we leverage RAVDESS, an existing audio dataset of 24 actors: 12 women and 12 men, who each repeat the same two phrases twice, in eight different emotions, for a total of 768 audio clips. The RAVDESS dataset comes from the psychology community and was curated and validated for use in scientific analyses of voices and tones. We build an automated testing environment to play audio clips to a Halo and record Halo's output. We then analyze the *valence* — the positivity or negativity of a word — associated with the adjectives Halo uses to describe voices and we assess whether the Halo reports women and men's voices (their tones) in different ways.[2]

Foreshadowing our results, we find that the answer to RQ1 is yes, the Halo *does* exhibit gender biases in its tone output and, in fact, the differences are quite substantial. As an example of our results, over 75% of the words used by the Halo to describe men's emotions had positive valence (over all our experimental runs); in contrast, fewer than 50% of the words used by the Halo to describe women's voices had positive valence (see Section 4.2.1).

Not only do we believe it important to know *whether* gender biases exist in the Halo tone output, per RQ1, we believe that it is important to know *how* those biases manifest. As context, there are strong historical and societal stereotypes and harms associated with interplay between genders and emotions. Consider, for example, that researchers found that people preferred lower-pitched, 'masculine' voices for those in leadership roles [11, 25]. As another example, studies have found that professional women perceived as angry are conferred lower status than angry men [16], and that women who are assertive in the workplace receive backlash [10]. Thus, while exploring the psychological impact of the Halo's gendered outputs is outside the scope of this work, we believe it important to know which tones the Halo disproportionately ascribes to women vs men. This leads to our second research question:

- **RQ2:** Of the Halo's outputs, which tones are more associated with women than with men? And which tones are more associated with men than women?

Sampling from the Halo's outputs with the lowest valence, we find that the Halo is significantly more likely to label women as 'disappointed' and 'annoyed' than men (respectively, 33 vs 1 times, 60 vs 2 times). And, the Halo simply doesn't use other words, such as 'angry' or 'uncomfortable' at all to describe mens' voices despite using them 17 and 46 times respectively to label women. Sampling from the Halo's outputs with the highest valence, we see find that the Halo is significantly more likely to label men as 'knowledgeable,' 'confident,' and 'focused' than women (respectively 531 vs 240 times, 752 vs 501 times, 561 vs 312 times).

Lastly, recall that our source dataset of 768 audio recordings from professional actors has those actors speaking with 8 source emotions. While our analysis of RQ1 and RQ2 were independent of the source emotion, we next ask:

- **RQ3:** Of the eight source emotions in the professional actor voice dataset, are there some for which the gender biases are particularly strong?

We believe that these findings are not just of academic interest. While a study of the psychological impacts of the Halo's output on women and men is outside the scope of this work, as noted above, we believe that there *could* be negative impacts (especially on women). Such negative impacts are not pure conjecture. They are supported by existing literature, e.g., [15, 17], and all authors have experienced or observed the interplay between gender and the words used to describe emotional tone. The fact that the Halo is now no longer commercially available — Amazon announced that it is shutting down the Halo program in the midst of widespread tech industry layoffs in late April 2023 (shortly before the submission of this paper, reportedly due to "an increasingly crowded segment and an uncertain economic environment") [42] — does not, in our minds, diminish the significance of our findings. The Halo *was* a commercially deployed system and it presumably represented what a company (Amazon) believed to be state-of-the-art engineering. And, at the time of this writing, although the Halo is no longer being sold, those who purchased the product can and likely are still using it. Thus, we consider our investigation a contribution to a growing body of work calling attention to machine learning biases, especially in deployed commercial products.

2 BACKGROUND

Biases in Algorithms. It is by now well-known that biases exist in machine learning algorithms, and in computing systems in general, and that these biases can cause harms. In foundational work, Buolamwini and Gebru [18] showed that image-based gender classification products had much higher error rates for women with darker skin. Imana et al. [23] found that Facebook ad algorithms withheld certain job opportunities from women. Recent work by Wolfe et al. [46] found that a synthetic image generator generated sexualized imagery significantly more for the prompt 'a 17 year old girl' and than 'a 17 year old boy'. These are just some examples of the harmful biases in ML systems; see Srinivasan and Chandler [43] for a survey of additional findings, as well as Mehrabi et al. for a taxonomy of bias in algorithms [32]. Our work adds to this body of knowledge by examining what is to our knowledge a new modality: the potential bias in algorithms that use a person's voice to describe their emotions. Further, we do so in the context of a commercial product: the Amazon Halo.

Baises in (Western) Society. In some cases, e.g., as with Buolamwini and Gebru's results with face recognition systems [18], the biases result from an inequitable representation of people in the machine learning model's training process. Indeed, the creation of more diverse datasets resulted in a decrease in the racial biases of face recognition-based systems [38]. In other cases, the biases in algorithms may perpetuate existing biases and stereotypes within society. For example, as foundational work, Caliskan et al. [19] found that a language model trained on web content would contain recoverable and accurate imprints of historic societal biases.

While it is not our goal to compare the Halo's biases (if any) with societal biases, it is important to acknowledge that societal gender biases exist. Prior work has examined, for example, societal stereotypes that men exhibit more competence, while women exhibit more warmth [20]. Further, the existence of these biases can cause real harm. Existing work has also examined the impact of these

[2] As examples of valence, the word 'confident' has valence 7.56 and is considered 'positive' and the word 'stern' has valence 3.9 and is considered 'negative' [45].

stereotypes, from internal self-perception [21], to job opportunities and career growth [13, 14, 30, 33, 41]. As a concrete example, researchers found that gender stereotypes of emotion can lead to biased evaluation of women leaders [15]. Thus, whether the Halo perpetuates existing societal biases or creates new one, the results of biases in its output can lead to harms. It is thus imperative to understand what biases the Halo, and by inference possible future voice-based tone analysis system might manifest.

Emotional Valence. In this work, we analyze the Halo's output when it analyzes different voices. This output comes in the form of three emotion words, as shown in Figure 1b. To analyze whether the words output by Halo are positive or negative, rather than relying on our own interpretation of these words, we rely on lexicon-based sentiment analysis. More specifically, we draw on valence values, first established as part of the valence-arousal-dominance framework in psychology for analyzing emotions [39]. Valence values describe the positive or negative meaning of a word: positive words are associated with higher valence, and negative words with a lower valence. Prior work in computer science has used this framework for sentiment analysis in several fields, including natural language processing [22, 28, 35]. For example, Preoţiuc-Pietro et al. used a valence-arousal approach to analyze the sentiment of Facebook posts [37]. And, more recently, Alonso et al. examined how sentiment analysis can be used to detect misinformation [9]. In our work, we rely on a corpus of valence values for emotion words (e.g., 'disgust', 'fearful', 'calm') from Warriner et al. [45] that have been crowdsourced from 1,827 people.

3 METHODOLOGY

The focus of our research is to understand whether the Halo describes men's and women's voices using different adjectives. This section describes the data we used to analyze the Halo, how we played clips of audio to the device, and how we collected the corresponding output.

3.1 Data Set

3.1.1 Preliminary Experiments: Word Choice of Audio Clips. Prior to testing the Halo in an automated fashion, we first conducted preliminary experiments to familiarize ourselves with factors that might affect the Halo's output. According to anecdotal news reports, the words a user says might impact the words Halo uses to report their voice [8, 26]. For example, a user that says something negative, regardless of their tone, would receive correspondingly negative feedback from the device.

In order to test the role of spoken words, we selected two text-to-speech voices from Microsoft Azure [5], one 'masculine,' labeled 'Christopher,' and one 'feminine,' labeled 'Jenny'. With both, we repeated the same sentence with the following format: 'I am feeling really [*emotion*] today.' We used three words for neutral emotions: 'normal', 'typical', and 'ordinary', and recorded the output from the Halo of the voice-to-text saying each sentence with baseline emotions. With each of the neutral emotions, Halo reported that the both text-to-speech voices sounded 'focused', 'knowledgeable', and 'confident'. We then used 122 words based on Plutchik's wheel of emotions [36], and found that approximately 20% of the emotional words resulted in output that differed from the neutral baseline, with results constant over both test voices. For instance, saying the word 'happy' in the sentence resulted in the Halo reporting voices as 'affectionate', 'caring', and 'proud'. While saying the word 'sad' resulted in the same voice reported as sounding 'discouraged', 'sad', and 'uncomfortable'. As such, we concluded from our initial tests that words do indeed, have an impact on the output of the Halo.

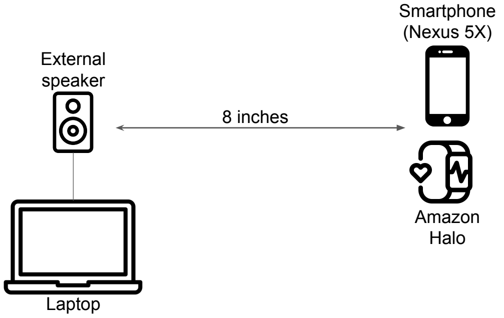


Figure 2: A diagram of the physical testing setup for our experiments.

Our underlying research questions focus on assessing gender biases in the Halo's output. Thus, for our research, we do not dive more deeply into the analysis of the role of specific words in the Halo's output. Rather, since our preliminary experiments above confirm that the spoken words might impact the Halo's output, we concluded that it was essential to control for spoken words in our analyses (Section 3.1.2).

3.1.2 Selecting the Final Audio Set. Given that the words a person might say could affect Halo's output (Section 3.1.1), we sought a data set in which both women and men said the same words. We used audio clips from RAVDESS data set, a gender-balanced set of 12 women and 12 men actors with North American English accents.[3] These actors repeat two phrases two times each: 'kids are talking by the door' and 'dogs are sitting by the door' [31]. Actors convey these sentences in eight different types of emotions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. And, each emotion (except for neutral) is repeated in two intensities: normal (representing 'everyday' speech) and strong ('clear, unambiguous emotional exemplars' of emotions). To validate that the intended emotions were conveyed, the creators of the RAVDESS data set had the audio clips labeled by 247 workers. The paper introducing the RAVDESS dataset has, at the time of this writing (May 2023) been cited over 1000 times and has been used as the dataset for numerous studies.

For our experiments, we only use audio clips where actors convey emotions at normal intensities. In total, our data set covers 768 audio clips, which represents 32 clips per actor. By using audio clips from RAVDESS data set, we are thus able to focus on gender as the main feature that we change when testing the Halo.

[3] RAVDESS paper [31] refers to the actors as female and male; our assessment is that it does so because it was written prior to today's convergence on the use of "women" and "men" to denote gender and "female" and "male" to denote sex. Additionally, as noted earlier, although gender is a spectrum, our experiments focus on the binary genders of women and men because of the availability of the RAVDESS dataset.

3.2 Testing the Halo

3.2.1 Physical Setup. We experimented with the first-generation Amazon Halo Band, with which users interact through an app on an associated smartphone device. (Later versions of the device include an integrated display while also supporting Tone Analysis through the app.) For our experiments, we paired a Halo Band with a Nexus 5X phone. To simulate users speaking to the Halo, we played audio clips from each actor through a set of external speakers connected to a laptop. We placed both the Halo and the smartphone eight inches away from the speakers. All experiments were done in the same room, with the Halo, smartphone, and speakers on the same surface and in the same physical configuration. Figure 2 diagrams the physical set up used to test audio.

3.2.2 Training Voice Profiles. In order to use the Tone Analysis feature in Halo, users must set up a voice profile called a 'Voice ID' by reading six specific phrases out loud to the device. Amazon uses this Voice ID to identify the user's voice in conversation [7], so that it only analyzes tones associated with their speech. In preliminary experiments, we found that Voice ID training was important: the Halo sometimes failed to detect the actor's voice if we did Voice ID training without using the full range of their audio clips. Since we do not have the RAVDESS actors speaking the specific sentences that the Halo uses to train the Voice ID, we used voice cloning methods in order to simulate each actor's voice. To create each actor's voice clone, we used the coqui-ai text-to-speech Python library [3], which creates synthetic voice clones from audio clips that are a maximum of 30 seconds long. Given a voice clone, we can then use coqui-ai's text-to-speech feature to have these synthetic voices to "say" the Voice ID training phrases required by the Halo.

To create a voice clone for a RAVDESS actor, we needed to stitch together some of their audio clips as input. In preliminary experiments, we found that using only the neutral emotion audio clips was insufficient. The Voice ID trained with the neutral-only clones lead to the Halo failing to recognize the actor as the device's 'owner' nearly 20% of the time in our preliminary experiments (with 168 audio clips, 7 emotions each from 24 actors).

As we learned that an actor's full range of audio clips is necessary to create their synthetic voices, we separated the RAVDESS data set into two equal parts for the full set of experiments: a *training* set, comprised of the first iteration of each phrase said in every emotion. And, a *testing* set, comprised of the second iteration of each phrase said in every emotion. We used the audio clips from the training set to create the actor's synthetic voice on coqui-ai.

Our separation of the RAVDESS dataset into the training set and the testing set ensures that the audio clips used to train the synthetic voice for for each actor (1) captures the full range of pitch individuals use, and (2) remains separate from the final set of audio clips that we test for potential bias. We then used this simulated voice to repeat the necessary phrases to train each actor's Voice ID on the Halo.

3.2.3 Interacting with the Halo. Our procedure for testing audio clips was as follows. For each of the 24 actors in our dataset, we tested their 16 audio clips (two phrases, eight emotions) from the testing set. For each actor-clip pair, we:

(1) Trained the Voice ID as described above.
(2) Played the audio clip to the Halo.
(3) Took a screenshot of the Halo results on the smartphone app (from which we later extracted the three adjectives Halo used to describe the audio clip using OCR).
(4) Repeat steps (2) and (3) four more times with the same clip.
(5) Reset the Halo device by deleting the actor's Voice ID and Tone data.

The reason we repeated each audio clip five times (step 4) is because we found in preliminary experiments that the Halo does not deterministically provide the same output for the same input. The reason we reset the device (step 5) not only between each actor but between each actor's unique audio clip is that we wanted to avoid any potential 'contamination' between experiments, in case the Halo's outputs depend on previous speech from the user.

To streamline our testing, we automated interactions with the Halo (through its smartphone app) using python-pure-adb, a Python implementation of Android Debug Bridge [6].

3.2.4 Data Analysis. Our analysis of the Halo's output centers on the words (adjectives) it uses to describe women's and men's voices. As discussed in Section 2, the valence of a word is a measure of the word's positivity or negativity. We use the valence-arousal-dominance scores from Warriner et al. [45] for our analyses. Foreshadowing our experimental results in Section 4, the Halo outputs 44 adjectives for tones in our experiments; 15 of these adjectives have positive valence, 25 have negative velance, and two have neutral valence. The valence of all these adjectives are listed in Appendix A. The Halo also outputs 2 adjectives ('disinterested' and 'unconfident') that do not have a valence in the [45] list of known valence values. When conducting valence-based analyses, we exclude these words because (1) Warriner et al. [45] does not offer a valence value for them and (2) the Halo outputs these words infrequently (11 times and one time, respectively, out of a total of 5,760 words output).

3.3 Positionality

None of the authors are themselves users of the Amazon Halo. Two authors are women; two authors are men. Each author has experienced and/or observed the interplay between the words used to describe emotional tone and gender. Hence, the authors believe that if biases exist in an automated tone-labeling system, like the Halo, those biases could be problematic, thus motivating this project's research questions.

4 RESULTS

We now turn to our results. Recall our main research question: does the Amazon Halo differ — that is, exhibit potential bias — in how it interprets the tones of women's voices versus men's voices? After finding the answer to this question to be 'yes' at a high level, we dive into characterizing those differences and their possible implications.

4.1 An Initial Look at Halo's Outputs on our Dataset

We begin by presenting an overview of Halo's outputs on our dataset, including an initial look at gender differences.

Dataset Overview. Recall that Halo's Tone analysis reports three adjectives to describe the emotions, from strongest to weakest, expressed in voices. In this paper, we call these reported adjectives *inferred emotions.* And recall (Section 3) that we tested 384 audio clips from the RAVDESS dataset representing eight *intended emotions*, recorded from 12 women and 12 men. We played each clip to the Halo five times, and recorded the corresponding three Halo-inferred emotions after each iteration. In total, our final dataset thus includes 5,760 (384 × 5 × 3) instances of (not necessarily unique) inferred emotions.

Halo's range of inferred emotions was broad, and exhibited some gender-based differences. Though there are only eight intended emotions in the RAVDESS dataset, the 5,760 inferred emotions in our data set corresponds to 44 unique emotion words. Figure 3 shows the distribution of how often these 44 unique inferred emotions were used to to describe audio clips, distinguishing between women's and men's voices.

We can draw two initial conclusions from this figure. First, we see that the Halo output a broad range of inferred emotions, with some output frequently and others in a long tail: the top 13 inferred emotions represent over 90% of the total set of (non-unique) words Halo used to describe people's voices. The most common inferred emotions included: 'confident', 'focused', 'knowledgeable', 'opinionated', and 'stern'.

Second, informally, we observe differences between how women's and men's voices were reported. For example, from Figure 3, it appears that some inferred emotions were infrequently used for men and were used substantially more for women (e.g., 'stubborn') whereas others were frequently used for men but infrequently for women (e.g., 'knowledgeable'). We will investigate these differences carefully in subsequent sections.

The valence of Halo's inferred emotions matched the expected output overall, but also exhibited some gender-based differences. Next, we consider whether the emotions inferred by the Halo were positive or negative. Recall that we use valence values from the valence-arousal-dominance scores [45] to interpret the positivity or negativity of emotions (Section 3.2.4).

Figure 4 shows the density histogram of valence values associated with the 5,760 instances of inferred emotions in our data set, including a breakdown of whether these emotions described women's or men's voices. Considering unique emotions (not instances), we find that of the 42 unique inferred emotions by Halo, 25 were negative, 15 were positive, and 2 were neutral. As our point of comparison, note that for the eight intended emotions from RAVDESS, three have a negative valence, one is neutral, and four are positive.

We can again draw two conclusions from this data. First, we find that the overall breakdown of valence values loosely aligns with what we would expect from RAVDESS. But despite representing a large number of the 42 unique inferred emotions, negative words comprise 27.55% (1,587 of 5,760) of the inferred emotion *instances.* Neutral words represent 9.79% (564 of 5,760) of inferred emotion instances, and positive words represents 62.45% (3,597 of 5,760).

Second, we again observe indications of gender-based differences, which we explore in more depth in the next section. For instance, of the 1,587 time the Halo inferred emotions with negative valence values, 1,150 (72.5%) of those were used to describe women's voices. And of the 3,597 times Halo inferred emotions with positive valence, it applied those to men's voices 2,188 times (60.8%).

4.2 A Closer Look at Gender-Based Differences

Recall that the RAVDESS dataset consists of 12 women and 12 men speaking the same sentences with the same intended emotions. Thus, if there were no gender bias in the Halo's analysis of their tones, we would expect to see no gender-based differences in the output. Above, we have already seen indications that there *are* gender-based differences. In this section, we dive more deeply into analyzing these differences.

4.2.1 *RQ1: Assessing the Existence of Gender-Based Differences.*

We begin by investigating the valence of Halo's inferred emotions for women and men in the overall dataset, towards answering RQ1.

The Halo interpreted men's voices in our dataset with a more positive mean valence than women's voices overall. Figure 5 shows the distribution of valence values associated with women and men's voices, across the whole dataset. Overall, inferred emotions from men's voices had higher valence values. More specifically, the mean (6.6) and median (6.9) valence values for men correspond to positive values. In contrast, the mean (5.5) and median (5.6) valence values for women correspond to neutral values. A Mann-Whitney U test showed that the difference between men and women's median valence value was statistically significant (Z=1275.5, $p<0.05$).

Men's speech in our dataset was interpreted with a positive valence substantially more often than women's speech; Women's speech was interpreted with a negative valence more frequently. Diving more deeply into these differences come from, Figure 6 shows how these valence values correspond to inferred emotions. In this split bar graph, inferred emotions are sorted by their valence value, from positive (top) to negative (bottom). The chart shows us the distribution of inferred emotions, by gender. The bottom horizontal line corresponds to a 25% line, e.g., for women, 25% of the words output by the Halo (over all our experiments) had the valence of 'stubborn' or below. The middle line corresponds to a 50% line, e.g., for women, 50% of the words output by the Halo had the valence of 'opinionated' or below. The top line corresponds to a 75% line.

From this data, we observe that the Halo reported men's voices using positive inferred emotions more than 75% of the time. That is, 75% of the inferred emotions Halo used to describe men's voices have valence values either equivalent to or greater than 'focused' (6.48). In contrast, the Halo reported women's voices as positive less than 50% of the time.

Meanwhile, the Halo described women's voices with negative inferred emotions more than one-third of the time (39.9%). This is much more frequently than for men's voices, which were interpreted with negative emotions less than a quarter (15.2%) of the time.

4.2.2 *RQ2: Specific Emotions Attributed to Women Versus Men.*

Having found gender-based differences in valence in the Halo's output, thereby answering RQ1 in the affirmative, we now investigate more specifically *which* emotions were inferred for women versus men.

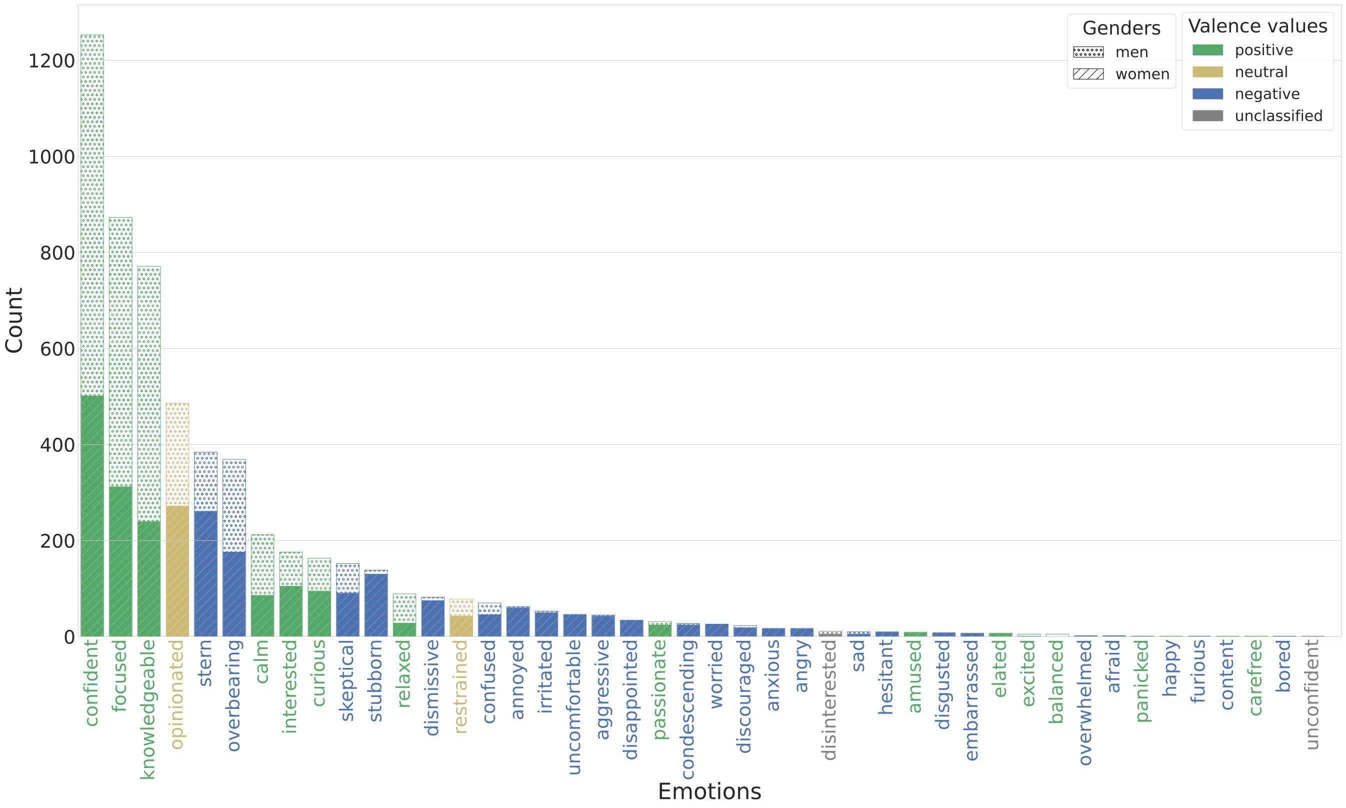


Figure 3: Distribution of inferred emotions.

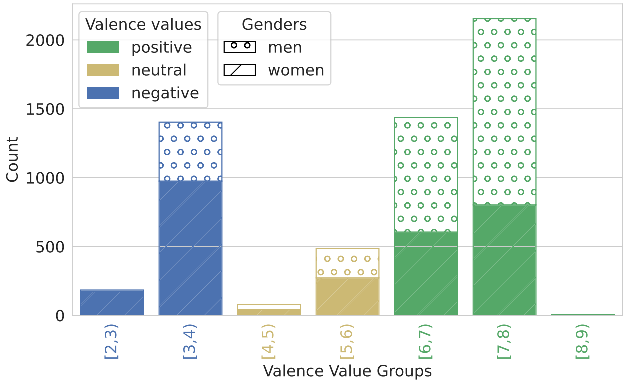


Figure 4: Distribution of valence values of instances of inferred emotions. The x-axis indicates that the lower number is inclusive, and the upper number is exclusive. For instance, [4, 5) indicates that the bucket is ≥ 4, and < 5.

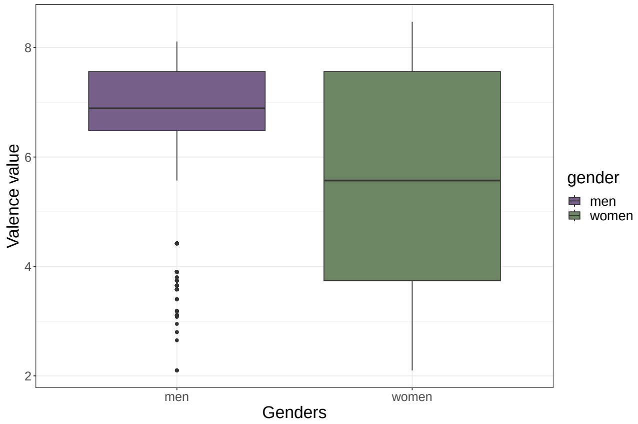


Figure 5: A box plot summarizing the valence values of inferred emotions for men and women.

Halo described women's voices with a greater variety in inferred emotions than men's voices. We begin by considering simply what inferred emotions the Halo produced for women's versus men's voices. As shown in Figure 7, of the 42 unique inferred emotions in our data set, we find that Halo used a wider range of words to describe women's voices than men's. More specifically, across all types of speech, Halo used 38 inferred emotions to describe the emotions expressed by women (14 unique only to women), and 28 for for men (four unique only to men).

Inferred emotions were distributed disproportionately by gender. We now look more closely — statistically — at whether an actor's

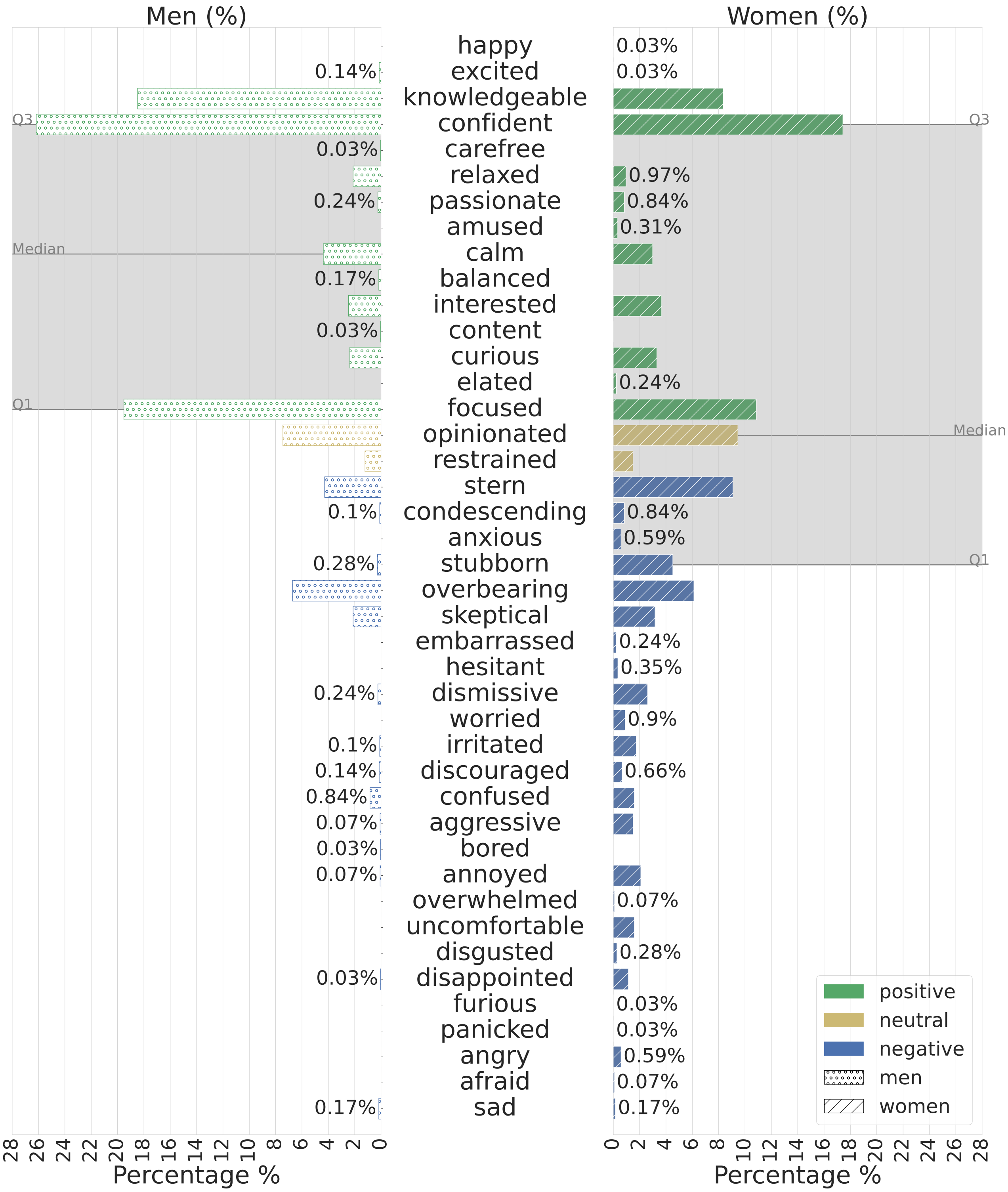


Figure 6: A horizontal split bar graph showing the distribution of inferred emotions for men and women. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

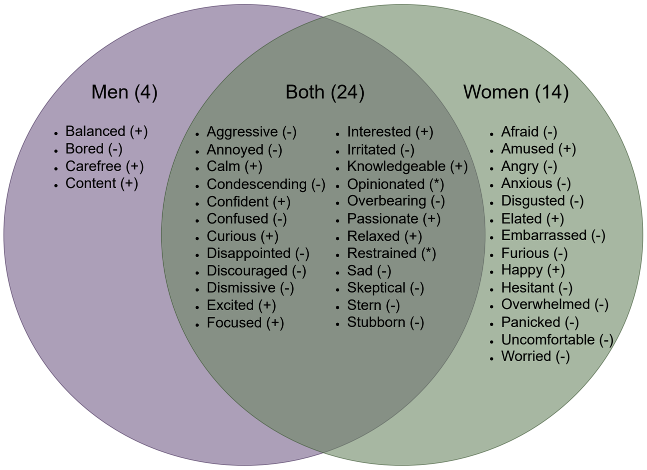


Figure 7: A Venn diagram showing the relationship between inferred emotions used to describe men and women's voices. Inferred emotions appear in alphabetic order. Words are marked as positive (+), negative (-), or neutral (*), according to valence value analysis.

Emotion	Men	Women
aggressive	*0.07%*	**1.49%**
angry	*0%*	**0.59%**
annoyed	*0.07%*	**2.08%**
anxious	*0%*	**0.59%**
condescending	*0.1%*	**0.83%**
confident	**26.11%**	*17.4%*
disappointed	*0.03%*	**1.15%**
dismissive	*0.24%*	**2.6%**
focused	**19.47%**	*10.83%*
irritated	*0.1%*	**1.74%**
knowledgeable	**18.44%**	*8.33%*
relaxed	**2.12%**	*0.97%*
stern	*4.27%*	**9.06%**
stubborn	*0.28%*	**4.51%**
uncomfortable	*0%*	**1.6%**
worried	*0%*	**0.9%**

Table 1: Percent of inferred emotions used more or less frequently than expected. Red/bold cells indicate a higher proportion than expected, and blue/italic cells indicate a lower proportion than expected. Darker colors indicate a larger difference.

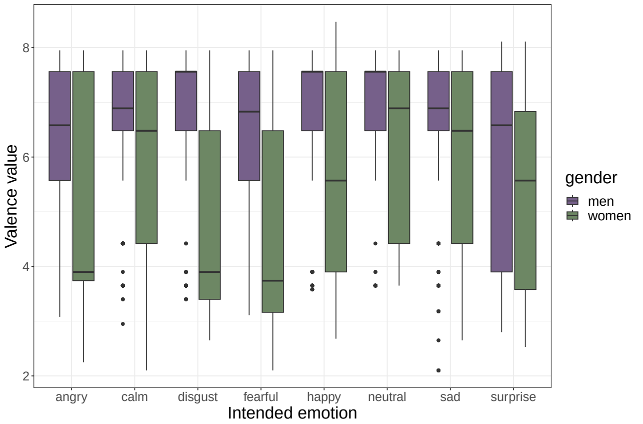


Figure 8: A box plot summarizing valence values of inferred emotions for men and women, separated by the intended emotions from the RAVDESS dataset.

gender in the RAVDESS dataset is related to the inferred emotions the Halo used to describe their voices. In order to do this, we conducted an omnibus chi-squared test of independence, which takes into account how frequently the Halo used each inferred emotion. The chi-squared test returned a statistically significant association between genders and inferred emotions (χ^2(43, N=5,760) = 838.8, p < 0.001). This means that an actor's gender does affect the emotions inferred by Halo.

Specific negative inferred emotions were disproportionately attributed to women (and vice versa). We now investigate more specifically *which* inferred emotions were used more or less frequently for women versus men. To do so, we conducted post-hoc Z tests on the standardized residuals of the chi-squared test. We corrected for multiple comparisons with the Bonferonni method to calculate the critical value (3.25). We observe that 16 of 42 inferred emotions exceeded this critical value, meaning that they were used more or less frequently than expected. These 16 inferred emotions are shown in Table 1, where the values in each cell show how often the word was used to describe actors of that gender, in percent format.

We observe that there were 12 inferred emotions used more commonly to describe women's voices, and they all have negative valence values: 'aggressive', 'angry', 'annoyed', 'anxious', 'condescending', 'disappointed', 'dismissive', 'irritated', 'stern', 'stubborn', 'uncomfortable', and 'worried'. The most disproportionately used inferred emotion to describe women is 'stubborn': we found that Halo was 16× more likely to use 'stubborn' to describe a women's voice than a man's voice.

In contrast, the Halo used four inferred emotions more commonly than expected to describe men's voices, and they all have positive valence values: 'confident', 'focused', 'knowledgeable', and 'relaxed'. We found that Halo used 'knowledgeable' to most disproportionately describe men's voices: it was 2× more likely to use 'knowledgeable' to describe a man's voice than a woman's.

4.2.3 RQ3: Intended Emotions with Most Biased Outputs.

Finally, we turn to our third research question: Of the eight source emotions in the professional actor voice dataset, are there some for which the gender biases are particularly strong?

Halo differed on women versus men's voices particularly on the intended emotions of disgust, fearful, and angry. Figure 8 breaks down the valence values of inferred emotions by the eight intended emotions from the RAVDESS dataset. As with the overall box plot (recall Figure 5), we can see that Halo consistently reports men's voices using inferred emotions associated with higher valence values, regardless of what emotion the actors were intending to convey. We see the biggest differences in median valence values for 'disgust', 'fearful', and 'angry'.

Below, we dive into the intended emotion of 'disgust', where we see the largest difference in median valence values between genders. Appendix B provides similar data for the other intended emotions as well. Overall, we find that in six of the eight intended source emotions, men's interquartile range falls entirely within inferred emotions with positive valence values. By contrast, women's interquartile range extends to inferred emotions with negative valence values in five of the eight intended source emotions.

Case study of women and men intending to sound 'disgusted'. As a case study, we dive more deeply into the intended emotion where the median valence values for men and women had the largest difference: disgust. Men voice actors voicing 'disgusted' source emotions had an inferred emotion median valence value of 7.56, while women had a median of 3.9. Figure 9 shows a split bar chart for the distribution of inferred emotions for men and women. Not only does men's interquartile range fall entirely within inferred emotions with positive valence values, we also observe that men's interquartile range ends where women's begin. That is, 75% of inferred emotions used by Halo to describe disgusted men have valence values equal to, or above 'focused' (6.48). In contrast, the Halo only uses words with valence values in the same range to describe women less than 25% of the time.

Statistically, considering only the Halo's outputs on inputs with the intended emotion of disgust, a chi-squared test of independence found a significant association between gender and inferred emotion ($\chi^2(28, N=720) = 272.45$, $p < 0.001$). Post-hoc Z-tests on the standardized residuals found that 10 of the 29 inferred emotions exceeded the critical value of 3.13 ($p<0.05$). Table 2 shows the 29 inferred emotions, as well as the 10 that were used disproportionately to describe either men or women when they intended to sound disgusted. We find that the Halo used 'calm', 'confident', 'focused', and 'knowledgeable' more frequently to describe men conveying disgust. On the other hand, it used 'annoyed', 'dismissive, irritated', 'stern' and 'stubborn' more crequently to describe women. Notably, all of the inferred emotions associated with men have positive valence values, while all of the inferred emotions associated with women have negative. Of the words most disproportionately use to describe either gender, the Halo was almost 3× more likely to report that a disgusted man sounded 'knowledgeable'. And, while disgusted women were reported as sounding 'stubborn' 10.28% of the time, the Halo never used the word to describe men's voices.

Emotion	Men	Women
aggressive	0%	1.39%
annoyed	*0%*	**5.28%**
anxious	0%	2.5%
calm	**4.17%**	*0%*
carefree	0.28%	0%
condescending	0%	1.39%
confident	**27.78%**	*13.33%*
confused	0%	0.56%
curious	0.28%	0.56%
disappointed	0%	1.39%
discouraged	0%	0.28%
disgusted	0%	1.39%
disinterested	1.39%	0%
dismissive	*1.39%*	**8.33%**
embarrassed	0%	0.56%
focused	**20.83%**	*7.22%*
interested	1.39%	1.67%
irritated	*0%*	**3.89%**
knowledgeable	**22.5%**	*8.06%*
opinionated	8.33%	9.72%
overbearing	5%	3.61%
passionate	1.11%	0%
relaxed	2.5%	0%
restrained	0.56%	0%
skeptical	0%	0.56%
stern	*2.5%*	**11.67%**
stubborn	*0%*	**10.28%**
uncomfortable	0%	3.61%
worried	*0%*	**2.78%**

Table 2: Inferred emotions used to describe actors conveying disgust. Percent of inferred emotions used more or less frequently than expected. Red/bold cells indicate a higher proportion than expected, and blue/italic cells indicate a lower proportion than expected. Darker colors indicate a larger difference.

5 DISCUSSION

We now take a step back from our findings — which suggest that the Halo interprets women's and men's voices in different, potentially biased ways — to discuss implications, limitations, future work.

5.1 Implications

At the highest level, our findings show that significant (in this case, gender-based) biases exist in deployed machine learning models in a commercial product, the Amazon Halo. This work thus supports a growing number of calls that ML models, especially the ones that are deployed in commercial products, should be thoroughly tested before they are widely released to the public.

More specifically, gender-biased recommendations from the Halo have potential negative implications both for individual users and at a societal level. We cannot be certain without further study what (and how severe) these implications might be, but we nevertheless believe that it is important for Amazon (in this case) and other companies to anticipate and minimize biases that come with such potential implications in advance. For example, individual harms to

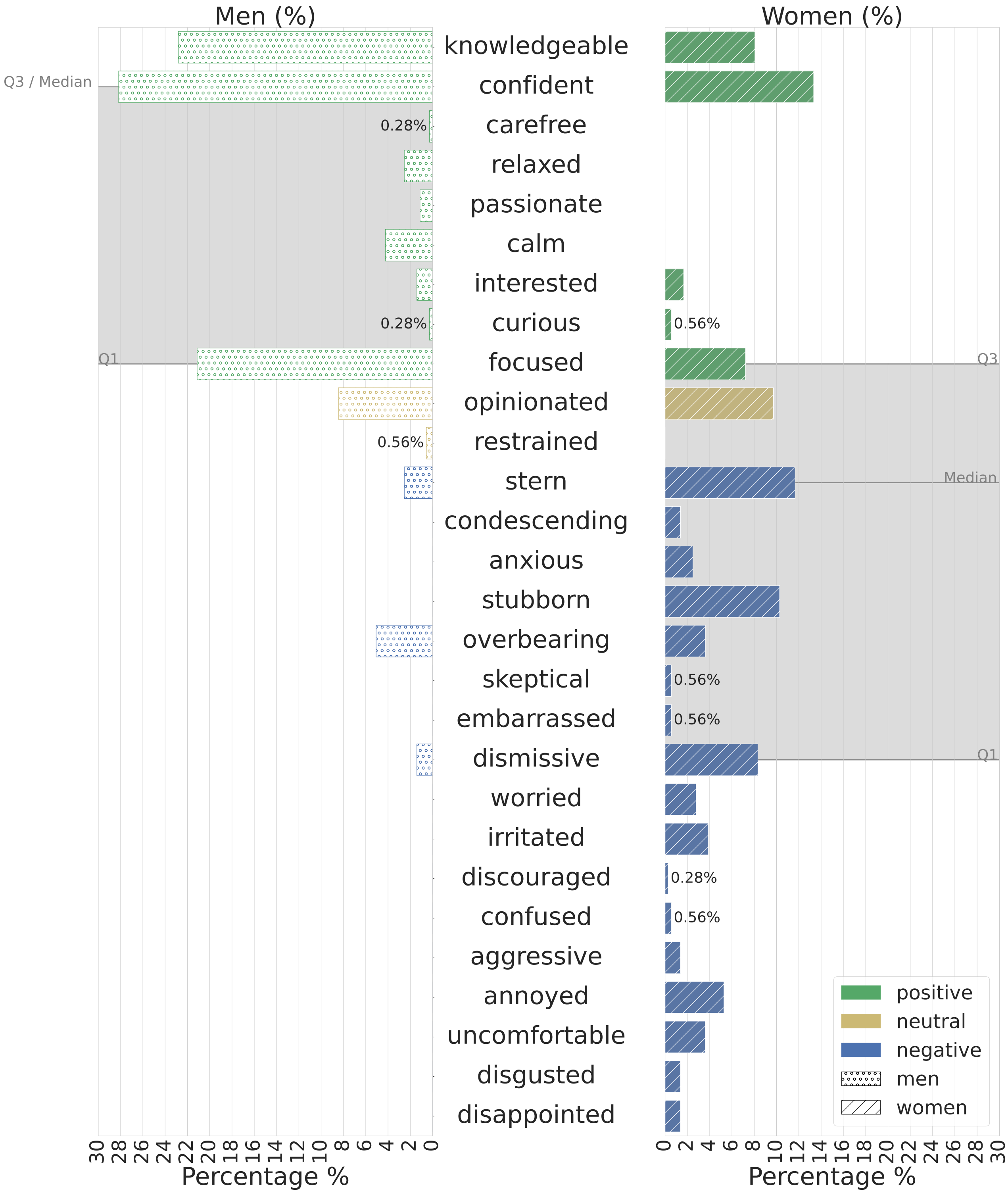


Figure 9: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *disgust*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

users whose tone is mis-labeled (e.g., the Halo informs women expressing happiness that they often sound 'opinionated' and 'stern') might include emotional distress, loss of confidence, or ineffectively attempting to adapt their speech patterns. Moreover, if widely deployed, devices like the Halo could potentially solidify and even amplify existing societal biases. And while the Halo is marketed as a device for people to analyze *themselves*, future products might allow people to analyze *others*, where biased outputs may lead to additional harms. We encourage future work that directly studies these potential implications, with end users, for future products with features like Halo's tone analysis.

5.2 Scope, Limitations, and Future Work

In our study we establish that there are biases in Halo's output, but we do not investigate the underlying reason for these biases. One possible explanation is that the training data behind the Halo's ML is itself biased, encoding existing societal gender biases. Future work could compare the Halo's outputs to human judgements on the same datasets. Another possibility is that the Halo's ML model relies on basic characteristics of humans voice, such as pitch and frequency, which naturally vary for men and women [44]. Future work could reverse engineer Halo's output (or a product with similar features), e.g., by establishing correlations between voice characteristics and Halo's inferences.

Another limitation of this work is that we do not have ground truth for tone analysis. Different listeners may have different interpretations of the tone or emotion conveyed in an audio clip, due to cultural factors, personal biases, or other context. In our experiments and analysis, we mitigate this issue in several ways. We use the widely-used RAVDESS dataset, which was previously validated for conveying the intended emotions and controls for word choice. We use valence values (created by previous research) rather than our own qualitative analysis to evaluate the positivity or negativity of emotion words. Finally, we conducted our experiments in a controlled setting to minimize the impact of any other variables (e.g., all experiments within a short time window to minimize the possibility that the Halo's model changed, resetting the device for each audio clip to avoid any potential impact of user-specific learning, and conducting all experiments in the same environment without background noise). Future work with the Halo or similar devices might also explore more realistic speech settings, e.g., real user speech or a less controlled audio dataset than RAVDESS.

Finally, we evaluate the Halo's outputs using a perspective and methods from computer science. There is a rich literature around gender bias, perception, and speech and language in psychology (e.g., [11, 25]), and we look forward to scholars from those fields connecting with our findings here.

6 CONCLUSION

Our work here, analyzing the Amazon Halo's tone analysis feature, adds to a growing body of work auditing and calling attention to biases in deployed machine learning models and the technologies that rely on them. Using an existing dataset of women's and men's voices saying the same sentences with different intended emotions, we found that the Halo's output for women's and men's voices differs systematically. Among other results, we found that the Halo interprets the women's voices as having more negative emotions and the men's voices as having more positive emotions, both with respect to each other and with respect to the intended emotion. We believe these potentially biased output have individual and societal implications in the context of this specific technology, especially if it (or something similar) becomes more widely deployed — as well as more generally adds to a chorus of warnings about bias in deployed machine learning models.

This research was supported in part by NSF Award #2205171.

REFERENCES

[1] Amazon halo band – measure how you move, sleep, and sound – designed with privacy in mind. https://web.archive.org/web/20230101175501/https://www.amazon.com/Amazon-Halo-Fitness-And-Health-Band/dp/B07QK955LS.
[2] Apple Fitness +. https://www.apple.com/apple-fitness-plus/. Accessed on 05/07/2023.
[3] Coqui. https://coqui.ai/. Accessed on 9/20/2022.
[4] Google Fitness Trackers. https://store.google.com/us/category/watches?hl=en-US#watches-cat-trackers. Accessed on 05/07/2023.
[5] Microsoft Azure. https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/#features. Accessed on 8/18/2022.
[6] pure-python-adb. https://pypi.org/project/pure-python-adb/. Accessed on 8/20/2022.
[7] A new tool to help you understand and improve your social wellbeing, Aug. 2020. Accessed on 9/21/2022.
[8] Akhtar, I. Amazon halo's tone of voice feature made me a better person (or at least sound like one). CNET. https://www.cnet.com/health/fitness/amazon-halo-fitness-band-tone-of-voice-feature-made-me-a-better-person-or-at-least-sound-like-one/. Accessed on 5/6/2023,.
[9] Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., and Vilares, J. Sentiment analysis for fake news detection. *Electronics 10*, 11 (2021), 1348.
[10] Amanatullah, E. T., and Tinsley, C. H. Punishing female negotiators for asserting too much…or not enough: Exploring why advocacy moderates backlash against assertive female negotiators. *Organizational Behavior and Human Decision Processes 120*, 1 (2013), 110–122.
[11] Anderson, R. C., and Klofstad, C. A. Preference for leaders with masculine voices holds in the case of feminine leadership roles. *PloS one 7*, 12 (2012), e51216.
[12] Bajorek, J. P. Voice recognition still has significant race and gender biases. *Harvard Business Review 10* (2019).
[13] Bertrand, M., and Hallock, K. F. The gender gap in top corporate jobs. *ILR Review 55*, 1 (2001), 3–21.
[14] Blau, F. D., and Kahn, L. M. The gender wage gap: Extent, trends, and explanations. *Journal of economic literature 55*, 3 (2017), 789–865.
[15] Brescoll, V. L. Leading with their hearts? how gender stereotypes of emotion lead to biased evaluations of female leaders. *The Leadership Quarterly 27*, 3 (2016), 415–428.
[16] Brescoll, V. L., and Uhlmann, E. L. Can an angry woman get ahead?: Status conferral, gender, and expression of emotion in the workplace. *Psychological Science 19*, 3 (2008), 268–275.
[17] Brescoll, V. L., and Uhlmann, E. L. Can an angry woman get ahead? status conferral, gender, and expression of emotion in the workplace. *Psychological science 19*, 3 (2008), 268–275.
[18] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (2018), PMLR, pp. 77–91.
[19] Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186.
[20] Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology 82*, 6 (2002), 878.
[21] Hayes, S. D., Crocker, P. R., and Kowalski, K. C. Gender differences in physical self-perceptions, global self-esteem and physical activity: Evaluation of the physical self-perception profile model. *Journal of Sport Behavior 22*, 1 (1999), 1.
[22] Hutto, C., and Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (2014), vol. 8, pp. 216–225.
[23] Imana, B., Korolova, A., and Heidemann, J. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021* (2021), pp. 3767–3778.
[24] Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic fairness. In *Aea papers and proceedings* (2018), vol. 108, pp. 22–27.
[25] Klofstad, C. A., Anderson, R. C., and Peters, S. Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women.

Proceedings of the Royal Society B: Biological Sciences 279, 1738 (2012), 2698–2704.

[26] Ko, M. Amazon halo - creepiest thing i've ever reviewed). https://www.youtube.com/watch?v=HL4JYfn4C5Y. Accessed 5/6/2023.

[27] Kordzadeh, N., and Ghasemaghaei, M. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems 31*, 3 (2022), 388–409.

[28] Kouloumpis, E., Wilson, T., and Moore, J. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the international AAAI conference on web and social media* (2011), vol. 5, pp. 538–541.

[29] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. Accessed on 5/7/2023.

[30] Lippa, R. A., Preston, K., and Penner, J. Women's representation in 60 occupations from 1972 to 2010: More women in high-status jobs, few women in things-oriented jobs. *PloS one 9*, 5 (2014), e95960.

[31] Livingstone, S. R., and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one 13*, 5 (2018), e0196391.

[32] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*, 6 (2021), 1–35.

[33] Michie, S., and Nelson, D. L. Barriers women face in information technology careers: Self-efficacy, passion and gender biases. *Women in management review 21*, 1 (2006), 10–27.

[34] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application 8* (2021), 141–163.

[35] Nielsen, F. Å. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903* (2011).

[36] Plutchik, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist 89*, 4 (2001), 344–350.

[37] Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis* (2016), pp. 9–15.

[38] Raji, I. D., and Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2019), ACM, pp. 429–435.

[39] Russell, J. A. A circumplex model of affect. *Journal of personality and social psychology 39*, 6 (1980), 1161.

[40] Sapiezynski, P., Ghosh, A., Kaplan, L., Rieke, A., and Mislove, A. Algorithms that" don't see color" measuring biases in lookalike and special ad audiences. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), pp. 609–616.

[41] Soklaridis, S., Kuper, A., Whitehead, C. R., Ferguson, G., Taylor, V. H., and Zahn, C. Gender bias in hospital leadership: a qualitative study on the experiences of women ceos. *Journal of health organization and management 31*, 2 (2017), 253–268.

[42] Song, V. Amazon shuts down Halo division and discontinues all devices. The Verge, Apr. 2023. https://www.theverge.com/2023/4/26/23699459/amazon-layoffs-halo-fitness-tracking-sleep-tracking. Accessed on 5/7/2023.

[43] Srinivasan, R., and Chander, A. Biases in ai systems. *Communications of the ACM 64*, 8 (2021), 44–49.

[44] Titze, I. R., Svec, J. G., and Popolo, P. S. Vocal dose measures.

[45] Warriner, A. B., Kuperman, V., and Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods 45*, 4 (2013), 1191–1207.

[46] Wolfe, R., Yang, Y., Howe, B., and Caliskan, A. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2023).

A VALENCE VALUES FOR HALO OUTPUT WORDS

We list below the adjectives output by the Halo to describe the tone of voices in our experiments. Each item in the list below includes a word (the adjective describing tone), the valence of the word, whether the word is considered positive, neutral, or negative, and how often we observed each word. The list is sorted in decreasing positivity. The valence scores and positivity-negativity denotations are from [45]. In our data set, we evaluate positive words as those with valence values greater than or equal to 6. Neutral words have valence greater or equal to 4 and less than 6. And, negative words have valence values less than 4. At the end of the list are two adjectives output by the Halo that do not have a valence score in [45]. As noted in the body of the paper, these adjectives appear infrequently in our experimental results (11 times and one time, respectively, out of a total of 5,760 adjectives output).

- happy, 8.47, positive, 1
- excited, 8.11, positive, 5
- knowledgeable, 7.95, positive, 771
- confident, 7.56, positive, 1253
- carefree, 7.32, positive, 1
- relaxed, 7.25, positive, 89
- passionate, 7.17, positive, 31
- amused, 7.05, positive, 9
- calm, 6.89, positive, 212
- balanced, 6.84, positive, 5
- interested, 6.83, positive, 176
- content, 6.7, positive, 1
- curious, 6.58, positive, 163
- elated, 6.56, positive, 7
- focused, 6.48, positive, 873
- opinionated, 5.57, neutral, 486
- restrained, 4.42, neutral, 78
- stern, 3.9, negative, 384
- anxious, 3.8, negative, 17
- condescending, 3.8, negative, 27
- stubborn, 3.74, negative, 138
- overbearing, 3.65, negative, 369
- skeptical, 3.58, negative, 152
- embarrassed, 3.51, negative, 7
- hesitant, 3.48, negative, 10
- dismissive, 3.4, negative, 82
- worried, 3.27, negative, 26
- irritated, 3.19, negative, 53
- discouraged, 3.18, negative, 23
- confused, 3.11, negative, 70
- aggressive, 3.08, negative, 45
- bored, 2.95, negative, 1
- annoyed, 2.8, negative, 62
- overwhelmed, 2.8, negative, 2
- uncomfortable, 2.7, negative, 46
- disgusted, 2.68, negative, 8
- disappointed, 2.65, negative, 34
- furious, 2.57, negative, 1
- panicked, 2.56, negative, 1
- angry, 2.53, negative, 17
- afraid, 2.25, negative, 2
- sad, 2.1, negative, 10
- disinterested, N/A, unclassified, 11
- unconfident, N/A, unclassified, 1.

B SPLIT BAR GRAPHS FOR EACH SOURCE EMOTION

In this appendix, we provide the split-plot graphs for each intended source emotion. We omit 'disgust', which appeared in Figure 9 in Section 4.

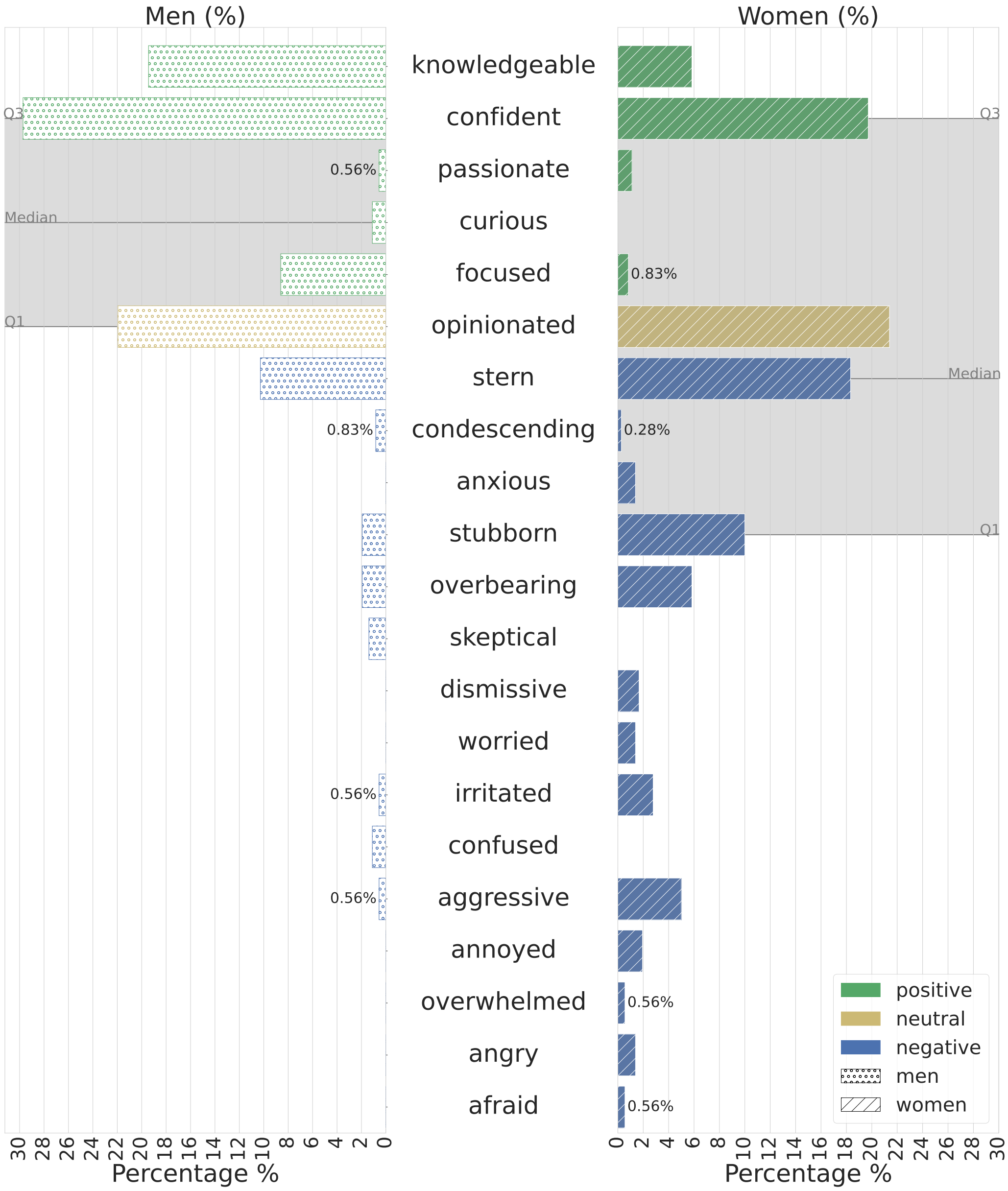


Figure 10: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *anger*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

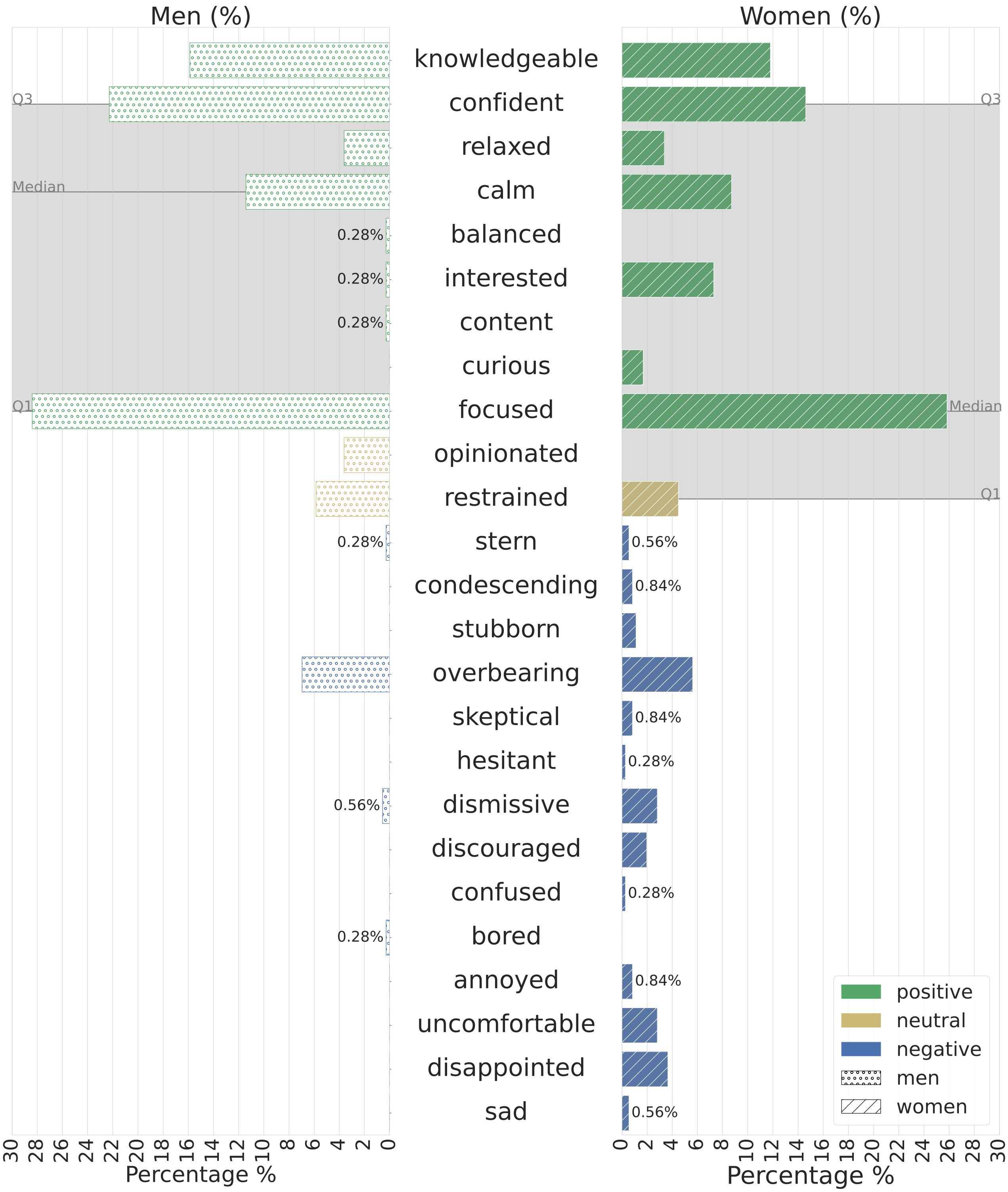


Figure 11: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *calmness*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

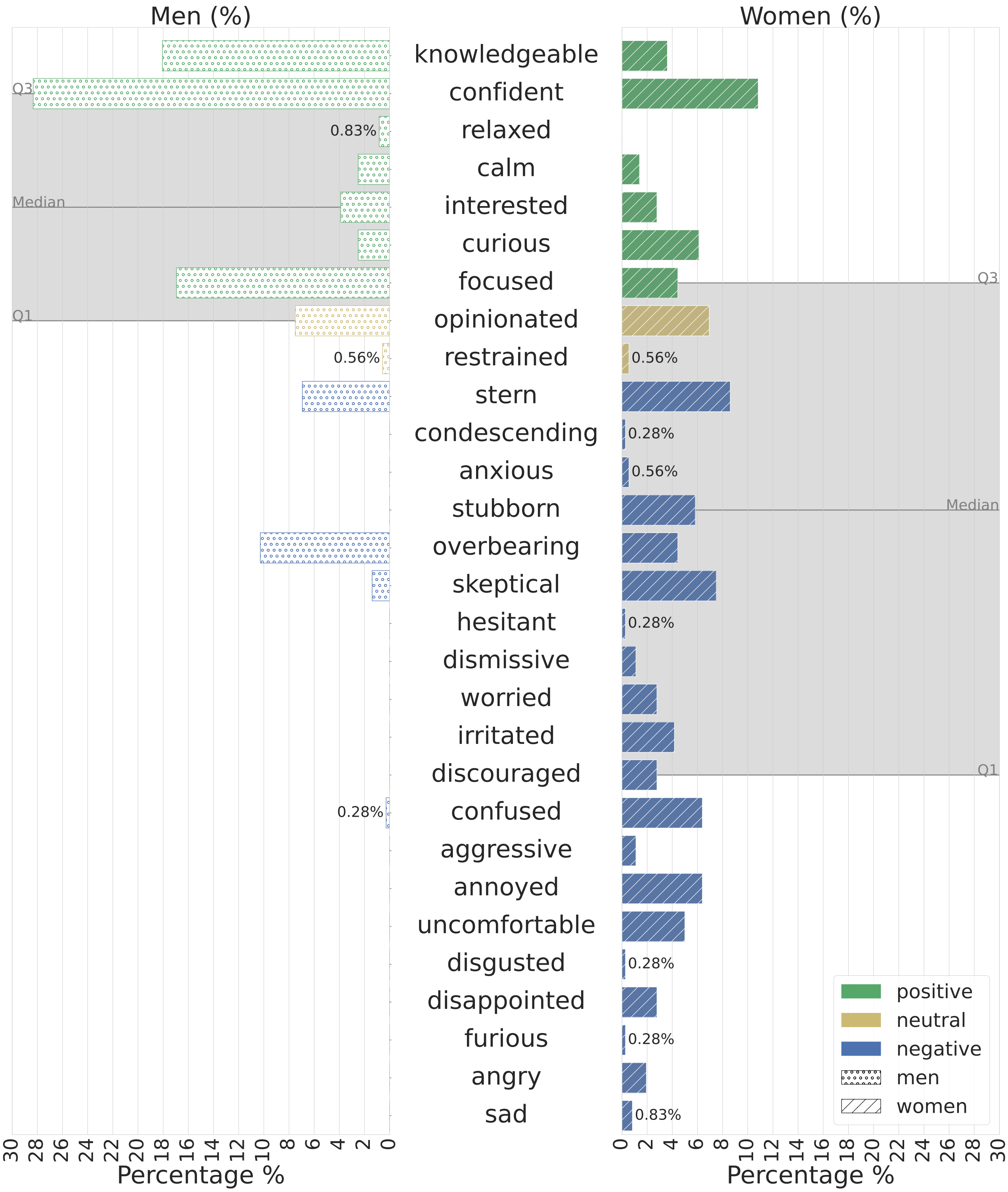


Figure 12: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *fear*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

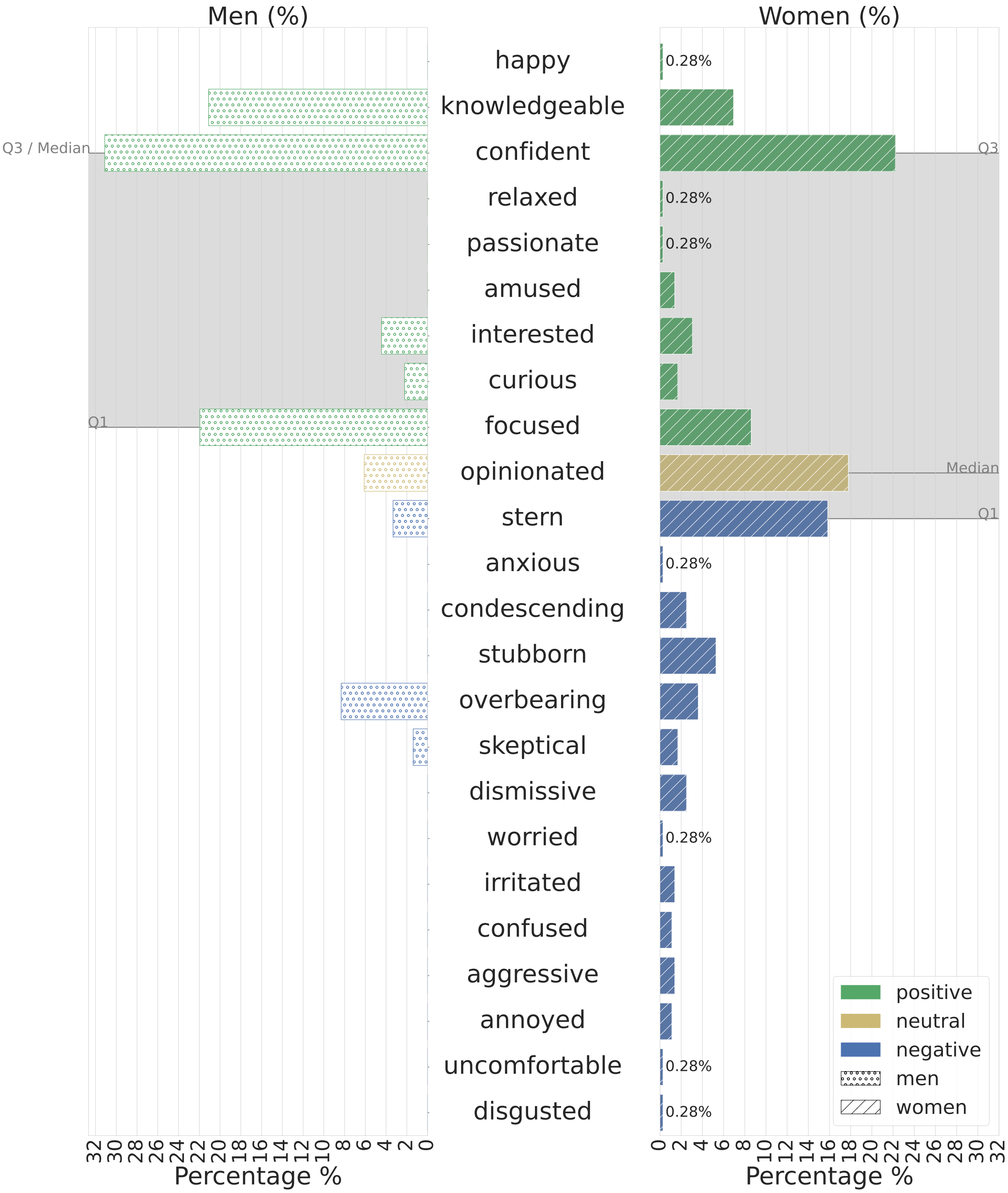


Figure 13: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *happiness*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

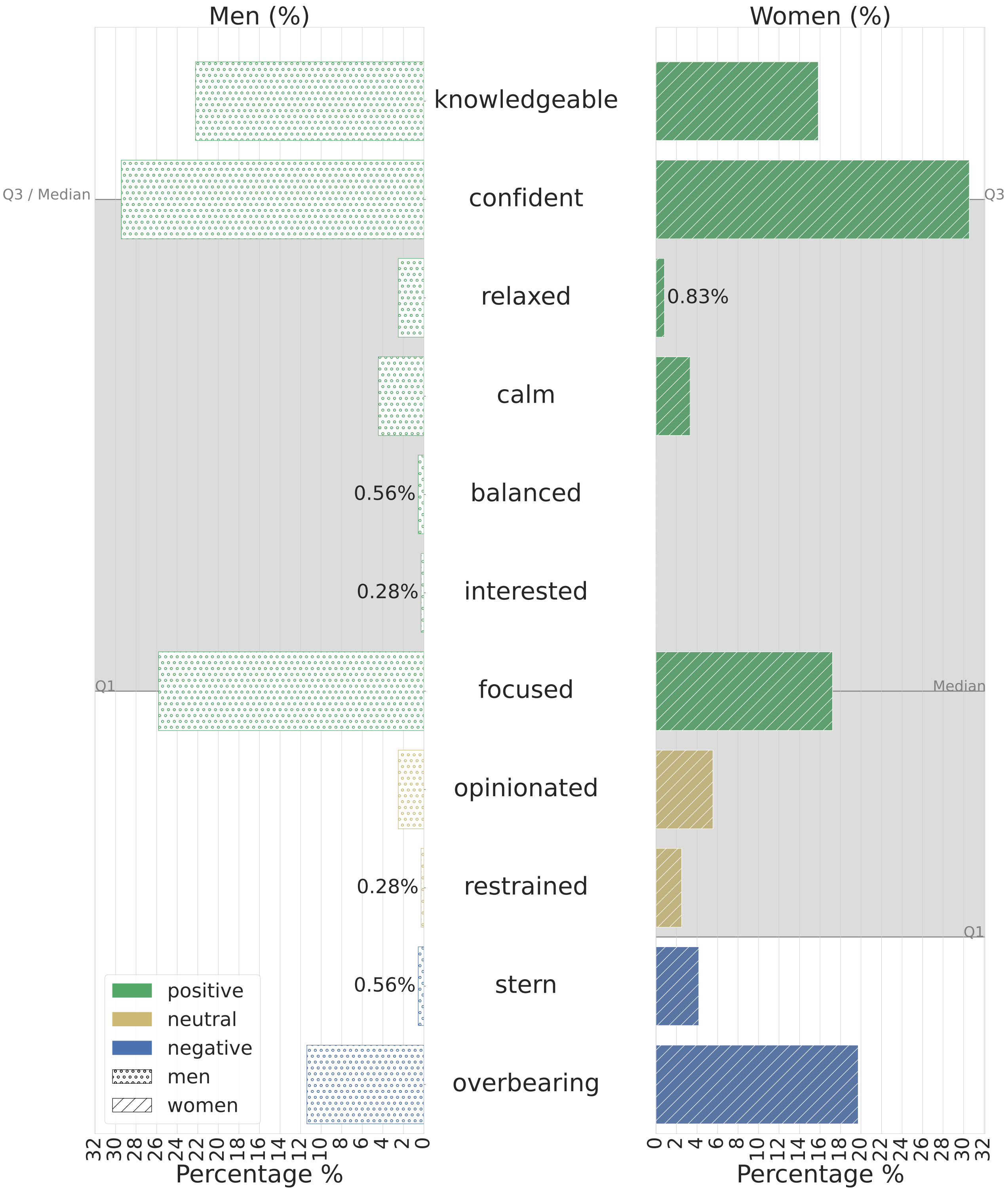


Figure 14: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *neutrality*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

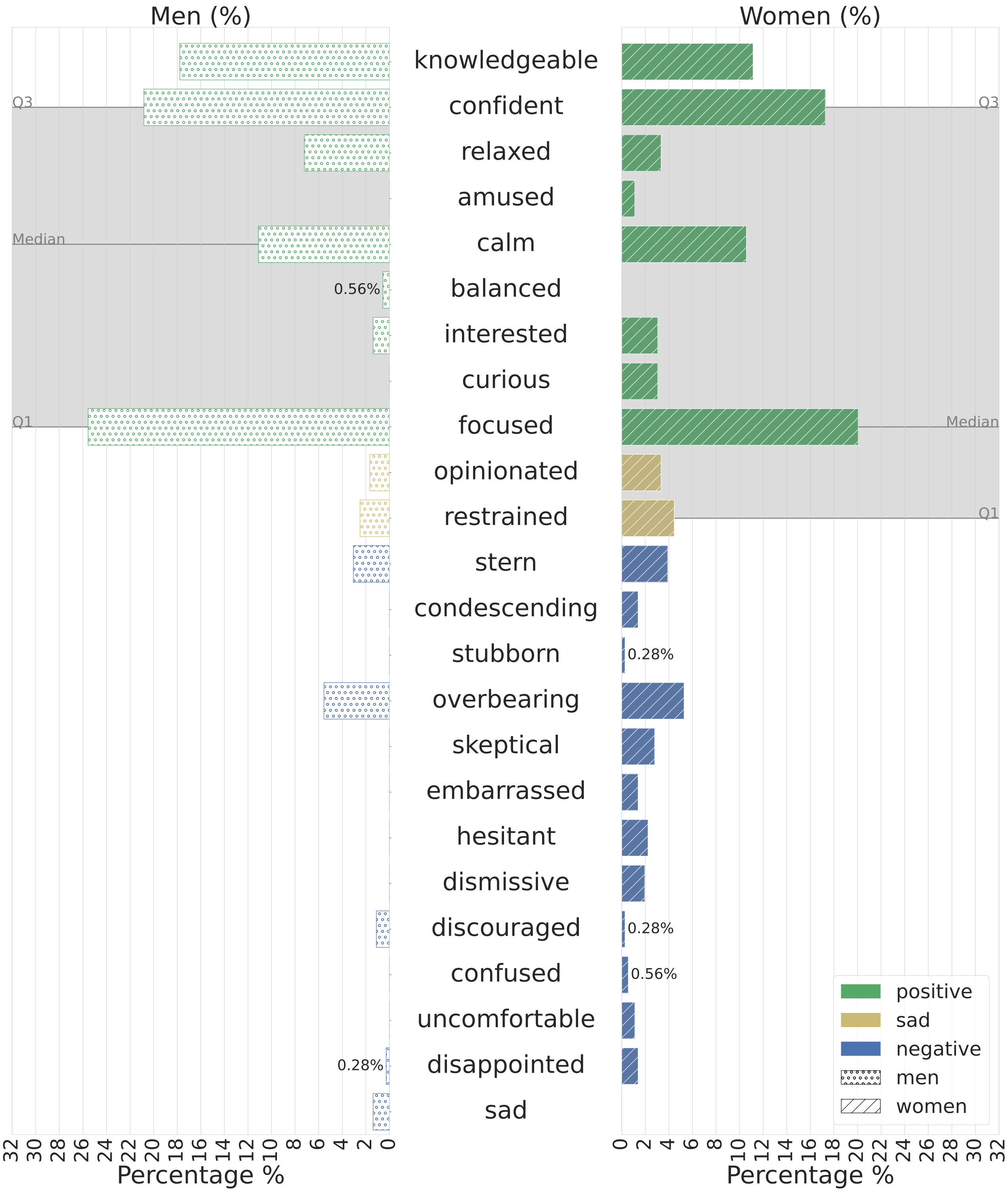


Figure 15: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *sadness*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.

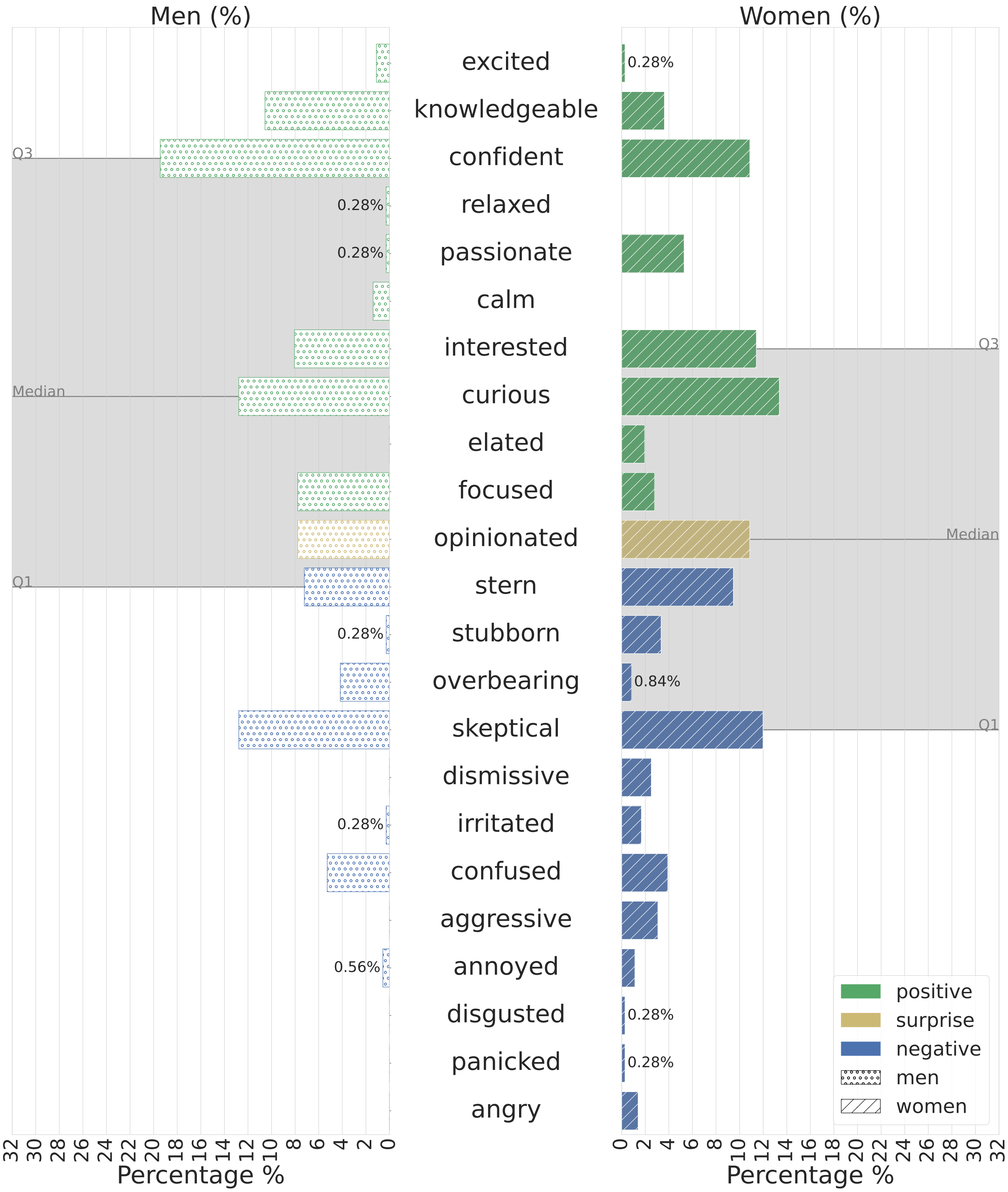


Figure 16: A horizontal split bar graph showing the distribution of inferred emotions for men and women expressing *surprise*. The interquartile range is shaded in gray for each gender in their respective sides of the graph.